UNIVERSITY OF CALIFORNIA, SAN DIEGO

Bayesian Methods for Finding Sparse Representations

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in

Electrical Engineering (Intelligent Systems, Robotics & Control)

by

David Paul Wipf

Committee in charge:

Professor Bhaskar D. Rao, Chair
Assistant Professor Sanjoy Dasgupta
Professor Charles Elkan
Professor Kenneth Kreutz-Delgado
Professor Terrence J. Sejnowski
Assistant Professor Nuno Vasconcelos

2006

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT OF THE DISSERTATION

Bayesian Methods for Finding Sparse Representations

by

David Paul Wipf

Doctor of Philosophy in Electrical Engineering

(Intelligent Systems, Robotics & Control)

University of California, San Diego, 2006

Professor Bhaskar D. Rao, Chair

Finding the sparsest or minimum ℓ_0-norm representation of a signal given a (possibly) overcomplete dictionary of basis vectors is an important problem in many application domains, including neuroelectromagnetic source localization, compressed sensing, sparse component analysis, feature selection, image restoration/compression, and neural coding. Unfortunately, the required optimization is typically NP-hard, and so approximate procedures that succeed with high probability are sought.

Nearly all current approaches to this problem, including orthogonal matching pursuit (OMP), basis pursuit (BP) (or the LASSO), and minimum ℓ_p quasi-norm methods, can be viewed in Bayesian terms as performing standard MAP estimation using a fixed, sparsity-inducing prior. In contrast, we advocate empirical Bayesian approaches such as sparse Bayesian learning (SBL), which use a parameterized prior to encourage sparsity through a process called evidence maximization. We prove several

results about the associated SBL cost function that elucidate its general behavior and provide solid theoretical justification for using it to find maximally sparse representations. Specifically, we show that the global SBL minimum is always achieved at the maximally sparse solution, unlike the BP cost function, while often possessing a more limited constellation of local minima than comparable MAP methods which share this property. We also derive conditions, dependent on the distribution of the nonzero model weights embedded in the optimal representation, such that SBL has no local minima. Finally, we demonstrate how a generalized form of SBL, out of a large class of latent-variable models, uniquely satisfies two minimal performance criteria directly linked to sparsity. These results lead to a deeper understanding of the connections between various Bayesian-inspired strategies and suggest new sparse learning algorithms.

Several extensions of SBL are also considered for handling sparse representations that arise in spatio-temporal settings and in the context of covariance component estimation. Here we assume that a small set of common features underly the observed data collected over multiple instances. The theoretical properties of these SBL-based cost functions are examined and evaluated in the context of existing methods. The resulting algorithms display excellent performance on extremely large, ill-posed, and ill-conditioned problems in neuroimaging, suggesting a strong potential for impacting this field and others.

Chapter I

Introduction

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

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

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 Gaussian noise.[1] The goal is to estimate $\boldsymbol{w}$ given $\boldsymbol{t}$ and Φ.

Perhaps the most ubiquitous estimator used for this task is one that maximizes the likelihood of the data $p(\boldsymbol{t}|\boldsymbol{w})$ and is equivalent to the least squares solution. When the dimensionality of $\boldsymbol{w}$ is small relative to the signal dimension (i.e., $M \ll N$), then the ML solution is very effective. However, a rich set of applications exist where the

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

opposite is true, namely, the dimensionality of the unknown $\boldsymbol{w}$ significantly exceeds the signal dimension N. In this situation, the inverse mapping from $\boldsymbol{t}$ to $\boldsymbol{w}$ is said to be *underdetermined*, leading to a severely more complicated estimation task since there are now an infinite number of solutions that could have produced the observed signal $\boldsymbol{t}$ with equal likelihood.

A Bayesian remedy to this indeterminacy assumes that nature has drawn $\boldsymbol{w}$ from some distribution $p(\boldsymbol{w})$ that allows us to narrow the space of candidate solutions in a manner consistent with application-specific assumptions. For example, if we assume that $\boldsymbol{w}$ has been drawn from a zero-mean Gaussian prior with covariance $\sigma_w^2 I$ while $\boldsymbol{\epsilon}$ is independently Gaussian with covariance $\sigma_\epsilon^2 I$, then the *maximum a posteriori* (MAP) estimator of $\boldsymbol{w}$ is given by

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w}) = \Phi^T \left(\lambda I + \Phi\Phi^T\right)^{-1} \boldsymbol{t}, \tag{I.2}$$

where $\lambda \triangleq \sigma_\epsilon^2/\sigma_w^2$. Here the inverse mapping $\Phi^T \left(\lambda I + \Phi\Phi^T\right)^{-1}$ is linear like the forward (generative) model; however, in general this need not be the case.

Use of (I.2) favors estimates $\hat{\boldsymbol{w}}$ with a large number of small nonzero coefficients. Instead, assume now that we have some prior belief that $\boldsymbol{t}$ has been generated by a sparse coefficient expansion, meaning that most of the elements in $\boldsymbol{w}$ are equal to zero. Such inverse solutions can be encouraged by the incorporation of a so-called sparsity-inducing prior, characterized by fat tails and a sharp, possibly infinite, peak at zero [79]. An alternative route to sparsity is to use special so-called *empirical priors* characterized

by flexible parameters that must be estimated (somewhat counterintuitively) from the data itself [66]. The problem in both situations, however, is that the ensuing inverse problem from $\boldsymbol{t}$ to $\boldsymbol{w}$ becomes highly non-linear. Moreover, although as M increases there is a greater possibility that a highly sparse representation exists, the associated estimation task becomes exponentially more difficult, with even modest sized problems becoming insolvable.

In the next section, we will discuss a few relevant applications where sparse representations as described are crucial. We will then more precisely define the types of sparse inverse problems we wish to solve followed by detailed descriptions of several popular Bayesian solutions to these problems. We will conclude by providing an outline of the remainder of this thesis.

I.A Applications

Numerous applications can effectively be reduced to the search for tractable sparse solutions to (I.1) and the associated interpretation of the coefficients that result. Three interrelated examples are signal denoising, compression/coding of high dimensional data, and dictionary learning or sparse component analysis. In the first, the goal is to find a mapping such that signal energy is concentrated in a few coefficients while the noise energy remains relatively distributed, or is relegated to a few noise components of an appropriately fashioned overcomplete dictionary. This allows for thresholding in the transform domain to remove noise while limiting the signal degradation [15, 43]. Secondly, for coding purposes, sparsity can play an important role in redundancy reduc-

tion, leading to efficient representations of signals [64, 68, 96]. It has also been argued that such representations are useful for modelling high dimensional data that may lie in some lower-dimensional manifold [69]. Thirdly, a large number of overcomplete dictionary learning algorithms rely heavily on the assumption that the unknown sources are sparse [31, 50, 52, 53]. These methods typically interleave a dictionary update step with a some strategy for estimating sparse sources at each time point. Here the distinction arises between learning the optimal sources at every time point for a given dictionary and blindly learning an unknown dictionary, which does not necessarily require that we learn the optimal source reconstruction.

Applications of sparsity are not limited to the above as will be discussed in the following subsections. These descriptions represent topics particularly germane to the research contained in this thesis.

I.A.1 Nonlinear Parameter Estimation and Source Localization

Sparse solutions to (I.1) can be utilized to solve a general class of nonlinear estimation problems. Suppose we are confronted with the generative model

$$\boldsymbol{t} = g(\boldsymbol{\alpha}, \Theta) + \boldsymbol{\epsilon} = \sum_{d=1}^{D} \alpha_d f\left(\boldsymbol{\theta}_d\right) + \boldsymbol{\epsilon} \tag{I.3}$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_D]^T$ is an unknown coefficient vector, $\Theta = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D] \in \mathbb{R}^{R \times D}$ is an unknown parameter matrix, and $f : \mathbb{R}^R \rightarrow \mathbb{R}^N$ is a known nonlinear function. Given $\boldsymbol{t}$ and $f(\cdot)$, the goal here is to learn $\boldsymbol{\alpha}$ and Θ. A surprisingly large number of

parameter estimation tasks, including many ML problems, can be expressed in this form. We will refer to this problem as *source localization*, since often the parameters Θ and α correspond with the location and amplitude of some source activity of interest. Note also that D, which can be considered the number of active sources, may be unknown.

Assuming that $f(\cdot)$ is highly nonlinear, then estimation of $\boldsymbol{\alpha}$ and Θ can be extremely difficult and subject to numerous local optima. However, by densely sampling Θ space, this estimation task can be mapped into the sparse representation framework, assuming D is sufficiently smaller than N. This requires a dictionary to be formed with columns $\boldsymbol{\phi}_i = f(\boldsymbol{\theta}_i)$, with sampling sufficiently dense to obtain the required accuracy. The nonzero coefficients obtained from learning a sparse solution $\hat{\boldsymbol{w}}$ correspond with the unknown α_d, while the corresponding selected columns of Φ signify, to within the quantization accuracy, the values of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D$.

This method generally has a significant advantage over more traditional nonlinear optimization techniques, in that results are much less dependent on the initialization that is used and the local minimum profile of (I.3). This occurs because, in some sense, the sparse approximation framework considers 'all' source locations initially and then prunes away unsupported values in a competitive process. While local minima may still exist, they are local minima with respect to a more global solution space and typically a reasonable solution is obtainable. In contrast, minimizing (I.3) directly using some descent method considers only a single solution at a time and proceeds based only on local information in the neighborhood of this solution. Moreover, it requires explicit knowledge of D, whereas in theory, the sparse approximation framework can

learn this value from the data (i.e., upon convergence, the number of nonzero elements in $\hat{\boldsymbol{w}}$ approximately equals D).

The next section, in part, addresses a particular instance of this methodology related to neuroimaging. Another very relevant example (not discussed) involving this framework is direction-of-arrival estimation [34, 60].

I.A.2 Neuroelectromagnetic Source Imaging

Recent non-invasive imaging techniques based on electroencephalography (EEG) and magnetoencephalography (MEG) draw heavily on the resolution of underdetermined inverse problems using (implicitly or explicitly) a sparse Bayesian formulation [33, 40, 74, 75, 100]. At least two fundamental issues can be addressed under a Bayesian sparse recovery framework. The first relates to source localization, the second uses sparse component analysis to remove artifacts and analyze macro-level brain dynamics.

MEG and EEG use an array of sensors to take EM field measurements from on or near the scalp surface with excellent temporal resolution. In both cases, the observed field is generated by the same synchronous, compact current sources located within the brain. Because the mapping from source activity configuration to sensor measurement is many to one, accurately determining the spatial locations of these unknown sources is extremely difficult. In terms of the generative model (I.1), the relevant localization problem can be posed as follows: The measured EM signal is $\boldsymbol{t}$ where the dimensionality N is equal to the number of sensors. The unknown coefficients $\boldsymbol{w}$ are the (discretized) current values at M candidate locations distributed throughout the cortical surface. These

candidate locations are obtained by segmenting a structural MR scan of a human subject and tesselating the gray matter surface with a set of vertices. The i-th column of Φ then represents the signal vector that would be observed at the scalp given a unit current source at the i-th vertex. Multiple methods (based on the physical properties of the brain and Maxwell's equations) are available for this computation [88].

To obtain reasonable spatial resolution, the number of candidate source locations will necessarily be much larger than the number of sensors. The salient inverse problem then becomes the ill-posed estimation of these activity or source regions. Given the common assumption that activity can be approximated by compact cortical regions, or a collection of equivalent current dipoles, the sparse recovery framework is particularly appropriate. Source localization using a variety of implicit Bayesian priors have been reported with varying degrees of success [33, 42, 71, 75, 100]. This problem can also be viewed as an instance of (I.3), where $\boldsymbol{\theta}_d$ represents the 3D coordinates of a particular current dipole and the corresponding α_d is the source amplitude, which is assumed to be oriented orthogonal to the cortical surface. The case of unconstrained dipoles can be handled by adding two additional source components tangential to the cortex.

Direct attempts to solve (I.3) using nonlinear optimization exhibit rather poor performance, e.g., only two or three sources can be reliably estimated in simulation, due to the presence of numerous local minima. In contrast, using the sparse representation framework upwards of fifteen sources can be consistently recovered [75]. Regardless, the estimation task remains a challenging problem.

A second application of sparse signal processing methods to EEG/MEG in-

volves artifact removal and source separation. Whereas the dictionary Φ is computed directly using standard physical assumptions to solve the localization task, here we assume an unknown decomposition Φ that is learned from a series of observed EEG/MEG signals $\boldsymbol{t}(n)$ varying over the time index n. The dimensionality of the associated $\boldsymbol{w}(n)$ is interpreted as the number of unknown neural sources or causes plus the number of artifactual sources and noise. A variety of algorithms exist to iteratively estimate both Φ (dictionary update) and $\boldsymbol{w}(n)$ (signal update) using the a priori assumption that the latter time courses are sparse. In practice, it has been observed that the resulting decomposition often leads to a useful separation between unwanted signals (e.g., eye blinks, heart beats, etc.) and distinct regions of brain activity or event-related dynamics [48, 75]. Note that all of the sparse Bayesian methods discussed in this thesis, when combined with a dictionary update rule, can conceivably be used to address this problem.

In summary, high-fidelity source localization and dynamic source detection/separation serve to advance non-invasive, high temporal resolution electromagnetic brain imaging technologies that heretofore have suffered from inadequate spatial resolution and ambiguous dynamics. The solution of a possibly underdetermined system using the assumption of sparsity plays a crucial role is solving both problems.

I.A.3 Neural Coding

This section focuses on the role of sparse representations operating at the level of individual neurons within a population. A mounting collection of evidence, both experimental and theoretical, suggests that the mammalian cortex employs some type of

sparse neural code to efficiently represent stimuli from the environment [67, 72, 101]. In this situation, the observed data $\boldsymbol{t}$ represent a particular stimuli such as a visual scene projected onto the retina. Each column of the matrix Φ models the receptive field of a single neuron, reflecting the particular feature (e.g., such as an oriented edge) for which the neuron is most responsive. The vector $\boldsymbol{w}$ then contains the response properties of a set of M neurons to the input stimulus $\boldsymbol{t}$, with a sparse code implying that most elements of $\boldsymbol{w}$, and therefore most neurons, are inactive at any given time while a small set with stimulus-correlated receptive fields maintain substantial activity or firing rates. In many situations the number of neurons available for coding purposes is much greater than the intrinsic dimensionality of the stimulus, possibly reflecting the existence of a large number of potential causes underlying the space of potential stimuli [69]. This requires that the response properties of many cortical neurons are effectively nonlinear, consistent with sparse inverse mappings associated with (I.1) and a variety of empirical data.

A key pointer to the potential role of sparse coding in the processing of sensory data came in the seminal work by Olshausen and Field [67, 68]. Here an iterative algorithm is proposed to learn a matrix Φ that encourages/faciliates sparse representations $\boldsymbol{w}$ when presented with patches from natural images $\boldsymbol{t}$.[2] With no other assumptions, the Φ that results from this procedure contains columns representing a full set of spatially localized, oriented, and bandpass receptive fields consistent with those observed in the simple cells of the mammalian primary visual cortex. This result reinforces the notion

[2] We will briefly discuss learning the dictionary Φ in Section VIII.C

that a sparse coding principle could underly the brain's neural representation of natural stimuli.

As summarized in [69], sparse coding strategies offers several advantages to an individual organism. For example, sparse codes from overcomplete dictionaries lead to efficient, less redundant representations and may make it easier for higher areas of the brain to learn relevant structure and causal relationships embedded in sensory inputs. Recent work using overcomplete representations in a biologically motivated recognition systems support this assertion [65]. Moreover, in understanding how the brain processes information, the possibility exists for building better artificial systems for robust compression and recognition.

I.A.4 Compressed Sensing

Compressed sensing begins with the assumption that some sparse data vector of interest $\boldsymbol{w}$ exists in a high-dimensional space [8, 20, 102]. We would like to have access to $\boldsymbol{w}$ but direct measurement of each element in $\boldsymbol{w}$ is assumed to be very expensive. As such, the objective is to obtain an accurate estimate by measuring only a few random projections of $\boldsymbol{w}$. In this situation, each row of Φ becomes a random vector and each element of $\boldsymbol{t}$ is the associated measurement/projection. The goal is then to recover $\boldsymbol{w}$ using only the observed projections $\boldsymbol{t}$ and the knowledge that $\boldsymbol{w}$ is sparse.

I.B Definitions and Problem Statement

To simplify matters, it is useful to define

$$\|\boldsymbol{w}\|_0 \triangleq \sum_{i=1}^{M} \mathcal{I}\left[|w_i| > 0\right], \tag{I.4}$$

where $\mathcal{I}[\cdot]$ denotes the indicator function. $\|\cdot\|_0$ is a *diversity* measure since it counts the number of elements in $\boldsymbol{w}$ that are not equal to zero. It is also commonly referred to as the ℓ_0 norm, although it is not actually a true norm. This is in contrast to *sparsity*, which counts the number of elements that are strictly equal to zero. The two are related by

$$\text{diversity} = M - \text{sparsity}. \tag{I.5}$$

The nonzero elements of any weight vector are referred to as *active sources*.

With regard to the dictionary Φ, *spark* is defined as the smallest number of linearly dependent columns [17]. 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 [34], 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 sparse representation problem, we begin with the most straightforward case where $\boldsymbol{\epsilon} = 0$. If Φ is overcomplete, then we are presented with an ill-posed inverse problem unless further assumptions are made. For example, if a matrix of gen-

erating weights $\boldsymbol{w}_{\text{gen}}$ satisfies

$$\|\boldsymbol{w}_{\text{gen}}\|_0 \ < \ \text{spark}(\Phi)/2, \tag{I.6}$$

then no other solution $\boldsymbol{w}$ can exist such that $\boldsymbol{t} = \Phi\boldsymbol{w}$ and $\|\boldsymbol{w}\|_0 \leq \|\boldsymbol{w}_{\text{gen}}\|_0$. Results of this nature have been derived in [34] and later discussed in [17]. Furthermore, if we assume suitable randomness on the nonzero entries of $\boldsymbol{w}_{\text{gen}}$, then this result also holds under the alternative inequality

$$\|\boldsymbol{w}_{\text{gen}}\|_0 < \text{spark}(\Phi) - 1, \tag{I.7}$$

which follows from the analysis in Section II.B.2. Given that one or both of these conditions hold, then recovering $\boldsymbol{w}_{\text{gen}}$ is tantamount to solving

$$\boldsymbol{w}_{\text{gen}} = \boldsymbol{w}_0 \triangleq \arg\min_{\boldsymbol{w}} \ \|\boldsymbol{w}\|_0, \qquad \text{s.t. } \boldsymbol{t} = \Phi\boldsymbol{w}. \tag{I.8}$$

This has sometimes been called the *exact sparse recovery problem*, since any solution forces exact (strict) equality. In general, (I.8) is NP-hard so approximate procedures are in order. In Chapters II and III, we will examine the solution of (I.8) in further detail, which has also been studied exhaustively by others [17, 29, 35, 95]. For the remainder of this thesis, whenever $\boldsymbol{\epsilon} = 0$, we will assume that $\boldsymbol{w}_{\text{gen}}$ satisfies (I.6) or (I.7), and so $\boldsymbol{w}_0$ (the maximally sparse solution) and $\boldsymbol{w}_{\text{gen}}$ can be used interchangeably.

When $\boldsymbol{\epsilon} \neq 0$, things are decidedly more nebulous. Because noise is present,

we typically do not expect to represent $\boldsymbol{t}$ exactly, suggesting the relaxed optimization problem

$$\boldsymbol{w}_0(\lambda) \triangleq \arg\min_{\boldsymbol{w}} \ \|\boldsymbol{t} - \Phi\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_0, \tag{I.9}$$

where λ is a trade-off parameter balancing estimation quality with sparsity. Unfortunately, solving (I.9) 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 $\boldsymbol{w}_{\text{gen}}$, 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 (e.g., in the source localization problem). Although not the central focus of this thesis, if the ultimate goal is compression of $\boldsymbol{t}$, then the solution of (I.9) may trump other concerns.

From a conceptual standpoint, (I.9) can be recast in Bayesian terms by adding constants and applying a $\exp[-(\cdot)]$ transformation. This leads to a Gaussian likelihood function $p(\boldsymbol{t}|\boldsymbol{w})$ with λ-dependent variance

$$p(\boldsymbol{t}|\boldsymbol{w}) \propto \exp\left[-\frac{1}{\lambda}\|\boldsymbol{t} - \Phi\boldsymbol{w}\|_2^2\right] \tag{I.10}$$

and a prior distribution given by

$$p_0(\boldsymbol{w}) \propto \exp\left[-\|\boldsymbol{w}\|_0\right]. \tag{I.11}$$

In weight space, this improper prior maintains a sharp peak when a weight equals zero

and heavy (in fact uniform) 'tails' everywhere else. The optimization problem from (I.9) can equivalently be written as

$$\boldsymbol{w}_0(\lambda) \equiv \arg\max_{\boldsymbol{w}} p(\boldsymbol{t}|\boldsymbol{w})p_0(\boldsymbol{w}) = \arg\max_{\boldsymbol{w}} \frac{p(\boldsymbol{t}|\boldsymbol{w})p_0(\boldsymbol{w})}{p(\boldsymbol{t})} = \arg\max_{\boldsymbol{w}} p_0(\boldsymbol{w}|\boldsymbol{t}). \quad \text{(I.12)}$$

Therefore, (I.9) can be viewed as a challenging MAP estimation task, with a posterior characterized by numerous locally optimal solutions.

I.C Finding Sparse Representations vs. Sparse Regression

Before proceeding to discuss various Bayesian strategies for solving (I.8) and (I.9), it is important to make the following distinction. In many ways, the problem of finding sparse representations can be thought of as regression where sparsity acts as a regularization mechanism to avoid overfitting the training data $\boldsymbol{t}$ as well as potentially leading to more interpretable model structures. Nonetheless, there remains one subtle difference: while the ultimate goal of regression is to minimize generalization error (i.e., error on evaluation data not available during model training), here we are more concerned with the actual sparse representation of the training data $\boldsymbol{t}$. This distinction is reflected in the results of this paper, which focus on how well a particular method is likely to solve (I.8) or (I.9). With the exception of Chapter IV, which discusses how certain sparse Bayesian strategies relate to probability mass in a full predictive distribution, performance on unseen data is not emphasized. However, for the interested reader, there is a known relationship between sparsity of fit and generalization performance as

discussed in [39]. And so many sparsity-based regression schemes have demonstrated marked success [86, 94] relative to predictive accuracy.

I.D Bayesian Methods

Directly solving (I.8) or (I.9) poses a difficult optimization challenge both because of the sharp discontinuity at zero and the combinatorial number of local minima. However, simple greedy methods offer a convenient means for providing at least locally optimal solutions. For example, there are forward sequential selection methods based on some flavor of Matching Pursuit (MP) [61]. 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. Although not our focus, we will sometimes consider *Orthogonal Matching Pursuit* (OMP), a popular variant of MP that can be viewed as finding a local minimum to (I.8) or (I.9) [12].

An alternative strategy is to replace the troublesome prior $p_0(\boldsymbol{w})$ with a distribution that, while still encouraging sparsity, is somehow more computationally convenient. Bayesian approaches to the sparse approximation problem that follow this route have typically been divided into two categories: (i) *maximum a posteriori* (MAP) estimation using a fixed, computationally tractable family of priors and, (ii) *empirical Bayesian* approaches that employ a flexible, parameterized prior that is 'learned' from the data. We discuss both techniques in turn.

I.D.1 MAP Estimation

A natural solution to the computational difficulty associated with $p_0(\boldsymbol{w})$ is to choose the best possible convex relaxation, which turns out to be the standardized Laplacian distribution

$$p_1(\boldsymbol{w}) \propto \exp\left(-\sum_{i=1}^{M} |w_i|\right). \tag{I.13}$$

Often referred to as Basis Pursuit (BP)[10], the LASSO [93], or ℓ_1-norm regularized regression, MAP estimation using this prior involves solving

$$\boldsymbol{w}_{\text{BP}} = \arg\min_{\boldsymbol{w}} \ \|\boldsymbol{t} - \Phi\boldsymbol{w}\|_2^2 + \lambda \sum_{i=1}^{M} |w_i|. \tag{I.14}$$

This convenient convex cost function can be globally minimized using a variety of standard optimization packages. The properties of the BP cost function and algorithms for its minimization have been explored in [17, 79, 96]. While often effective, the BP solution sometimes fails to be sufficiently sparse in practice. This subject will be discussed more in later sections.

A second prior that is sometimes chosen in place of the Lalpacian is the scale-invariant Jeffreys prior given by

$$p_J(\boldsymbol{w}) \propto \prod_{i=1}^{M} \frac{1}{|w_i|}. \tag{I.15}$$

Although technically an improper prior [4], the heavy tails and sharp (in fact infinite) peak at zero mirror the characteristics of a sparse distribution. The MAP estimation

problem then becomes

$$\boldsymbol{w}_{\text{Jeffreys}} = \arg\min_{\boldsymbol{w}} \ \|\boldsymbol{t} - \Phi\boldsymbol{w}\|_2^2 + \lambda \sum_{i=1}^{M} \log|w_i|, \tag{I.16}$$

where the regularization term here has sometimes been referred to as the Gaussian entropy [78]. This Jeffreys-based cost function suffers from numerous local minima, but when given a sufficiently good initialization, can potentially find solutions that are closer to $\boldsymbol{w}_{\text{gen}}$ than $\boldsymbol{w}_{\text{BP}}$. From an implementational standpoint, (I.16) can be solved using the algorithms derived in [27, 34].

Thirdly, we weigh in the generalized Gaussian prior

$$p(\boldsymbol{w}) \propto \exp\left(-\sum_{i=1}^{M} |w_i|^p\right), \tag{I.17}$$

where $p \in [0, 1]$ is a user-defined parameter. The corresponding optimization problem, which is sometimes called the FOCUSS algorithm, involves solving

$$\boldsymbol{w}_{\text{FOCUSS}} = \arg\min_{\boldsymbol{w}} \ \|\boldsymbol{t} - \Phi\boldsymbol{w}\|_2^2 + \lambda \sum_{i=1}^{M} |w_i|^p. \tag{I.18}$$

This is very similar to a procedure originally outlined in [51] based on work in [2]. If $p \to 0$, the FOCUSS cost function approaches (I.9). 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 $\boldsymbol{w}'$ such that $\boldsymbol{t} = \Phi\boldsymbol{w}'$ and $\|\boldsymbol{w}'\|_0 \leq N$, regardless of λ. The set of initial conditions whereby we will actually

converge to $\boldsymbol{w}_0(\lambda)$ has measure zero. When $p = 1$, FOCUSS reduces to an interior point method for implementing BP [78]. The FOCUSS framework also includes the Jeffreys approach as a special case as shown in Appendix VI.H.1. In practice, it is sometimes possible to jointly select values of p and λ such that the algorithm outperforms both BP and Jeffreys. In general though, with BP, Jeffreys, and FOCUSS, λ must be tuned with regard to a particular application. Also, in the limit as λ becomes small, we can view each MAP algorithm as minimizing the respective diversity measure subject to the constraint $\boldsymbol{t} = \Phi\boldsymbol{w}$. This is in direct analogy to (I.8).

Because the FOCUSS framework can accommodate all the the sparsity priors mentioned above, and for later comparison purposes with other methods, we include the FOCUSS update rules here. These rules can be derived in a variety of settings, including the EM algorithm [70]. This requires expressing each prior in terms of a set of latent variables $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_M]^T$ which are treated as hidden data. Details will be discussed further in Chapter V. The E-step requires computing the expected value of $\boldsymbol{\gamma}$ given $\boldsymbol{t}$ and the current weight estimate $\hat{\boldsymbol{w}}$ using

$$\gamma_i = |\hat{w}_i|^{2-p}, \quad \forall i, \tag{I.19}$$

while the M-step updates $\hat{\boldsymbol{w}}$ via

$$\hat{\boldsymbol{w}} = \Gamma\Phi^T \left(\lambda I + \Phi\Gamma\Phi^T\right)^{-1} \boldsymbol{t}, \tag{I.20}$$

where $\Gamma \triangleq \mathrm{diag}(\boldsymbol{\gamma})$. These updates are guaranteed to converge monotonically to a local

minimum (or saddle point) of (I.18).

In the low-noise limit, i.e., as $\lambda \to 0$, the M-step can be seamlessly replaced with

$$\hat{\boldsymbol{w}} = \Gamma^{1/2} \left(\Phi \Gamma^{1/2} \right)^{\dagger} \boldsymbol{t}, \tag{I.21}$$

where $(\cdot)^{\dagger}$ denotes the Moore-Penrose pseudo-inverse. This result follows from the general identity

$$\lim_{\varepsilon \to 0} U^T \left(\varepsilon I + U U^T \right)^{-1} = U^{\dagger}. \tag{I.22}$$

In this manner, all of the methods from above can be used to approximate (I.8). We observe that at each iteration $\hat{\boldsymbol{w}}$ is feasible, i.e., $\boldsymbol{t} = \Phi \hat{\boldsymbol{w}}$. This assumes that $\boldsymbol{t}$ is in the span of the columns of Φ associated with nonzero elements in $\boldsymbol{\gamma}$, which will always be the case if $\boldsymbol{t}$ is in the span of Φ and all elements of $\boldsymbol{\gamma}$ are initialized to nonzero values.

I.D.2 Empirical Bayes

All of the methods discussed in the previous section for estimating $\boldsymbol{w}_{\text{gen}}$ involve searching some implicit posterior distribution for the mode by solving $\arg\max_{\boldsymbol{w}} p(\boldsymbol{w}, \boldsymbol{t}) = \arg\max_{\boldsymbol{w}} p(\boldsymbol{t}|\boldsymbol{w}) p(\boldsymbol{w})$, where $p(\boldsymbol{w})$ 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 as with BP, a unimodal posterior results and mode-finding is greatly simplified; however, the resultant posterior mode may not be sufficiently sparse, and therefore $\boldsymbol{w}_{\text{BP}}$ may be unrepresentative of $\boldsymbol{w}_{\text{gen}}$ (or the maximally sparse solution $\boldsymbol{w}_0$). 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 $\boldsymbol{w}_{\text{gen}}$, 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 sparse representation problem is developed using the concept of *automatic relevance determination* (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 [57, 66]. These ideas have also been applied to Bayesian kernel machines [94]. 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 $p(\boldsymbol{t}|\boldsymbol{w})$ to be Gaussian with noise variance λ consistent with the likelihood model (I.10) and previous Bayesian methods. Generally, λ is assumed to be known; however, the case where λ is not known will be discussed briefly in Section VIII.A. Next, application of ARD involves assigning to each coefficent w_i the independent Gaussian prior

$$p(w_i;\gamma_i) \triangleq \mathcal{N}(0,\gamma_i)\,, \tag{I.23}$$

where γ_i is an unknown variance parameter [94]. (In Chapter V we will address how these γ_i parameters relate to those from the MAP section.) By combining each of these

priors, we arrive at a full weight prior

$$p(\boldsymbol{w};\boldsymbol{\gamma}) = \prod_{i=1}^{M} p(w_i;\gamma_i), \tag{I.24}$$

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 $\boldsymbol{w}$ then becomes

$$p(\boldsymbol{w}|\boldsymbol{t};\boldsymbol{\gamma}) = \frac{p(\boldsymbol{w},\boldsymbol{t};\boldsymbol{\gamma})}{\int p(\boldsymbol{w},\boldsymbol{t};\boldsymbol{\gamma}) d\boldsymbol{w}} = \mathcal{N}(\boldsymbol{\mu}, \Sigma), \tag{I.25}$$

with mean and covariance given by

$$\begin{aligned} \Sigma &\triangleq \text{Cov}[\boldsymbol{w}|\boldsymbol{t};\boldsymbol{\gamma}] = \Gamma - \Gamma\Phi^T\Sigma_t^{-1}\Phi\Gamma, \\ \boldsymbol{\mu} &\triangleq \text{E}[\boldsymbol{w}|\boldsymbol{t};\boldsymbol{\gamma}] = \Gamma\Phi^T\Sigma_t^{-1}\boldsymbol{t}, \end{aligned} \tag{I.26}$$

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

Since it is typically desirable to have a point estimate for $\boldsymbol{w}_{\text{gen}}$, we may enlist $\boldsymbol{\mu}$, the posterior mean, for this purpose. Sparsity is naturally achieved whenever a γ_i is equal to zero. This forces the posterior to satisfy $\text{Prob}(w_i = \boldsymbol{0}|\boldsymbol{t};\gamma_i = 0) = 1$, ensuring that the posterior mean of the i-th element, μ_i, will be zero as desired. Thus, estimating the sparsity profile of some $\boldsymbol{w}_{\text{gen}}$ is 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.

Hyperparameter Estimation: The 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 $\boldsymbol{t}$. 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 $\boldsymbol{w}$ as nuisance parameters and integrate them out [56]. The marginal likelihood that results is then maximized with respect to $\boldsymbol{\gamma}$, leading to the ARD-based cost function

$$\begin{aligned} \mathcal{L}(\boldsymbol{\gamma}) &\triangleq -2\log\int p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w};\boldsymbol{\gamma})d\boldsymbol{w} = -2\log p\left(\boldsymbol{t};\boldsymbol{\gamma}\right) \\ &\equiv \log|\Sigma_t| + \boldsymbol{t}^T\Sigma_t^{-1}\boldsymbol{t}, \end{aligned} \tag{I.27}$$

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 $\boldsymbol{w}$ [36, 37]. 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 [4, 56, 66].

Two ways have been proposed to minimize $\mathcal{L}(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$. (Section VII.B.1 briefly discusses additional possibilities.) First, treating the unknown weights

$\boldsymbol{w}$ as hidden data, we can minimize this expression over $\boldsymbol{\gamma}$ using a simple EM algorithm as proposed in [13, 37, 94] for covariance estimation. For the E-step, this requires computation of the posterior moments using (I.26), while the M-step is expressed via the update rule

$$\gamma_i^{(\text{new})} = \mu_i^2 + \Sigma_{ii}, \qquad \forall i = 1, \ldots, M. \tag{I.28}$$

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

Secondly, at the expense of proven convergence, we may instead optimize (I.27) by taking the derivative with respect to $\boldsymbol{\gamma}$, equating to zero, and forming a fixed-point equation that typically leads to faster convergence [56, 94]. Effectively, this involves replacing the M-step from above with

$$\gamma_i^{(\text{new})} = \frac{\mu_i^2}{1 - \gamma_i^{-1}\Sigma_{ii}}, \qquad \forall i = 1, \ldots, M. \tag{I.29}$$

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 (I.28). In the context of kernel regression using a complete dictionary (meaning $N = M$), use of (I.29), along with a modified form of (I.26),[3] 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) [94]. Similar update rules have also been effectively applied

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

to an energy prediction competition under the guise of ARD [57]. For our purposes, we choose the label 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 SBL iterations can be easily adapted to handle the limit as $\lambda \to 0$ using the modified moments

$$\Sigma = \left[I - \Gamma^{1/2} \left(\Phi \Gamma^{1/2} \right)^{\dagger} \Phi \right] \Gamma, \qquad \boldsymbol{\mu} = \Gamma^{1/2} \left(\Phi \Gamma^{1/2} \right)^{\dagger} \boldsymbol{t}, \tag{I.30}$$

This is particularly useful if we wish to solve (I.8). Again we are ensured a feasible solution will be produced at each iteration with a sparsity profile dictated by $\boldsymbol{\gamma}$.

I.D.3 Summary of Algorithms

Given observation data $\boldsymbol{t}$ and a dictionary Φ, all of the MAP and SBL procedures can be summarized by the following collection of steps:

1. Initialize the hyperparameters $\boldsymbol{\gamma}$, e.g., $\boldsymbol{\gamma} := \mathbf{1}$ or perhaps a non-negative random initialization.

2. For SBL, compute Σ and $\boldsymbol{\mu}$ using (I.26), or in the noiseless case, using (I.30).[4] For MAP estimation, only $\boldsymbol{\mu}$ need be computed, which equals the $\hat{\boldsymbol{w}}$ update.

[4] Note that off-diagonal elements of Σ need not be computed.

3. For SBL update $\boldsymbol{\gamma}$ using the EM rule (I.28) or the faster fixed-point rule (I.29). For MAP estimation, use the update rule (I.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 $\boldsymbol{w}_{\text{gen}}$, choose $\boldsymbol{\mu}^*$, the value of $\boldsymbol{\mu}$ evaluated at $\boldsymbol{\gamma}^*$.

6. Given that $\boldsymbol{\gamma}^*$ is sparse, the resultant estimator $\boldsymbol{\mu}^*$ will necessarily be sparse as well.

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).

I.E Thesis Outline

The remainder of this thesis is organized as follows. In Chapter II we provide a detailed, comparative analysis of the global and local minima of both MAP and empirical Bayesian methods for finding sparse representations. Special focus is placed on SBL, which is shown to have a number of attractive features. Chapter III then describes how the distribution of the nonzero weights affects the SBL algorithm's ability to find maximally sparse solutions and provides evidence for its superior performance over popular competing methods.

Chapter IV switches gears and provides a more intuitive motivation for why SBL is able to achieve sparse solutions. It also details how the SBL model relates to the

probability mass in a full Bayesian model with a sparse prior.

The analysis thus far centers on comparing SBL with a few popular MAP-based methods. However, recent results have shown that a rich set of latent variable models with sparse priors can be efficiently optimized leading to alternative MAP and empirical Bayesian approaches. Chapter V provides a theoretical examination of these types of models and demonstrates a unique procedure, out of all the possibilities, that satisfies two minimal performance criteria related to the recovery of sparse sources. SBL and BP can be viewed as special cases. The distinction between factorial and non-factorial priors is also discussed.

While the standard sparse recovery model is sufficient for many applications, additional flexibility is needed in certain cases. The next two chapters extend this framework to handle more general modelling assumptions. In Chapter VI, we assume that multiple response vectors $\boldsymbol{t}$ are available that were putatively generated by the same underlying set of features. The goal is then to assimilate the information contained in each response so as to more reliably estimate the correct sparsity profile. An extension of SBL for solving this problem is derived and analyzed.

Chapter VII further extends the flexibility of sparsity-based MAP and empirical Bayesian methods to handle the more general problem of covariance component estimation. This added generality is especially salient in the context of neuroelectromagnetic source imaging, where we derive some new algorithms and point to connections between existing source imaging approaches.

Chapter VIII addresses some practical issues that arise in the search for sparse

representations. It also discusses the extension to dictionary learning, deriving a particularly effective algorithm for learning orthogonal transforms that encourage sparisty.

Chapter IX contains brief concluding remarks.

Chapter II

Analysis of Global and Local Minima

This chapter is primarily aimed at evaluating the properties of global and local minima of the empirial Bayesian SBL algorithm and its relationship with more established MAP methods. Ideally, a given method should have a global minimum that closely coincides with $\boldsymbol{w}_{\text{gen}}$ while maintaining as few suboptimal local minima as possible. The major result of this chapter is showing that, while the global minimum of both SBL and certain MAP procedures are guaranteed to correspond with maximally sparse solutions under certain conditions, the former has substantially fewer local minima. We also derive necessary conditions for local minima to occur and quantify worst-case performance at locally minimizing solutions. Several empirical results (here and in later chapters) corroborate these findings.

Much of the analysis will focus on the exact recovery problem (I.8), i.e., finding the maximally sparse solution $\boldsymbol{w}_0$ which we will assume equals $\boldsymbol{w}_{\text{gen}}$. The exact recovery case is a useful starting point for comparing methods because the analysis is

more straightforward and there is no ambiguity involved regarding how the trade-off parameter λ should be chosen. Moreover, many of the insights gained through this process carry over to the case where noise is present and we are forced to accept some error between $\boldsymbol{t}$ and the estimate $\Phi\hat{\boldsymbol{w}}$. Extensions to the noisy case are briefly addressed, but will be explored further in Section VI.E.3 in a slightly more general context.

II.A Preliminaries

Consistent with previous discussion, we say that a dictionary Φ satisfies the *unique representation property* (URP) if every subset of N columns of Φ forms a basis in $\mathbb{R}^N$. This property will be satisfied almost surely for dictionaries composed of iid Gaussian elements, or dictionaries whose columns have be drawn uniformly from the surface of a unit hypersphere. A *basic feasible solution* (BFS) is defined as a solution vector $\boldsymbol{w}$ such that $\boldsymbol{t} = \Phi\boldsymbol{w}$ and $\|\boldsymbol{w}\|_0 \leq N$. As will be explained more below, the locally minimizing solutions for all algorithms considered are achieved at BFS. A *degenerate* BFS has strictly less than N nonzero entries; however, the vast majority of local minima are non-degenerate, containing exactly N nonzero entries.

Regarding algorithmic performance in obtaining sparse solutions, we define two types of errors. First, a *convergence error* refers to the situation where an algorithm converges to a non-global minimum of its cost function that does not equal $\boldsymbol{w}_0$. In contrast, a *structural error* implies that an algorithm has reached the global minimum of its cost function (or a local minima with lower cost than is achievable at the maximally sparse solution $\boldsymbol{w}_0$), but this solution does not equal $\boldsymbol{w}_0$.

II.B MAP Methods

This section discusses the properties of global and local solutions using standard MAP procedures. This leads to a discussion of what we term a *local sparsity maximization* (LSM) algorithm.

II.B.1 MAP Global Minima and Maximally Sparse Solutions

The MAP methods from Section I.D.1 applied to the exact sparse problem (I.8) reduce to solving either

$$\min_{w} \sum_{i=1}^{M} \log |w_i| \quad \text{s.t. } \boldsymbol{t} = \Phi \boldsymbol{w} \tag{II.1}$$

assuming the Jeffreys prior, or

$$\min_{w} \sum_{i=1}^{M} |w_i|^p \quad \text{s.t. } \boldsymbol{t} = \Phi \boldsymbol{w} \tag{II.2}$$

assuming a generalized Gaussian prior, for which the Laplacian is a special case (equivalent to assuming $p = 1$). With regard to globally minimizing solutions, the analysis is very simple. In the limit as $p \to 0$, it can be shown that (II.1) is a special case of (II.2) both in terms of the resulting cost function and the associated update rules that result (see [78] and Appendix VI.H.1 for more discussion). Consequently, we only need consider (II.2) without loss of generality. As described in [51], there exists a p' sufficiently small such that, for all $0 < p < p'$, the global minimum of (II.2) will equal the maximally sparse solution to $\boldsymbol{t} = \Phi \boldsymbol{w}$. However, this 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. But the point here is that p need not equal zero exactly in order to be guaranteed that (II.2) produces $\boldsymbol{w}_0$ when globally optimized. Equivalently, we can say there will always be a p sufficiently small such that no structural errors are possible.

As $p \rightarrow 1$, there is increasingly less likelihood that the global minimum to (II.2) will be maximally sparse. While significant attention has been given to establishing equivalence conditions whereby the global $p = 1$ solution will in fact equal $\boldsymbol{w}_0$ [17, 18, 29, 35, 95], these conditions tend to be extremely restrictive, and therefore difficult to apply, in many practical situations. And so in general, structural errors can be frequent as shown empirically in Sections II.D.2, III.D, and VI.D.

In the neuroimaging applications with which we are concerned, most existing equivalence conditions only allow for trivial sparse recovery problems, if any, to be solved. Appendix II.F.2 contains a brief example of a BP equivalence condition and the associated difficulty applying it in the context of MEG/EEG source imaging. Chapter VII evaluates neuroimaging-specific issues in greater detail.

II.B.2 Analysis of Local Minima

When $p = 1$, it is well known that the resulting optimization problem is convex, whether $\lambda \rightarrow 0$ or not. Consequently, convergence to undesirable local solutions is not generally an issue.[1] When $p < 1$, things are decidedly different. Local minima pose

[1] It is possible, however, to have multiple globally minimizing solutions, all confined to the same basin of attraction, in certain nuanced situations.

a clear impediment to achieving globally optimal solutions, and therefore, quantifying the number and extent of such minima is important. While strictly deterministic results may be evasive in general, the issue can be addressed probabilistically. To facilitate this goal, we first present the following result (see Appendix II.F.1 for proof):

Lemma 1. If Φ satisfies the URP, then the set of BFS to $\boldsymbol{t} = \Phi\boldsymbol{w}$ equals the set of locally minimizing solutions to (II.2) assuming $p \in [0, 1)$.

Assuming the URP holds (as is the case almost surely for dictionaries formed from iid elements drawn from a continuous, bounded probability density), we can conclude from Lemma 1 that we need only determine how many BFS exist when counting the number of local minima for the case where $p < 1$.

The number of BFS, and therefore the number of local minima, is bounded between $\binom{M-1}{N} + 1$ and $\binom{M}{N}$; the exact number depends on $\boldsymbol{t}$ and Φ [34]. Given that usually $M \gg N$, even the lower bound will be huge. The exact number can be assessed more precisely in certain situations. For example, if we assume there exists only a single degenerate sparse solution with $D_0 < N$ nonzero elements, then this solution is by definition the maximally sparse solution $\boldsymbol{w}_0$. Under these circumstances, it is a simple matter to show that the total number of BFS, denoted $\mathcal{N}_{\text{BFS}}$, is given by $\binom{M}{N} - \binom{M-D_0}{N-D_0} + 1$. But in what situations is our assumption of a single degenerate BFS valid? The following Lemma addresses this question (see Appendix II.F.1 for the proof):

Lemma 2. Let $\Phi \in \mathbb{R}^{N\times M}, M > N$ be constructed such that it satisfies the URP. Additionally, let $\boldsymbol{t} \in \mathbb{R}^N$ satisfy $\boldsymbol{t} = \Phi\boldsymbol{w}_0$ for some $\boldsymbol{w}_0$ such that $\|\boldsymbol{w}_0)\|_0 \triangleq D_0 < N$,

with non-zero entries of $\boldsymbol{w}_0$ drawn independently and identically from a continuous, bounded density. Then there is almost surely no other solution $\boldsymbol{w} \neq \boldsymbol{w}_0$ such that $\boldsymbol{t} = \Phi\boldsymbol{w}$ and $\|\boldsymbol{w}\|_0 = D < N$.

Given that the conditions of Lemma 2 are satisfied, we may then conclude that,

$$\mathrm{P}\left[\mathcal{N}_{\text{BFS}} = \binom{M}{N} - \binom{M - D_0}{N - D_0} + 1\right] = 1. \tag{II.3}$$

So for an arbitrary initialization $\boldsymbol{w}$ and assuming $M > N$, we cannot guarantee (i.e., with probability one) that the FOCUSS algorithm (or any other descent method) will avoid converging to one of the $\mathcal{N}_{\text{BFS}} - 1$ suboptimal local minima (i.e., a convergence error per our previous definition), each with suboptimal diversity given by $\|\boldsymbol{w}\|_0 = N$. However, while the number of local minima is the same for all $p < 1$, the relative sizes of the basins of attraction for each corresponding local minima is not. As $p \to 1$, the basins of attraction favor solutions resembling the $p = 1$ case become much larger, thereby increasing the likelihood that a random initialization will produce the minimum ℓ_1-norm solution.

In the more general case where we would like to relax the restriction $\boldsymbol{t} = \Phi\boldsymbol{w}$ exactly, the analysis of global and local minima is decidedly more complex. However, it has been shown that the corresponding MAP estimation problem

$$\min_{\boldsymbol{w}} \|\boldsymbol{t} - \Phi\boldsymbol{w}\|_2^2 + \lambda \sum_{i=1}^{M} |w_i|^p \tag{II.4}$$

will have at most N nonzero entries at any local minimum for any $p \in [0, 1]$ and $\lambda \in (0, \infty)$ [79].

II.B.3 Discussion

The discussion in the preceding sections leads to a natural dichotomy between two strategy extremes for obtaining sparse solutions via MAP estimation. We can either, (a) Choose to keep a cost function whose global minimum produces the maximally sparse solution and then deal with the local minima that ensue. This implies we will incur no structural errors but may encounter frequent convergence errors. Or we can, (b) Substitute a convex surrogate measure in place of the troublesome ℓ_0 norm (i.e., the ℓ_1 norm which follows from a Laplacian prior) that leads to a more tractable optimization problem but whose global minimum often does not equal $\boldsymbol{w}_0$. This means there will be no convergence errors but potentially many structural errors.

The first case leads to what will we call *local sparsity maximization* (LSM) algorithms. We will use this label to refer to any descent algorithm that employs a cost function whose global and local minima can be achieved by solutions that globally and locally minimize (I.8) respectively. From the preceding analysis we know that MAP estimation (in the noiseless limit) using the EM algorithm qualifies when either a generalized Gaussian with p sufficiently small or a Jeffreys prior is chosen. While all of these methods are potentially very useful candidates for finding sparse solutions, their Achilles heel is that a combinatorial number of local minima exist.

One potential advantage of using an LSM algorithm is that poor solutions

can be thrown out and the algorithm reinitialized repeatedly until a suitable estimate is found. This method has been shown to be somewhat successful in [78] and benefits from the fact that many LSM methods converge rapidly, meaning that multiple runs are feasible. In contrast, if the minimum ℓ_1-norm solution is computed and found to be unacceptable, reinitialization is fruitless, since every starting point leads to the same solution. This is the price we must pay for incorporating an objective function whose global minimum need not coincide with the global minimum of (I.8), and so regardless of initialization option (b) may fail.

II.C Sparse Bayesian Learning

In an ideal setting, we would like to experience no convergence errors or structural errors such that we would be sure of always finding maximally sparse solutions. While this is generally not possible given the NP-hardness of the sparse recovery problem, perhaps there is a better way to manage the trade-off than is currently available using the MAP framework. In this section, we prove that SBL is also a LSM algorithm when $\lambda \rightarrow 0$, implying that it will never produce structural errors. This may come as somewhat of a surprise since the SBL objective function is seemingly unrelated to (I.8). We then show that it maintains provably fewer local minima and therefore, displays vastly fewer convergence errors than previous LSM algorithms. The net result is fewer total errors than any of the MAP procedures discussed above.

II.C.1 SBL Global Minima and Maximally Sparse Solutions

To qualify as an LSM, we must first show how global minima of the SBL cost function

$$\mathcal{L}(\boldsymbol{\gamma}; \lambda = \varepsilon) = \log \left| \varepsilon I + \Phi \Gamma \Phi^T \right| + \boldsymbol{t}^T \left(\varepsilon I + \Phi \Gamma \Phi^T \right)^{-1} \boldsymbol{t} \tag{II.5}$$

relate to maximally sparse solutions in the limit as ε approaches zero.[2] Since SBL operates in hyperparameter space, the connection is less transparent than in the MAP case, but no less viable. The following theorem quantifies this relationship.

Theorem 1. Let $\mathcal{W}_0$ denote the set of weight vectors that globally minimize (I.8) with Φ satisfying the URP. Furthermore, let $\mathcal{W}(\varepsilon)$ be defined as the set of weight vectors

$$\left\{ \boldsymbol{w}_{**} : \;\; \boldsymbol{w}_{**} = \Gamma_{**} \Phi^T \left(\varepsilon I + \Phi \Gamma_{**} \Phi^T \right)^{-1} \boldsymbol{t}, \;\; \boldsymbol{\gamma}_{**} = \arg\min_{\boldsymbol{\gamma}} \mathcal{L}(\boldsymbol{\gamma}; \lambda = \varepsilon) \right\}. \tag{II.6}$$

Then in the limit as $\varepsilon \to 0$, if $\boldsymbol{w} \in \mathcal{W}(\varepsilon)$, then $\boldsymbol{w} \in \mathcal{W}_0$.

The weight estimator used for $\boldsymbol{w}_{**}$ is just the posterior mean derived in Section I.D.2. A full proof of this result is available in Appendix II.F.3; however, we provide a brief sketch here. First, as shown in the next section, every local minimum of $\mathcal{L}(\boldsymbol{\gamma}; \lambda = \varepsilon)$ is achieved at a basic feasible solution $\boldsymbol{\gamma}_*$, i.e., a solution with N or fewer nonzero entries, regardless of ε. Therefore, in our search for the global minimum, we only need examine the space of basic feasible solutions. As we allow ε to become sufficiently small, we

[2] When convenient, we will use $\mathcal{L}(\boldsymbol{\gamma}; \lambda)$ to denote the SBL cost function when $\lambda > 0$ and reserve $\mathcal{L}(\boldsymbol{\gamma})$ for the specific case where $\lambda = 0$.

show that

$$\mathcal{L}(\boldsymbol{\gamma}_*; \lambda = \varepsilon) = (N - \|\boldsymbol{\gamma}_*\|_0) \log(\varepsilon) + \mathrm{O}(1) \tag{II.7}$$

at any such solution. This result is minimized when $\|\boldsymbol{\gamma}_*\|_0$ is as small as possible. A maximally sparse basic feasible solution, which we denote $\boldsymbol{\gamma}_{**}$, can only occur with nonzero elements aligned with the nonzero elements of some $\boldsymbol{w} \in \mathcal{W}_0$. In the limit as $\varepsilon \to 0$, $\boldsymbol{w}_{**}$ becomes feasible while maintaining the same sparsity profile as $\boldsymbol{\gamma}_{**}$, leading to the stated result.

This result demonstrates that the SBL framework can provide an effective proxy to direct ℓ_0-norm minimization. More importantly, in the next section we will show that the limiting SBL cost function, which we will henceforth denote

$$\mathcal{L}(\boldsymbol{\gamma}) \triangleq \lim_{\varepsilon \to 0} \mathcal{L}(\boldsymbol{\gamma}; \lambda = \varepsilon) = \log \left| \Phi \Gamma \Phi^T \right| + \boldsymbol{t}^T \left(\Phi \Gamma \Phi^T \right)^{-1} \boldsymbol{t}, \tag{II.8}$$

often maintains a much more attractive local minima profile than comparable MAP methods.

II.C.2 Analysis of Local Minima

Like the MAP approaches, we will now show that SBL local minima are achieved at BFS which, when combined with Theorem 1, ensures that SBL is also an LSM algorithm per our definition. But not all LSM algorithms are created equal. We will also show that the converse is not true: *Every BFS need not represent an SBL local minimum*. Necessary conditions are derived for local minima to occur leading to a

simple geometric example of how SBL can have many fewer than previous MAP-based LSM methods. This is a key factor in SBL's superior performance as demonstrated later in Section II.D.2. Additionally, we show that even when noise is present, SBL local minima produce solutions with at most N nonzero elements.

Local Minima and BFS

This section proves that all local minima of $\mathcal{L}(\boldsymbol{\gamma};\lambda)$ are achieved at solutions with at most N nonzero elements, regardless of the value of λ. This leads to a simple bound on the number of local minima and demonstrates that SBL is also an LSM. First we introduce two lemmas that are necessary for the main result.

Lemma 3. $\log|\Sigma_t|$ 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., [41]). Furthermore, based on Theorem 5.7 in [85], if a function $f(\cdot)$ is concave on $\mathbb{R}^m$ and $\mathcal{A}$ is an affine transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$, then $f(\mathcal{A}(\cdot))$ is also concave. Therefore, by defining

$$f(X) \triangleq \log|X| \tag{II.9}$$

$$\mathcal{A}(\Gamma) \triangleq \lambda I + \Phi\Gamma\Phi^T, \tag{II.10}$$

we achieve the desired result. ■

Lemma 4. 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

$$\boldsymbol{b} \triangleq \boldsymbol{t} - \lambda\boldsymbol{u} \tag{II.11}$$

$$A \triangleq \Phi\text{diag}(\Phi^T\boldsymbol{u}) \tag{II.12}$$

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

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

$$\boldsymbol{t} - \lambda\boldsymbol{u} = \Phi\text{diag}(\Phi^T\boldsymbol{u})\boldsymbol{\gamma}, \tag{II.13}$$

completing the proof. ■

Theorem 2. Every local minimum of $\mathcal{L}(\boldsymbol{\gamma};\lambda)$ is achieved at a solution with at most N nonzero elements, regardless of the value of λ.[3]

Proof: Consider the optimization problem

$$\begin{aligned} \min: \quad & f(\boldsymbol{\gamma}) \\ \text{subject to:} \quad & A\boldsymbol{\gamma} = \boldsymbol{b}, \ \ \boldsymbol{\gamma} \geq 0, \end{aligned} \tag{II.14}$$

[3]This does not rule out the possibility that another $\boldsymbol{\gamma}$ will also obtain the same local minimum, i.e., a given basin could potentially include multiple minimizing $\boldsymbol{\gamma}$ at the bottom if certain conditions are met. However, included in this local minimizing set, will be a solution with at most N nonzero elements.

where $\boldsymbol{b}$ and A are defined as in (II.11) and (II.12) and $f(\boldsymbol{\gamma}) = \log|\Sigma_t|$. From Lemma 4, 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 $\mathcal{L}(\boldsymbol{\gamma};\lambda)$ while holding the second term constant to some C). Also, Lemma 3 dictates that the objective function $f(\boldsymbol{\gamma})$ is concave.

Clearly, any local minimum of $\mathcal{L}(\boldsymbol{\gamma};\lambda)$, e.g., Γ_*, must also be a local minima of (II.14) with

$$C = \boldsymbol{t}^T\boldsymbol{u} = \boldsymbol{t}^T(\lambda I + \Phi\Gamma_*\Phi^T)^{-1}\boldsymbol{t}. \tag{II.15}$$

However, based on [55] Theorem 6.5.3, a minimum of (II.14) is achieved at an extreme point and additionally, Theorem 2.5 establishes the equivalence between extreme points and BFS. Consequently, all local minima must be achievable at BFS or a solution with $\|\boldsymbol{\gamma}\|_0 \leq N$. ■

Corollary 1. If $\lambda = 0$ and Φ satisfies the URP, then every local minimum of $\mathcal{L}(\boldsymbol{\gamma})$ is achieved at a solution $\boldsymbol{\gamma}_* = \boldsymbol{w}_*^2$ where $\boldsymbol{w}_*$ is some BFS to $\boldsymbol{t} = \Phi\boldsymbol{w}$ and the $(\cdot)^2$ operator is understood to apply elementwise.

Proof: Assume some local minima $\boldsymbol{\gamma}_*$ is obtained such that $\|\boldsymbol{\gamma}_*\|_0 = N$. Define $\widetilde{\boldsymbol{\gamma}}$ to be the vector of nonzero elements in $\boldsymbol{\gamma}_*$ and $\widetilde{\Phi}$ to be the associated dictionary columns. Let $\widetilde{\boldsymbol{w}} \triangleq \widetilde{\Phi}^{-1}\boldsymbol{t}$, and so $\widetilde{\boldsymbol{w}}$ represents the nonzero elements of some BFS. Then if $\boldsymbol{\gamma}_*$ is a local minimum to $\mathcal{L}(\boldsymbol{\gamma})$, $\widetilde{\boldsymbol{\gamma}}$ must (locally) minimize the constrained cost function

$$\mathcal{L}(\widetilde{\boldsymbol{\gamma}}) \;=\; \log\left|\widetilde{\Phi}\widetilde{\Gamma}\widetilde{\Phi}^T\right| + \boldsymbol{t}^T\left(\widetilde{\Phi}\widetilde{\Gamma}\widetilde{\Phi}^T\right)^{-1}\boldsymbol{t}$$

$$= \sum_{i=1}^{N} \left(\log \widetilde{\gamma}_i + \frac{\widetilde{w}_i^2}{\widetilde{\gamma}_i} \right). \tag{II.16}$$

The unique minimum is easily seen to be $\widetilde{\gamma}_i = \widetilde{w}_i^2$ for all i. Upon padding with the appropriate zeros, we obtain the desired result. Finally, the case where $\|\boldsymbol{\gamma}_*\|_0 < N$ can be handled in a similar manner by arbitrarily adding $N - \|\boldsymbol{\gamma}_*\|_0$ columns to $\widetilde{\Phi}$ and proceeding as before. ■

Corollary 2. If $\lambda = 0$ and Φ satisfies the URP, then

$$1 \;\leq\; \begin{array}{c}\text{\# of SBL} \\ \text{Local Minima}\end{array} \;\leq\; \begin{array}{c}\text{\# of BFS to} \\ \boldsymbol{t} = \Phi \boldsymbol{w}\end{array} \;\in\; \left[\binom{M-1}{N} + 1, \binom{M}{N} \right]. \tag{II.17}$$

Proof: From Corollary 1, there can be at most one SBL local minimum associated with each BFS to $\boldsymbol{t} = \Phi \boldsymbol{w}$. It follows then that the total number of SBL local minima[4] cannot be greater than the number of BFS. The lower bound is of course trivial. ■

Along with Theorem 1, these results imply that SBL is also an LSM, assuming a proper descent method is used to optimize its cost function. However, in the remainder of this chapter we will show that the actual number of SBL local minima can be well below the upper bound of (II.17) in many practical situations (unlike previous LSM methods). In fact, only in particularly nuanced situations will the upper bound be

[4]By local minima here, we implicitly mean separate basins (which could potentially have multiple minimizing solutions at the bottom). Of course the relative sizes of these basins, as well as the relative proximity of any initialization point to the basin containing the global minimum are very important factors.

reached. Later, Chapter III will demonstrate conditions whereby the lower bound can be reached.

Eliminating Local Minimum

Thus far, we have demonstrated that there is a close affiliation between the limiting SBL framework and the the minimization problem posed by (I.8). We have not, however, provided any concrete reason why SBL should be preferred over current MAP methods of finding sparse solutions. In fact, this preference is not established until we more carefully explore the problem of convergence to local minima.

As discussed in Section II.B, the problem with MAP-based LSM methods is that every BFS, of which there exist a combinatorial number, unavoidably becomes a local minimum. However, what if we could somehow eliminate all or most of these extrema? For example, consider the alternate objective function $f(\boldsymbol{w}) \triangleq \min(\|\boldsymbol{w}\|_0, N)$, leading to the optimization problem

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}), \qquad \text{s.t. } \boldsymbol{t} = \Phi\boldsymbol{w}. \tag{II.18}$$

While the global minimum remains unchanged, we observe that all local minima occurring at non-degenerate BFS have been effectively removed. In other words, at any solution $\boldsymbol{w}_*$ with N nonzero entries, we can always add a small component $\alpha\boldsymbol{w}' \in \text{Null}(\Phi)$ and maintain feasibility without increasing $f(\boldsymbol{w})$, since $f(\boldsymbol{w})$ can never be greater than N. Therefore, we are free to move from BFS to BFS without increasing $f(\boldsymbol{w})$. Also,

the rare degenerate BFS that do remain, even if suboptimal, are sparser by definition. Therefore, locally minimizing our new problem (II.18) is clearly superior to locally minimizing (I.8). But how can we implement such a minimization procedure, even approximately, in practice?

Although we cannot remove all non-degenerate local minima and still retain computational feasibility, it is possible to remove many of them, providing some measure of approximation to (II.18). This is effectively what is accomplished using SBL as will be demonstrated below. (Chapter V deals indirectly with this issue as well when we talk about non-factorial priors such as $\exp\left[-f(\boldsymbol{w})\right]$.) Specifically, we will derive necessary conditions required for a non-degenerate BFS to represent a local minimum to $\mathcal{L}(\boldsymbol{\gamma})$. We will then show that these conditions are frequently *not* satisfied, implying that there are potentially many fewer local minima. Thus, locally minimizing $\mathcal{L}(\boldsymbol{\gamma})$ comes closer to (locally) minimizing (II.18) than traditional MAP-based LSM methods, which in turn, is closer to globally minimizing $\|\boldsymbol{w}\|_0$.

Necessary Conditions for Local Minima

As previously stated, all local minima to $\mathcal{L}(\boldsymbol{\gamma})$ must occur at BFS $\boldsymbol{\gamma}_*$ (in the sense described in the previous section). Now suppose that we have found a (non-degenerate) $\boldsymbol{\gamma}_*$ with associated $\boldsymbol{w}_*$ computed using (I.30) and we would like to assess whether or not it is a local minimum to our SBL cost function. For convenience, let $\widetilde{\boldsymbol{w}}$ again denote the N nonzero elements of $\boldsymbol{w}_*$ and $\widetilde{\Phi}$ the associated columns of Φ (therefore, $\boldsymbol{t} = \widetilde{\Phi}\widetilde{\boldsymbol{w}}$ and $\widetilde{\boldsymbol{w}} = \widetilde{\Phi}^{-1}\boldsymbol{t}$). Intuitively, it would seem likely that if we are not

at a true local minimum, then there must exist at least one additional column of Φ not in $\widetilde{\Phi}$, e.g., some $\boldsymbol{x}$, that is somehow aligned with or in some respect similar to $\boldsymbol{t}$. Moreover, the significance of this potential alignment must be assessed relative to $\widetilde{\Phi}$. But how do we quantify this relationship for the purposes of analyzing local minima?

As it turns out, a useful metric for comparison is realized when we decompose $\boldsymbol{x}$ with respect to $\widetilde{\Phi}$, which forms a basis in $\mathbb{R}^N$ under the URP assumption. For example, we may form the decomposition $\boldsymbol{x} = \widetilde{\Phi}\widetilde{\boldsymbol{v}}$, where $\widetilde{\boldsymbol{v}}$ is a vector of weights analogous to $\widetilde{\boldsymbol{w}}$. As will be shown below, the similarity required between $\boldsymbol{x}$ and $\boldsymbol{t}$ (needed for establishing the existence of a local minimum) may then be realized by comparing the respective weights $\widetilde{\boldsymbol{v}}$ and $\widetilde{\boldsymbol{w}}$. In more familiar terms, this is analogous to suggesting that similar signals have similar Fourier expansions. Loosely, we may expect that if $\widetilde{\boldsymbol{v}}$ is 'close enough' to $\widetilde{\boldsymbol{w}}$, then $\boldsymbol{x}$ is sufficiently close to $\boldsymbol{t}$ (relative to all other columns in $\widetilde{\Phi}$) such that we are not at a local minimum. We formalize this idea via the following theorem:

Theorem 3. Let Φ satisfy the URP and let $\boldsymbol{\gamma}_*$ represent a vector of hyperparameters with N and only N nonzero entries and associated basic feasible solution $\widetilde{\boldsymbol{w}} = \widetilde{\Phi}^{-1}\boldsymbol{t}$. Let $\mathcal{X}$ denote the set of $M - N$ columns of Φ not included in $\widetilde{\Phi}$ and $\mathcal{V}$ the set of weights given by $\left\{\widetilde{\boldsymbol{v}} : \widetilde{\boldsymbol{v}} = \widetilde{\Phi}^{-1}\boldsymbol{x}, \boldsymbol{x} \in \mathcal{X}\right\}$. Then $\boldsymbol{\gamma}_*$ is a local minimum of $\mathcal{L}(\boldsymbol{\gamma})$ only if

$$\sum_{i \neq j} \frac{\widetilde{v}_i \widetilde{v}_j}{\widetilde{w}_i \widetilde{w}_j} \leq 0 \qquad \forall \widetilde{\boldsymbol{v}} \in \mathcal{V}. \tag{II.19}$$

Proof: If $\boldsymbol{\gamma}_*$ truly represents a local minimum of our cost function, then the following

condition must hold for all $\boldsymbol{x} \in \mathcal{X}$:

$$\frac{\partial \mathcal{L}(\boldsymbol{\gamma}_*)}{\partial \gamma_x} \;\geq\; 0, \tag{II.20}$$

where γ_x denotes the hyperparameter corresponding to the basis vector $\boldsymbol{x}$. In words, we cannot reduce $\mathcal{L}(\boldsymbol{\gamma}_*)$ along a positive gradient because this would push γ_x below zero. Using the matrix inversion lemma, the determinant identity, and some algebraic manipulations, we arrive at the expression

$$\frac{\partial \mathcal{L}(\boldsymbol{\gamma}_*)}{\partial \gamma_x} = \frac{\boldsymbol{x}^T B \boldsymbol{x}}{1 + \gamma_x \boldsymbol{x}^T B \boldsymbol{x}} - \left(\frac{\boldsymbol{t}^T B \boldsymbol{x}}{1 + \gamma_x \boldsymbol{x}^T B \boldsymbol{x}} \right)^2, \tag{II.21}$$

where $B \triangleq (\widetilde{\Phi}\widetilde{\Gamma}\widetilde{\Phi}^T)^{-1}$. Since we have assumed that we are at a local minimum, it is straightforward to show that $\widetilde{\Gamma} = \text{diag}(\widetilde{\boldsymbol{w}})^2$ leading to the expression

$$B = \widetilde{\Phi}^{-T} \text{diag}(\widetilde{\boldsymbol{w}})^{-2} \widetilde{\Phi}^{-1}. \tag{II.22}$$

Substituting this expression into (II.21) and evaluating at the point $\gamma_x = 0$, the above gradient reduces to

$$\frac{\partial \mathcal{L}(\boldsymbol{\gamma}_*)}{\partial \gamma_x} = \widetilde{\boldsymbol{v}}^T \left(\text{diag}(\widetilde{\boldsymbol{w}}^{-1} \widetilde{\boldsymbol{w}}^{-T}) - \widetilde{\boldsymbol{w}}^{-1} \widetilde{\boldsymbol{w}}^{-T} \right) \widetilde{\boldsymbol{v}}, \tag{II.23}$$

where $\widetilde{\boldsymbol{w}}^{-1} \triangleq [\widetilde{w}_1^{-1}, \ldots, \widetilde{w}_N^{-1}]^T$. This leads directly to the stated theorem. ■

This theorem provides a useful picture of what is required for local minima to exist and more importantly, why many BFS are not local minima. Moreover, there are several convenient ways in which we can interpret this result to accommodate a more intuitive perspective.

A Simple Geometric Interpretation

In general terms, if the signs of each of the elements in a given $\widetilde{\boldsymbol{v}}$ match up with $\widetilde{\boldsymbol{w}}$, then the specified condition will be violated and we cannot be at a local minimum. We can illustrate this geometrically as follows.

To begin, we note that our cost function $\mathcal{L}(\boldsymbol{\gamma})$ is invariant with respect to reflections of any basis vectors about the origin, i.e., we can multiply any column of Φ by -1 and the cost function does not change. Returning to a candidate local minimum with associated $\widetilde{\Phi}$, we may therefore assume, without loss of generality, that $\widetilde{\Phi} \equiv \widetilde{\Phi}\text{diag}\left(\text{sgn}(\boldsymbol{w})\right)$, giving us the decomposition $\boldsymbol{t} = \widetilde{\Phi}\boldsymbol{w}$, $\boldsymbol{w} > 0$. Under this assumption, we see that $\boldsymbol{t}$ is located in the convex cone formed by the columns of $\widetilde{\Phi}$. We can infer that if any $\boldsymbol{x} \in \mathcal{X}$ (i.e., any column of Φ not in $\widetilde{\Phi}$) lies in this convex cone, then the associated coefficients $\widetilde{\boldsymbol{v}}$ must all be positive by definition (likewise, by a similar argument, any $\boldsymbol{x}$ in the convex cone of $-\widetilde{\Phi}$ leads to the same result). Consequently, Theorem 3 ensures that we are not at a local minimum. The simple 2D example shown in Figure II.1 helps to illustrate this point.

Alternatively, we can cast this geometric perspective in terms of relative cone sizes. For example, let $C_{\widetilde{\Phi}}$ represent the convex cone (and its reflection) formed by $\widetilde{\Phi}$.

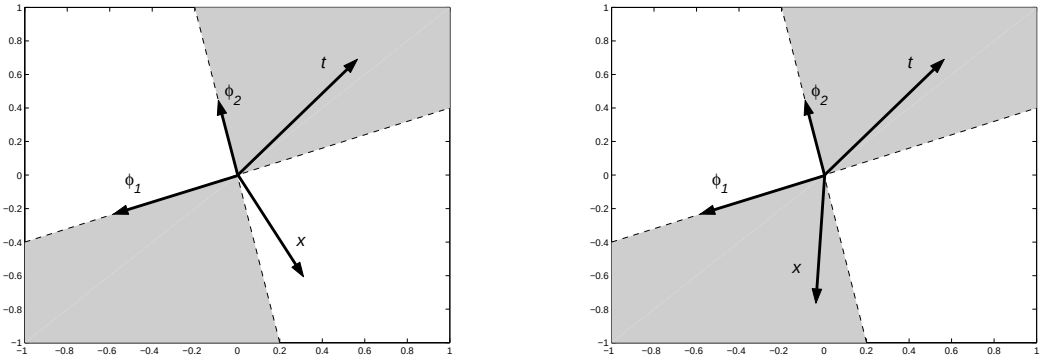

Figure II.1: 2D example with a 2×3 dictionary Φ (i.e., $N = 2$ and $M = 3$) and a basic feasible solution using the columns $\widetilde{\Phi} = [\boldsymbol{\phi}_1\ \boldsymbol{\phi}_2]$. *Left*: In this case, $\boldsymbol{x} = \boldsymbol{\phi}_3$ does not penetrate the convex cone containing $\boldsymbol{t}$, and we do not satisfy the conditions of Theorem 3. This configuration does represent a minimizing basic feasible solution. *Right*: Now $\boldsymbol{x}$ is in the cone and therefore, we know that we are not at a SBL local minimum; but this configuration *does* represent a local minimum to current LSM methods.

Then we are not at a local minimum to $L(\boldsymbol{\gamma})$ if there exists a second convex cone C formed from a subset of columns of Φ such that $\boldsymbol{t} \in C \subset C_{\widetilde{\Phi}}$, i.e., C is a tighter cone containing $\boldsymbol{t}$. In Figure II.1(*right*), we obtain a tighter cone by swapping $\boldsymbol{x}$ for ϕ_2.

While certainly useful, we must emphasize that in higher dimensions, these geometric conditions are *much* weaker than (II.19), e.g., if all $\boldsymbol{x}$ are *not* in the convex cone of $\widetilde{\Phi}$, we still may not be at a local minimum. In fact, to guarantee a local minimum, all $\boldsymbol{x}$ must be reasonably far from this cone as quantified by (II.19). Of course the ultimate reduction in local minima from the $\binom{M-1}{N} + 1$ to $\binom{M}{N}$ bounds is dependent on the distribution of basis vectors in $\boldsymbol{t}$-space. In general, it is difficult to quantify this reduction except in a few special cases. For example, in the special case where $\boldsymbol{t}$ is proportional to a single column of Φ, the number of BFS, and therefore the number of local minima to standard LSM algorithms, equals $\binom{M-1}{N} + 1$ (this assumes Φ satisfies the URP). In contrast, SBL is unimodal under these conditions, with the unique minimum

producing $\boldsymbol{w}_0$. The proof of this claim follows as a special case of Corollary 3 discussed in Section III.B, which addresses criteria whereby SBL has a unique minimum.[5] While detailed theoretical analysis is difficult in more general situations, the next section empirically demonstrates that the overall reduction in local minima can be very substantial, ultimately leading to a higher probability of recovering $\boldsymbol{w}_0$.

II.D Empirical Results

This section serves to empirically substantiate many of the theoretical ideas of this chapter.

II.D.1 Local Minima Comparison

To show that the potential reduction in local minima derived previously translates into concrete results, we conducted a simulation study using randomized dictionaries, with columns drawn uniformly from the surface of a unit hypersphere. Randomized dictionaries are of particular interest in signal processing and other disciplines [11, 17, 22, 80, 102]. Moreover, basis vectors from many real world measurements can often be modelled as random. In any event, randomized dictionaries capture a wide range of phenomena and therefore represent a viable benchmark for testing sparse recovery methods. At least we would not generally expect an algorithm to perform well with a random dictionary and poorly on everything else. Additionally, this particular mechanism for generating dictionaries is advocated in [18] as a useful benchmark and

[5] It can also be viewed as a special case of Theorem 9 presented in Section VI.E.1.

is exactly what is required in compressed sensing applications [20, 102]. Regardless, related experiments with other dictionary types, e.g., pairs of orthobases, yield similar results.

Our goal was to demonstrate that MAP-based LSM algorithms often converge to local minima that do not exist in the SBL cost function. To accomplish this, we repeated the following procedure for dictionaries of various sizes. First, we generate a random $N \times M$ dictionary Φ whose columns are each drawn uniformly from a unit sphere. Sparse weight vectors $\boldsymbol{w}_{\text{gen}}$ are randomly generated with $\|\boldsymbol{w}_{\text{gen}}\|_0 = 7$ (and uniformly distributed amplitudes on the nonzero components). The vector of target values is then computed as $\boldsymbol{t} = \Phi \boldsymbol{w}_{\text{gen}}$. The LSM algorithm is then presented with $\boldsymbol{t}$ and Φ and attempts to learn the minimum ℓ_0-norm solutions. The experiment is repeated a sufficient number of times such that we collect 1000 examples where the LSM algorithm converges to a local minimum. In all these cases, we check if the condition stipulated by Theorem 3 applies, allowing us to determine if the given solution is a local minimum to the SBL algorithm or not. The results are contained in Table II.1 for the FOCUSS LSM algorithm assuming $p \rightarrow 0$. We note that, the larger the overcompleteness ratio M/N, the larger the total number of LSM local minima (via the bounds presented earlier). However, there also appears to be a greater probability that SBL can avoid any given one.

In many cases where we found that SBL was not locally minimized, we initialized the SBL algorithm in this location and observed whether or not it converged to the optimal solution. In roughly 50% of these cases, *it escaped to find the maximally*

Table II.1: Given 1000 trials where FOCUSS (with $p \rightarrow 0$) has converged to a sub-optimal local minimum, we tabulate the percentage of times the local minimum is also a local minimum to SBL. M/N refers to the overcompleteness ratio of the dictionary used, with N fixed at 20.

M/N	1.3	1.6	2.0	2.4	3.0
SBL Local Minimum %	4.9%	4.0%	3.2%	2.3%	1.6%

sparse solution. The remaining times, it did escape in accordance with theory; however, it converged to another local minimum. In contrast, when we initialize other LSM algorithms at an SBL local minima, we always remain trapped as expected.

II.D.2 Performance Comparisons

While we have shown SBL has potentially many fewer local minima, we have not yet shown exactly to what degree this translates into improved performance finding $\boldsymbol{w}_{\text{gen}}$ over standard MAP methods, both LSM algorithms and BP. This section provides such a comparison. As before, we employ Monte Carlo simulations using randomized dictionaries for this purpose. We also use more structured dictionaries composed of pairs of orthobases as a further means of evaluation. For simplicity, noiseless tests were performed first, which facilitates direct comparisons because discrepancies in results cannot be attributed to poor selection of the trade-off parameter λ. Moreover, we have found that relative performance with the inclusion of noise remains essentially unchanged (see below). More extensive simulation results involving similar types of problems can be found in Sections III.D and VI.D.

Random Dictionaries

The experimental setup here is similar to that of the previous section, although here we used the fixed values $N = 20$ and $M = 40$ (related results using other dictionary sizes and sparsity levels can be found in Section III.D). 1000 independent trials were performed, each with a randomly generated dictionary and $\boldsymbol{w}_{\text{gen}}$. For every trial, three different MAP algorithms were compared with SBL; each method is presented with $\boldsymbol{t}$ and Φ and attempts to learn $\boldsymbol{w}_{\text{gen}}$, with a minimum ℓ_2-norm initialization being used in each case. An error is recorded whenever the estimate $\hat{\boldsymbol{w}}$ does not equal $\boldsymbol{w}_{\text{gen}}$.

Under the conditions set forth for the generation of Φ and $\boldsymbol{t}$, spark$(\Phi) = N+1$ and (I.7) is in force. Therefore, we can be sure that $\boldsymbol{w}_{\text{gen}} = \boldsymbol{w}_0$ with probability one. Additionally, we can be certain that when an algorithm fails to find $\boldsymbol{w}_{\text{gen}}$, it has not been lured astray by an even sparser representation.

The purpose of this study was to examine the relative frequency of cases where each algorithm failed to uncover the generating sparse weights. Also, 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 and yet has a lower cost function value than the generating solution (i.e., structural error). To this end, for each trial we compared cost function values at convergence with the 'ideal' cost function value at $\boldsymbol{w}_0$. Results are presented in the Table II.2.

Several items are worth noting with respect to these results. First, we see that with BP, only structural errors occur. This is expected since the BP cost function has no

Table II.2: 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 ($p = 0.001$)	**FOCUSS ($p = 0.9$)**	**Basis Pursuit ($p = 1.0$)**	**SBL**
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%**

local minima.[6] However, there is essentially a 22.3% chance that the minimum ℓ_1-norm solution of BP 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.

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. Also, these results were obtained using the fast (fixed-point) SBL update rules (see Section I.D.2). When the slower EM version of SBL is used, the error rate is reduced still further.

[6] And so these results hold whether we use interior-point or Simplex methods for BP.

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.,

$$\Phi = [\Theta, \Psi], \tag{II.24}$$

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 expansions among many others. The idea is that, while 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 [16, 17] 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 [17], 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, we cannot guarantee $\boldsymbol{w}_0$ is the sparsest possible solution unless $\|\boldsymbol{w}_0\|_0 < 5$. Nevertheless, we did find that in all cases where an algorithm failed, it converged to a solution $\boldsymbol{w}$ with $\|\boldsymbol{w}\|_0 = N > \|\boldsymbol{w}_0\|_0$. Results are

Table II.3: Comparative results from simulation study over 1000 independent trials using pairs of orthobases. Convergence errors and structural errors are defined as before.

	FOCUSS **($p = 0.001$)**	**FOCUSS** **($p = 0.9$)**	**Basis Pursuit** **($p = 1.0$)**	**SBL**
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%**

displayed in Table II.3.

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.

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 $20dB$. This relatively high number was selected so as 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 and so some criteria must be adopted for choosing the trade-off parameter λ, which balances

Table II.4: Comparative results from simulation study over 1000 independent trials using randomly generated dictionaries and the inclusion of additive white Gaussian noise to $20dB$.

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

sparsity with data fit. For all algorithms, λ was selected to roughly optimize the probability of recovering the generative weights per the criteria described below. In contrast, Section VI.D.1 contains related results plotted as λ is varied across a wide range of values.

Results are presented in Table II.4. 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 $\boldsymbol{w}_0$.

Again, SBL displays a much higher probability of recovering the generative basis vectors. These results corroborate our earlier theoretical and empirical findings suggesting the superiority of SBL in many situations.

II.D.3 Discussion

We have 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 application of SBL to sparse recovery problems. 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 using $p \ll 1$). 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. The next chapter will extend these ideas by examining criteria whereby *all* troublesome local minima are removed.

II.E Acknowledgements

This chapter is, in part, a reprint of material published in three articles: "Probabilistic Analysis for Basis Selection via ℓ_p Diversity Measures," *IEEE Int. Conf. Acoutics, Speech, and Signal Processing* (2004); "Sparse Bayesian Learning for Basis Selection," *IEEE Trans. Signal Processing* (2004); and "ℓ_0-Norm Minimization for Basis Selection," *Advances in Neural Information Processing Systems 17* (2005). In all cases I was the primary researcher and B.D. Rao supervised the research.

II.F Appendix

II.F.1 Proof of Lemmas 1 and 2

Proof of Lemma 1: That local minima are only achieved at BFS has been shown in [78]. We will now handle the converse. A vector $\boldsymbol{w}^*$ is a constrained local minimizer of $\|\boldsymbol{w}\|_p$ (s.t. $\boldsymbol{t} = \Phi\boldsymbol{w}$) if for every vector $\boldsymbol{w}' \in \text{Null}(\Phi)$, there is a $\delta > 0$ such that

$$\|\boldsymbol{w}^*\|_p < \|\boldsymbol{w}^* + \varepsilon\boldsymbol{w}'\|_p \quad \forall \varepsilon \in (0, \delta]. \tag{II.25}$$

We will now show that all BFS satisfy this condition. We first handle the case where $p > 0$ by defining $g(\varepsilon) \triangleq \|\boldsymbol{w}^* + \varepsilon\boldsymbol{w}'\|_p$ and then computing the gradient of $g(\varepsilon)$ at a feasible point in the neighborhood of $g(0) = \|\boldsymbol{w}^*\|_p$. We then note that at any feasible point $\boldsymbol{w} = \boldsymbol{w}^* + \varepsilon\boldsymbol{w}'$ we have

$$\begin{aligned} \frac{\partial g(\varepsilon)}{\partial \varepsilon} &= \left(\frac{\partial \|\boldsymbol{w}\|_p}{\partial \boldsymbol{w}}\right)^T \frac{\partial \boldsymbol{w}}{\partial \varepsilon} = \sum_{i=1}^{M} \frac{\partial \|\boldsymbol{w}\|_p}{\partial (w_i)} w_i' \\ &= \sum_{i=1}^{M} \text{sgn}(w_i^* + \varepsilon w_i') p |w_i^* + \varepsilon w_i'|^{p-1} w_i'. \end{aligned} \tag{II.26}$$

Since we have assumed we are at a BFS, we know that at least $M - N$ entries of $\boldsymbol{w}^*$ are equal to zero. Furthermore, let us assume without loss of generality that the first $M - N$ elements of $\boldsymbol{w}^*$ equal zero. This allows us to reexpress (II.26) as

$$\frac{\partial g(\varepsilon)}{\partial \varepsilon} = \sum_{i=1}^{M-N} \text{sgn}(w_i') p |\varepsilon w_i'|^{p-1} w_i' + \text{O}(1)$$

$$= \quad p \sum_{i=1}^{M-N} |w_i'|^p \left(\frac{1}{\varepsilon}\right)^{1-p} + \mathrm{O}(1). \tag{II.27}$$

At this point we observe that any $\boldsymbol{w}' \in \mathrm{Null}(\Phi)$ must have a nonzero element corresponding to a zero element in $\boldsymbol{w}^*$. This is a direct consequence of the URP assumption. Therefore, at least one $w_i', i \in [1, M-N]$ must be nonzero. As such, with ε sufficiently small, we can ignore terms of order $\mathrm{O}(1)$ (since $(1/\varepsilon)^{1-p}$ is unbounded for ε sufficiently small and $p < 1$) and we are left in (II.27) with a summation that must be positive.

Consequently, we see that for all $\varepsilon \in (0, \delta]$, $\partial g(\varepsilon)/\partial \varepsilon > 0$. By the Mean Value Theorem, this requires that $g(\delta) > g(0)$ or more explicitly,

$$\|\boldsymbol{w}^* + \delta \boldsymbol{w}'\|_p > \|\boldsymbol{w}^*\|_p. \tag{II.28}$$

Since w' is an arbitrary feasible vector, this completes the proof.

Finally, in the special case of $p = 0$, it is immediately apparent that all BFS must be local minima since $\|\boldsymbol{w}^*\|_0 < \|\boldsymbol{w}^* + \varepsilon \boldsymbol{w}'\|_0, \quad \forall \varepsilon \in (0, \delta]$. ■

Proof of Lemma 2: Let $\boldsymbol{w}_0'$ be a vector containing the amplitudes of the nonzero entries in $\boldsymbol{w}_0$ and Φ_1 the associated columns of Φ. Now let us suppose that there does exist a second solution $\boldsymbol{w}$ satisfying the conditions given above, with $\boldsymbol{w}'$ and Φ_2 being analogously defined. This implies that for some $\boldsymbol{w}'$,

$$\boldsymbol{t} = \Phi_1 \boldsymbol{w}_0' = \Phi_2 \boldsymbol{w}', \tag{II.29}$$

or equivalently, that $\boldsymbol{t}$ lies in both the span of Φ_1 and the span of Φ_2, both of which are full column rank by the URP assumption. Let us define this intersection as

$$\mathcal{A} = \text{span}(\Phi_1) \cap \text{span}(\Phi_2), \tag{II.30}$$

where we know by construction that

$$\begin{aligned} \dim(\mathcal{A}) &= \dim(\textbf{Null}([\Phi_1\ \Phi_2])) \\ &= \max(0, D + D_0 - N) \\ &< D_0. \end{aligned} \tag{II.31}$$

Note that the latter inequality follows since $D < N$ by assumption. At this point there are two possibilities. First, if $D \leq N - D_0$, then $\dim(\mathcal{A}) = 0$ and no solution $\boldsymbol{w}'$ (or $\boldsymbol{w}$ with $\|\boldsymbol{w}\|_0 = D$) can exist. Conversely, if $D > N - D_0$, the existence of a solution $\boldsymbol{w}'$ requires that $\Phi_1 \boldsymbol{w}_0'$ resides in a $(D + D_0 - N)$-dimensional subspace of the D_0-dimensional space Range(Φ_1). However, we know that with the entries of $\boldsymbol{w}_0'$ independently drawn from a continuous, bounded density function, $\Phi_1 \boldsymbol{w}_0$ also has a continuous and bounded density in Range(Φ_1) and the set $\{\boldsymbol{w}_0' : \Phi_1 \boldsymbol{w}_0' \in \mathcal{A}\}$ is of probability measure zero (see [83] for a discussion of probability measures). Therefore, we know that

$$\text{P}(\boldsymbol{w} \neq \boldsymbol{w}_0 \text{ exists s.t. } \|\boldsymbol{w}\|_0 < N) = \text{P}(\Phi_1 \boldsymbol{w}_0' \in \mathcal{A}) = 0, \tag{II.32}$$

completing the proof. ■

II.F.2 Performance Analysis with $p = 1$

As previously mentioned, when $p = 1$, a single minimum exists that may or may not correspond with the maximally sparse solution $\boldsymbol{w}_0$. However, in [17], a substantial result is derived that dictates when the minimum $\|\boldsymbol{w}\|_1$ solution is sufficient.

Theorem 4. (Equivalence Theorem [17]) Given an arbitrary dictionary Φ with columns $\boldsymbol{\phi}_i$ normalized such that $\boldsymbol{\phi}_i^T \boldsymbol{\phi}_i = 1, \forall i = 1, \ldots, M$, and given $G \triangleq \Phi^T \Phi$ and $\kappa \triangleq \max_{i \neq j} |G_{i,j}|$, if the sparsest representation of a signal by $\boldsymbol{t} = \Phi \boldsymbol{w}_0$ satisfies

$$\|\boldsymbol{w}_0\|_0 < 1/2(1 + 1/\kappa), \tag{II.33}$$

then the BP solution (which minimizes the $p = 1$ case) is guaranteed to equal $\boldsymbol{w}_0$.

This is a potentially powerful result since it specifies a computable condition by which the minimum $\|\boldsymbol{w}\|_1$-norm solution is guaranteed to produce $\boldsymbol{w}_0$. While elegant in theory, in practice it may be very difficult to apply. For example, the dictionaries required for MEG/EEG source localization (when suitably normalized as required by the theorem) typically have $\kappa \approx 1$. This implies that only sparse solutions with at most one nonzero element are guaranteed to be found. However, BP can still work effectively when the conditions of this theorem are violated.

As another consideration related to MEG/EEG, it may sometimes be desirable to leave the dictionary unnormalized or normalize with a metric other than the ℓ_2 norm, which can potentially mean that not even one nonzero element will be found with BP. In contrast, the SBL cost function is invariant to column normalization schemes, which only affect the implicit initialization used by the algorithm. More discussion of MEG/EEG related issues can be found in Chapter VII. Further analysis of general equivalence conditions and performance bounds for BP are derived in [18, 19, 96]. However, these are all more applicable to applications such as compressed sensing than to MEG/EEG source localization.

II.F.3 Proof of Theorem 1

In Section II.C.2 we demonstrate that every local minimum of $\mathcal{L}(\gamma;\lambda)$ is achieved at a $\boldsymbol{\gamma}$ with at most N nonzero entries. Consequently, we know that for any ε, the global minimum must occur at such a solution. At any such candidate local minimum, we only need be concerned with a subset of N basis vectors, denoted $\widetilde{\Phi}$ and the corresponding weights $\widetilde{\boldsymbol{w}}$ such that

$$\boldsymbol{t} = \Phi \boldsymbol{w} = \widetilde{\Phi}\widetilde{\boldsymbol{w}}. \tag{II.34}$$

Of course some of the elements of $\widetilde{\boldsymbol{w}}$ may be zero if we are at a degenerate basic feasible solution. Let us rewrite our cost function at this presumed local minimum (with ε treated

as a fixed parameter) as

$$\mathcal{L}(\widetilde{\Gamma};\lambda=\varepsilon)=\log|\varepsilon I+\widetilde{\Phi}\widetilde{\Gamma}\widetilde{\Phi}^T|+\boldsymbol{t}^T\left(\varepsilon I+\widetilde{\Phi}\widetilde{\Gamma}\widetilde{\Phi}^T\right)^{-1}\boldsymbol{t}, \tag{II.35}$$

where $\widetilde{\Gamma}$ is the diagonal matrix of hyperpriors associated with $\widetilde{\boldsymbol{w}}$. We now decompose each term as follows. First, we have

$$\begin{aligned}\boldsymbol{t}^T\left(\varepsilon I+\widetilde{\Phi}\widetilde{\Gamma}\widetilde{\Phi}^T\right)^{-1}\boldsymbol{t} &= \boldsymbol{t}^T\left[\widetilde{\Phi}\left(\varepsilon\widetilde{\Phi}^{-1}\widetilde{\Phi}^{-T}+\widetilde{\Gamma}\right)\widetilde{\Phi}^T\right]^{-1}\boldsymbol{t}\\ &= \widetilde{\boldsymbol{w}}^T\Phi^T\widetilde{\Phi}^{-T}\left[\varepsilon\left(\widetilde{\Phi}^T\widetilde{\Phi}\right)^{-1}+\widetilde{\Gamma}\right]^{-1}\widetilde{\Phi}^{-1}\widetilde{\Phi}\widetilde{\boldsymbol{w}}\\ &= \widetilde{\boldsymbol{w}}^T\left[\varepsilon\mathsf{S}+\widetilde{\Gamma}\right]^{-1}\widetilde{\boldsymbol{w}},\end{aligned} \tag{II.36}$$

where $\mathsf{S}\triangleq(\widetilde{\Phi}^T\widetilde{\Phi})^{-1}$ and the required inverse exists by the URP assumption. At this point we allow, without loss of generality, for $\widetilde{\boldsymbol{w}}$ to be expressed as $\widetilde{\boldsymbol{w}}=[\widetilde{\boldsymbol{w}}_{(D)};\ \mathbf{0}_{(N-D)}]$ where $\widetilde{\boldsymbol{w}}_{(D)}$ is a vector containing the $D\leq N$ nonzero entries in $\widetilde{\boldsymbol{w}}$ and $\mathbf{0}_{(N-D)}$ is a vector of $N-D$ zeros.

Defining $A\triangleq\varepsilon\mathsf{S}+\widetilde{\Gamma}$ and we can partition A as

$$A=\left[A^{11}A^{12};A^{21}A^{22}\right], \tag{II.37}$$

where A^{11} is a $D\times D$ block, A^{22} is $(N-D)\times(N-D)$, and so on.

Using the expression for the inverse of a partitioned matrix, we can expand (II.36) as

$$\begin{aligned}
\widetilde{\boldsymbol{w}}^T A^{-1} \widetilde{\boldsymbol{w}} &= \widetilde{\boldsymbol{w}}_{(D)}^T \left(A^{-1}\right)^{11} \widetilde{\boldsymbol{w}}_{(D)}^T \\
&= \widetilde{\boldsymbol{w}}_{(D)}^T \left[A^{11} - A^{12}\left(A^{22}\right)^{-1} A^{21}\right]^{-1} \widetilde{\boldsymbol{w}}_{(D)} \\
&= \widetilde{\boldsymbol{w}}_{(D)}^T \left[\varepsilon \mathsf{S}^{11} + \widetilde{\Gamma}^{11} - \varepsilon \mathsf{S}^{12}\left(\mathsf{S}^{22} + \widetilde{\Gamma}^{22}\right)^{-1} \mathsf{S}^{21} \varepsilon\right]^{-1} \widetilde{\boldsymbol{w}}_{(D)} \\
&= \widetilde{\boldsymbol{w}}_{(D)}^T \left[\varepsilon \Psi^{11} + \widetilde{\Gamma}^{11}\right]^{-1} \widetilde{\boldsymbol{w}}_{(D)}, \qquad \text{(II.38)}
\end{aligned}$$

where we have defined

$$\Psi^{11} \triangleq \mathsf{S}^{11} - \mathsf{S}^{12}\left(\mathsf{S}^{22} + \frac{\widetilde{\Gamma}^{22}}{\varepsilon}\right)^{-1} \mathsf{S}^{21}. \qquad \text{(II.39)}$$

Also, we observe that because S represents a non-degenerate (i.e., full-rank) covariance matrix, Ψ^{11} is full rank for all $\varepsilon \geq 0$ and all $\widetilde{\Gamma}^{22} \geq 0$.

We now turn to the second term in our cost function using

$$\begin{aligned}
\log|\varepsilon I_N + \widetilde{\Phi}\widetilde{\Gamma}\widetilde{\Phi}^T| &= \log|\widetilde{\Phi}||\varepsilon\widetilde{\Phi}^{-1}\widetilde{\Phi}^{-T} + \widetilde{\Gamma}||\widetilde{\Phi}^T| \\
&\equiv \log|\varepsilon \mathsf{S} + \widetilde{\Gamma}| \\
&= \log|\varepsilon \mathsf{S}^{11} + \widetilde{\Gamma}^{11}| + \log|\varepsilon \mathsf{S}^{22} + \widetilde{\Gamma}^{22} - \varepsilon \mathsf{S}^{21}\left(\mathsf{S}^{11} + \widetilde{\Gamma}^{11}\right)^{-1} \mathsf{S}^{12}\varepsilon| \\
&= \log\left|\varepsilon \mathsf{S}^{11} + \widetilde{\Gamma}^{11}\right| + \log\left|\varepsilon \Psi^{22} + \widetilde{\Gamma}^{22}\right|, \qquad \text{(II.40)}
\end{aligned}$$

where Ψ^{22} is defined in an analogous fashion as Ψ^{11} and likewise, is full-rank for all ε and $\widetilde{\Gamma}^{11}$. Combining terms, we have established that at an arbitrary local minimum, our

cost function is given by

$$L(\Gamma,\varepsilon)=\log\left|\varepsilon\mathsf{S}^{11}+\widetilde{\Gamma}^{11}\right|+\log\left|\varepsilon\Psi^{22}+\widetilde{\Gamma}^{22}\right|+\widetilde{\boldsymbol{w}}_{(D)}^{T}\left[\varepsilon\Psi^{11}+\widetilde{\Gamma}^{11}\right]^{-1}\widetilde{\boldsymbol{w}}_{(D)}. \quad \text{(II.41)}$$

At this point, we are poised to demonstrate two properties that hold for all $\varepsilon\in(0,\delta]$, where δ is sufficiently small yet greater than zero:

Lemma 5. There exists a constant $C>\delta$ such that $\gamma_i>C$ for all $i\in\{1,\ldots,D\}$ (i.e., the diagonal elements in $\widetilde{\Gamma}^{11}$ are all greater than C).

Proof: We observe that, for C sufficiently small (yet greater than δ), $i\in\{1,\ldots,D\}$, and $\gamma_i\leq C$, an upper bound for the gradient of (II.41) with respect to γ_i is given by

$$\frac{\partial L(\Gamma;\varepsilon<\delta)}{\partial\gamma_i}\leq \mathrm{O}\left(\frac{1}{C}\right)+\mathrm{O}\left(1\right)-\mathrm{O}\left(\frac{1}{C^2}\right)=-\mathrm{O}\left(\frac{1}{C^2}\right). \quad \text{(II.42)}$$

Since this gradient is necessarily negative for all $\gamma_i\leq C$, by the Mean Value Theorem, our local minimum must have γ_i greater than C. ■

Lemma 6. For all $i\in\{D+1,\ldots,N\}$, $\gamma_i=0$ (i.e., the diagonal elements in $\widetilde{\Gamma}^{22}$ are all equal to zero).

Proof: First, we observe that the minimum of $\mathcal{L}(\gamma;\lambda=\varepsilon)$, excluding the second term, is given by

$$\min_{\widetilde{\Gamma}^{11},\widetilde{\Gamma}^{22}}\quad \log\left|\varepsilon\mathsf{S}^{11}+\widetilde{\Gamma}^{11}\right|+\widetilde{\boldsymbol{w}}_{(D)}^{T}\left[\varepsilon\Psi^{11}+\widetilde{\Gamma}^{11}\right]^{-1}\widetilde{\boldsymbol{w}}_{(D)} \;=\; \mathrm{O}\left(1\right), \quad \text{(II.43)}$$

regardless of the value of $\widetilde{\Gamma}^{22}$. In contrast,

$$\min_{\widetilde{\Gamma}^{11},\widetilde{\Gamma}^{22}} \quad \log\left|\Psi^{22}+\widetilde{\Gamma}^{22}\right| \;=\; \log \mathrm{O}\left(\varepsilon^{N-D}\right). \tag{II.44}$$

with the minimum occurring with $\gamma_i \leq \mathrm{O}(\varepsilon)$ for all $i \in \{D+1,\ldots,N\}$, regardless of $\widetilde{\Gamma}^{11}$. But how do we know if these γ_i actually go to zero? If we compute the gradient of our cost function with respect to these γ_i, we obtain,

$$\frac{\partial L(\Gamma,\varepsilon)}{\partial \gamma_i} \leq \mathrm{O}\left(\frac{1}{\gamma_i}\right) - \mathrm{O}\left(1\right) = \mathrm{O}\left(\frac{1}{\gamma_i}\right). \tag{II.45}$$

This result is positive for all $\gamma_i \leq \mathrm{O}(\varepsilon)$ and therefore, at the local minimum, all must go to zero. ■

In conclusion, we can achieve an overall minimum of order $(N-D)\log\varepsilon$ with $\widetilde{\Gamma}^{11} > 0$ and $\widetilde{\Gamma}^{22} = 0$. Or more explicitly, at each local minimum, $\|\boldsymbol{\gamma}\|_0 = D$. Of course, the global minimum occurs when D is smallest. Therefore, at a solution achieving the global minimum $\boldsymbol{\gamma}_{**}$, we must have $\|\boldsymbol{\gamma}_{**}\|_0 = \|\boldsymbol{w}_0\|_0$ for all $\varepsilon \in (0,\delta]$.

Without loss of generality, by Lemma 6 we can then write

$$\begin{aligned}
\boldsymbol{w}_{**} &= \Gamma_{**}\Phi^T\left(\varepsilon I + \Phi\Gamma_{**}\Phi^T\right)^{-1}\boldsymbol{t} \\
&= \left(\Phi^T\Phi + \varepsilon\Gamma_{**}^{-1}\right)^{-1}\Phi^T\boldsymbol{t} \\
&= \left[\left(\Phi_{(\boldsymbol{w}_0)}^T\Phi_{(\boldsymbol{w}_0)} + \varepsilon\left(\Gamma_{**}^{11}\right)^{-1}\right)^{-1}\Phi_{(\boldsymbol{w}_0)}^T\boldsymbol{t};\;\; \mathbf{0}_{(N-\|\boldsymbol{w}_0\|_0)}\right],
\end{aligned} \tag{II.46}$$

where $\Phi_{(\boldsymbol{w}_0)}$ denotes the columns of Φ associated with nonzero elements in $\boldsymbol{w}_0$. Using Lemma 5, as $\varepsilon \to 0$ we have

$$\begin{aligned} \boldsymbol{w}_{**} &= \left[\left(\Phi_{(\boldsymbol{w}_0)}^T \Phi_{(\boldsymbol{w}_0)} \right)^{-1} \Phi_{(\boldsymbol{w}_0)}^T \boldsymbol{t};\ \ \mathbf{0}_{(N-\|\boldsymbol{w}_0\|_0)} \right] \\ &= \boldsymbol{w}_0, \end{aligned} \tag{II.47}$$

completing the proof.

Chapter III

Comparing the Effects of Different Weight Distributions

Previously, we have argued that the sparse Bayesian learning (SBL) framework is particularly well-suited for finding maximally sparse representations, showing that it has far fewer local minima than many other Bayesian-inspired strategies. In this Chapter, we provide further evidence for this claim by proving a restricted equivalence condition, based on the distribution of the nonzero generating model weights, whereby the SBL cost function is unimodal and will achieve the maximally sparse representation at the global minimum (in a noiseless setting). We also prove that if these nonzero weights are drawn from an approximate Jeffreys prior, then with probability approaching one, our equivalence condition is satisfied. Finally, we motivate the worst-case scenario for SBL and demonstrate that it is still better than the most widely used sparse representation algorithms. These include Basis Pursuit (BP), which is based on

a convex relaxation of the ℓ_0 (quasi)-norm, and Orthogonal Matching Pursuit (OMP), a simple greedy strategy that iteratively selects basis vectors most aligned with the current residual.

III.A Introduction

To review, the canonical form of the noiseless (exact) sparse recovery problem is given by,

$$\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_0, \qquad \text{s.t. } \boldsymbol{t} = \Phi \boldsymbol{w}, \tag{III.1}$$

where $\Phi \in \mathbb{R}^{N \times M}$ is a matrix whose columns represent an overcomplete or redundant basis (i.e., rank$(\Phi) = N$ and $M > N$), $\boldsymbol{w} \in \mathbb{R}^M$ is the vector of weights to be learned, and $\boldsymbol{t}$ is the signal vector. The cost function being minimized represents the ℓ_0 (quasi)-norm of $\boldsymbol{w}$ (i.e., a count of the nonzero elements in $\boldsymbol{w}$). The solution to (III.1) has been considered in [17, 18, 29, 35, 95].

Unfortunately, an exhaustive search for the optimal representation requires the solution of up to $\binom{M}{N}$ linear systems of size $N \times N$, a prohibitively expensive procedure for even modest values of M and N. Consequently, in practical situations there is a need for approximate procedures that efficiently solve (III.1) with high probability. To date, the two most widely used choices are Basis Pursuit (BP) [17] and Orthogonal Matching Pursuit (OMP) [95]. (Note that the later can be viewed as an LSM algorithm.) BP is based on a convex relaxation of the ℓ_0 norm, i.e., replacing $\|\boldsymbol{w}\|_0$ with $\|\boldsymbol{w}\|_1$, which leads to an attractive, unimodal optimization problem that can be read-

ily solved via linear programming or the EM algorithm. In contrast, OMP is a greedy strategy that iteratively selects the basis vector most aligned with the current signal residual. At each step, a new approximant is formed by projecting $\boldsymbol{t}$ onto the range of all the selected dictionary atoms. In the previous chapter, we demonstrated that SBL can be also be used to effectively solve (III.1) while maintaining several significant advantages over other, Bayesian-inspired strategies for finding sparse solutions (notably MAP-based LSM methods [27, 34]).

To compare BP, OMP, and SBL, we would ultimately like to know in what situations a particular algorithm is likely to find the maximally sparse solution. A variety of results stipulate rigorous conditions whereby BP and OMP are guaranteed to solve (III.1) [17, 29, 95]. All of these conditions depend explicitly on the number of nonzero elements contained in the optimal solution. Essentially, if this number is less than some Φ-dependent constant κ, the BP/OMP solution is proven to be equivalent to the minimum ℓ_0-norm solution. Unfortunately however, κ turns out to be restrictively small and, for a fixed redundancy ratio M/N, grows very slowly as N becomes large [18]. But in practice, both approaches still perform well even when these equivalence conditions have been grossly violated. To address this issue, a much looser bound has recently been produced for BP, dependent only on M/N. This bound holds for “most" dictionaries in the limit as N becomes large [18], where “most" is with respect to dictionaries composed of columns drawn uniformly from the surface of an N-dimensional unit hypersphere. For example, with $M/N = 2$, it is argued that BP is capable of resolving sparse solutions with roughly $0.3N$ nonzero elements with probability approaching

one as $N \rightarrow \infty$.

Turning to SBL, we have neither a convenient convex cost function (as with BP) nor a simple, transparent update rule (as with OMP); however, we can nonetheless come up with an alternative type of equivalence result that is neither unequivocally stronger nor weaker than those existing results for BP and OMP. This condition is dependent on the relative magnitudes of the nonzero elements embedded in optimal solutions to (III.1). Additionally, we can leverage these ideas to motivate which sparse solutions are the most difficult to find. Later, we provide empirical evidence that SBL, even in this worst-case scenario, can still outperform both BP and OMP.

III.B Equivalence Conditions for SBL

From Section I.D.2, we know that SBL can be used to solve (III.1) using the update rules

$$\begin{aligned} \boldsymbol{\gamma}_{(\text{new})} &= \text{diag}\left(\hat{\boldsymbol{w}}_{(\text{old})} \hat{\boldsymbol{w}}_{(\text{old})}^T + \left[I - \Gamma_{(\text{old})}^{1/2} \left(\Phi \Gamma_{(\text{old})}^{1/2} \right)^{\dagger} \Phi \right] \Gamma_{(\text{old})} \right) \\ \hat{\boldsymbol{w}}_{(\text{new})} &= \Gamma_{(\text{new})}^{1/2} \left(\Phi \Gamma_{(\text{new})}^{1/2} \right)^{\dagger} \boldsymbol{t}, \end{aligned} \tag{III.2}$$

where $(\cdot)^{\dagger}$ denotes the Moore-Penrose pseudo-inverse and $\Gamma \triangleq \text{diag}(\boldsymbol{\gamma})$. Based on EM convergence properties, these rules are guaranteed to reduce the SBL cost function at each iteration until a fixed point is reached. For the remainder of this chapter, $\boldsymbol{w}^{\text{SBL}}$ will

refer to a stable fixed point of (III.2), and therefore also a local minimum of

$$\mathcal{L}(\boldsymbol{\gamma}) = \log \left|\Phi\Gamma\Phi^T\right| + \boldsymbol{t}^T \left(\Phi\Gamma\Phi^T\right)^{-1} \boldsymbol{t}. \tag{III.3}$$

In this section, we establish conditions whereby any $\boldsymbol{w}^{\text{SBL}}$ will necessarily minimize (III.1). To state these results, we require some notation. First, we formally define a dictionary $\Phi = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_M]$ as a set of M unit ℓ_2-norm vectors (atoms) in $\mathbb{R}^N$, with $M > N$ and $\text{rank}(\Phi) = N$. We say that a dictionary satisfies the unique representation property (URP) if every subset of N atoms forms a basis in $\mathbb{R}^N$. We define $w_{(i)}$ as the i-th largest weight magnitude and $\widetilde{\boldsymbol{w}}$ as the $\|\boldsymbol{w}\|_0$-dimensional vector containing all the nonzero weight magnitudes of $\boldsymbol{w}$. The set of optimal solutions to (III.1) is $\mathcal{W}^*$ with cardinality $|\mathcal{W}^*|$. The *diversity* (or anti-sparsity) of each $\boldsymbol{w}^* \in \mathcal{W}^*$ is defined as $D^* \triangleq \|\boldsymbol{w}^*\|_0$.

Theorem 5. For a fixed dictionary Φ that satisfies the URP, there exists a set of $M-1$ scaling constants $\nu_i \in (0, 1]$ (i.e., strictly greater than zero) such that, for any $\boldsymbol{t} = \Phi\boldsymbol{w}'$ generated with

$$w'_{(i+1)} \leq \nu_i w'_{(i)} \qquad i = 1, \ldots, M-1, \tag{III.4}$$

any $\boldsymbol{w}^{\text{SBL}}$ must satisfy $\|\boldsymbol{w}^{\text{SBL}}\|_0 = \min(N, \|\boldsymbol{w}'\|_0)$ and $\boldsymbol{w}^{\text{SBL}} \in \mathcal{W}^*$.

The proof has been deferred to Appendix III.G.1. The basic idea is that, as the magnitude differences between weights increase, at any given scale, the covariance Σ_t embedded in the SBL cost function is dominated by a single dictionary atom such that problematic local minimum are removed. The unique, global minimum in turn achieves the stated

result.[1] The most interesting case occurs when $\|\boldsymbol{w}'\|_0 < N$, leading to the following:

Corollary 3. Given the additional restriction $\|\boldsymbol{w}'\|_0 < N$, then $\boldsymbol{w}^{\text{SBL}} = \boldsymbol{w}' \in \mathcal{W}^*$ and $|\mathcal{W}^*| = 1$, i.e., SBL has a unique stable fixed point that equals the unique, maximally sparse representation of the signal $\boldsymbol{t}$.

See the Appendix III.G.1 for the proof. These results are restrictive in the sense that the dictionary dependent constants ν_i significantly confine the class of signals $\boldsymbol{t}$ that we may represent. Moreover, we have not provided any convenient means of computing what the different scaling constants might be. But we have nonetheless solidified the notion that SBL is most capable of recovering weights of different scales (and it must still find all D^* nonzero weights no matter how small some of them may be). Additionally, we have specified conditions whereby we will find the unique $\boldsymbol{w}^*$ even when the diversity is as large as $D^* = N - 1$. The tighter BP/OMP bound from [17, 29, 95] scales as $\text{O}\left(N^{-1/2}\right)$, although this latter bound is much more general in that it is independent of the magnitudes of the nonzero weights.

In contrast, neither BP or OMP satisfy a comparable result; in both cases, simple 3D counter examples suffice to illustrate this point.[2] We begin with OMP. Assume the following:

[1] Because we have effectively shown that the SBL cost function must be unimodal, etc., any proven descent method could likely be applied in place of (III.2) to achieve the same result.

[2] While these examples might seem slightly nuanced, the situations being illustrated can occur frequently in practice and the requisite column normalization introduces some complexity.

$$\boldsymbol{w}^* = \begin{bmatrix} 1 \\ \epsilon \\ 0 \\ 0 \end{bmatrix} \quad \Phi = \begin{bmatrix} 0 & \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{1.01}} \\ 0 & 0 & 1 & \frac{0.1}{\sqrt{1.01}} \\ 1 & \frac{1}{\sqrt{2}} & 0 & 0 \end{bmatrix} \quad \boldsymbol{t} = \Phi \boldsymbol{w}^* = \begin{bmatrix} \frac{\epsilon}{\sqrt{2}} \\ 0 \\ 1 + \frac{\epsilon}{\sqrt{2}} \end{bmatrix}, \tag{III.5}$$

where Φ satisfies the URP and has columns $\boldsymbol{\phi}_i$ of unit ℓ_2 norm. Given any $\epsilon \in (0, 1)$, we will now show that OMP will necessarily fail to find $\boldsymbol{w}^*$. Provided $\epsilon < 1$, at the first iteration OMP will select $\boldsymbol{\phi}_1$, which solves $\max_i |\boldsymbol{t}^T \boldsymbol{\phi}_i|$, leaving the residual vector

$$\boldsymbol{r}_1 = \left(I - \boldsymbol{\phi}_1 \boldsymbol{\phi}_1^T\right) \boldsymbol{t} = [\ \epsilon/\sqrt{2} \quad 0 \quad 0\ ]^T. \tag{III.6}$$

Next, $\boldsymbol{\phi}_4$ will be chosen since it has the largest value in the top position, thus solving $\max_i |\boldsymbol{r}_1^T \boldsymbol{\phi}_i|$. The residual is then updated to become

$$\boldsymbol{r}_2 = \left(I - [\ \boldsymbol{\phi}_1 \quad \boldsymbol{\phi}_4\ ][\ \boldsymbol{\phi}_1 \quad \boldsymbol{\phi}_4\ ]^T\right) \boldsymbol{t} = \frac{\epsilon}{101\sqrt{2}} [\ 1 \quad -10 \quad 0\ ]^T. \tag{III.7}$$

From the remaining two columns, $\boldsymbol{r}_2$ is most highly correlated with $\boldsymbol{\phi}_3$. Once $\boldsymbol{\phi}_3$ is selected, we obtain zero residual error, yet we did not find $\boldsymbol{w}^*$, which involves only $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$. So for all $\epsilon \in (0, 1)$, the algorithm fails. As such, there can be no fixed constant $\nu > 0$ such that if $w^*_{(2)} \equiv \epsilon \leq \nu w^*_{(1)} \equiv \nu$, we are guaranteed to obtain $\boldsymbol{w}^*$ (unlike with SBL).

We now give an analogous example for BP, where we present a feasible solu-

tion with smaller ℓ_1 norm than the maximally sparse solution. Given

$$\boldsymbol{w}^* = \begin{bmatrix} 1 \\ \epsilon \\ 0 \\ 0 \end{bmatrix} \quad \Phi = \begin{bmatrix} 0 & 1 & \frac{0.1}{\sqrt{1.02}} & \frac{0.1}{\sqrt{1.02}} \\ 0 & 0 & \frac{-0.1}{\sqrt{1.02}} & \frac{0.1}{\sqrt{1.02}} \\ 1 & 0 & \frac{1}{\sqrt{1.02}} & \frac{1}{\sqrt{1.02}} \end{bmatrix} \quad \boldsymbol{t} = \Phi \boldsymbol{w}^* = \begin{bmatrix} \epsilon \\ 0 \\ 1 \end{bmatrix}, \tag{III.8}$$

it is clear that $\|\boldsymbol{w}^*\|_1 = 1 + \epsilon$. However, for all $\epsilon \in (0, 0.1)$, if we form a feasible solution using only $\boldsymbol{\phi}_1$, $\boldsymbol{\phi}_3$, and $\boldsymbol{\phi}_4$, we obtain the alternate solution

$$\boldsymbol{w} = \begin{bmatrix} (1 - 10\epsilon) & 0 & 5\sqrt{1.02}\epsilon & 5\sqrt{1.02}\epsilon \end{bmatrix}^T \tag{III.9}$$

with $\|\boldsymbol{w}\|_1 \approx 1 + 0.1\epsilon$. Since this has a smaller ℓ_1 norm for all ϵ in the specified range, BP will necessarily fail and so again, we cannot reproduce the result for a similar reason as before.

At this point, it remains unclear what probability distributions are likely to produce weights that satisfy the conditions of Theorem 5. It turns out that the Jeffreys prior, given by $p(x) \propto 1/x$, is appropriate for this task. This distribution has the unique property that the probability mass assigned to any given scaling is equal. More explicitly, for any $s \geq 1$,

$$P\left(x \in \left[s^i, s^{i+1}\right]\right) \propto \log(s) \quad \forall i \in \mathbb{Z}. \tag{III.10}$$

For example, the probability that x is between 1 and 10 equals the probability that it lies between 10 and 100 or between 0.01 and 0.1. Because this is an improper density, we

define an approximate Jeffreys prior with range parameter $a \in (0, 1]$. Specifically, we say that $x \sim J(a)$ if

$$p(x) = \frac{-1}{2\log(a)x} \quad \text{for } x \in [a, 1/a]. \tag{III.11}$$

With this definition in mind, we present the following result.

Theorem 6. For a fixed Φ that satisfies the URP, let $\boldsymbol{t}$ be generated by $\boldsymbol{t} = \Phi\boldsymbol{w}'$, where $\boldsymbol{w}'$ has magnitudes drawn iid from $J(a)$. Then as a approaches zero, the probability that we obtain a $\boldsymbol{w}'$ such that the conditions of Theorem 5 are satisfied approaches unity.

Appendix III.G.2 contains the proof. However, on a conceptual level this result can be understood by considering the distribution of order statistics. For example, given M samples from a uniform distribution between zero and some θ, with probability approaching one, the distance between the k-th and $(k+1)$-th order statistic can be made arbitrarily large as θ moves towards infinity. Likewise, with the $J(a)$ distribution, the relative scaling between order statistics can be increased without bound as a decreases towards zero, leading to the stated result.

Corollary 4. Assume that $D' < N$ randomly selected elements of $\boldsymbol{w}'$ are set to zero. Then as a approaches zero, the probability that we satisfy the conditions of Corollary 3 approaches unity.

In conclusion, we have shown that a simple, (approximate) noninformative Jeffreys prior leads to sparse inverse problems that are optimally solved via SBL with high probability. Interestingly, it is this same Jeffreys prior that forms the implicit weight

prior of SBL (see [94], Section 5.1). However, it is worth mentioning that other Jeffreys prior-based techniques, e.g., direct minimization of $p(\boldsymbol{w}) = \prod_i \frac{1}{|w_i|}$ subject to $\boldsymbol{t} = \Phi\boldsymbol{w}$, do *not* provide any SBL-like guarantees. Although several algorithms do exist that can perform such a minimization task (e.g., [27, 34]), they perform poorly with respect to (III.1) because of convergence to local minimum as shown in Chapter II. This is especially true if the weights are highly scaled, and no nontrivial equivalence results are known to exist for these procedures.

III.C Worst-Case Scenario

If the best-case scenario occurs when the nonzero weights are all of very different scales, it seems reasonable that the most difficult sparse inverse problem may involve weights of the same or even identical scale, e.g., $\widetilde{w}_1^* = \widetilde{w}_2^* = \ldots \widetilde{w}_{D^*}^*$. This notion can be formalized somewhat by considering the $\widetilde{\boldsymbol{w}}^*$ distribution that is furthest from the Jeffreys prior. First, we note that both the SBL cost function and update rules are independent of the overall scaling of the generating weights, meaning $\alpha\widetilde{\boldsymbol{w}}^*$ is functionally equivalent to $\widetilde{\boldsymbol{w}}^*$ provided α is nonzero. This invariance must be taken into account in our analysis. Therefore, we assume the weights are rescaled such that $\sum_i \widetilde{w}_i^* = 1$. Given this restriction, we will find the distribution of weight magnitudes that is most different from the Jeffreys prior.

Using the standard procedure for changing the parameterization of a probabil-

ity density, the joint density of the constrained variables can be computed simply as

$$p(\widetilde{w}_1^*, \ldots, \widetilde{w}_{D^*}^*) \propto \frac{1}{\prod_{i=1}^{D^*} \widetilde{w}_i^*} \quad \text{for} \quad \sum_{i=1}^{D^*} \widetilde{w}_i^* = 1, \;\; \widetilde{w}_i^* \geq 0, \forall i. \tag{III.12}$$

From this expression, it is easily shown that $\widetilde{w}_1^* = \widetilde{w}_2^* = \ldots = \widetilde{w}_{D^*}^*$ achieves the global minimum. Consequently, equal weights are the absolute *least* likely to occur from the Jeffreys prior. Hence, we may argue that the distribution that assigns $\widetilde{w}_i^* = 1/D^*$ with probability one is furthest from the constrained Jeffreys prior.

Nevertheless, because of the complexity of the SBL framework, it is difficult to prove axiomatically that $\widetilde{\boldsymbol{w}}^* \sim \mathbf{1}$ is overall the most problematic distribution with respect to sparse recovery. We can however provide additional motivation for why we should expect it to be unwieldy. As proven in Section II.C.1, the global minimum of the SBL cost function is guaranteed to produce some $\boldsymbol{w}^* \in \mathcal{W}^*$. This minimum is achieved with the hyperparameters $\gamma_i^* = (w_i^*)^2$, $\forall i$. We can think of this solution as forming a collapsed, or degenerate covariance $\Sigma_t^* = \Phi \Gamma^* \Phi^T$ that occupies a proper D^*-dimensional subspace of N-dimensional signal space. Moreover, this subspace must necessarily contain the signal vector $\boldsymbol{t}$. Essentially, Σ_t^* proscribes infinite density to $\boldsymbol{t}$, leading to the globally minimizing solution.

Now consider an alternative covariance $\Sigma_t^\diamond$ that, although still full rank, is nonetheless ill-conditioned (flattened), containing $\boldsymbol{t}$ within its high density region. Furthermore, assume that $\Sigma_t^\diamond$ is not well aligned with the subspace formed by Σ_t^*. The mixture of two flattened, yet misaligned covariances naturally leads to a more volu-

minous (less dense) form as measured by the determinant $|\alpha\Sigma_t^* + \beta\Sigma_t^\diamond|$. Thus, as we transition from $\Sigma_t^\diamond$ to Σ_t^*, we necessarily reduce the density at $\boldsymbol{t}$, thereby increasing the cost function $\mathcal{L}(\boldsymbol{\gamma})$. So if SBL converges to $\Sigma_t^\diamond$ it has fallen into a local minimum.

So the question remains, what values of $\widetilde{\boldsymbol{w}}^*$ are likely to create the most situations where this type of local minima occurs? The issue is resolved when we again consider the D^*-dimensional subspace determined by Σ_t^*. The volume of the covariance *within* this subspace is given by $\left|\widetilde{\Phi}\widetilde{\Gamma}^*\widetilde{\Phi}^{*T}\right|$, where $\widetilde{\Phi}^*$ and $\widetilde{\Gamma}^*$ are the basis vectors and hyperparameters associated with $\widetilde{\boldsymbol{w}}^*$. The larger this volume, the higher the probability that other basis vectors will be suitably positioned so as to both (i), contain $\boldsymbol{t}$ within the high density portion and (ii), maintain a sufficient component that is misaligned with the optimal covariance.

The maximum volume of $\left|\widetilde{\Phi}^*\widetilde{\Gamma}^*\widetilde{\Phi}^{*T}\right|$ under the constraints $\sum_i \widetilde{w}_i^* = 1$ and $\widetilde{\gamma}_i^* = (\widetilde{w}^*)_i^2$ occurs with $\widetilde{\gamma}_i^* = 1/(D^*)^2$, i.e., all the $\widetilde{w}_i^*$ are equal. Consequently, geometric considerations support the notion that deviance from the Jeffreys prior leads to difficulty recovering $\boldsymbol{w}^*$. Moreover, empirical analysis (not shown) of the relationship between volume and local minimum avoidance provide further corroboration of this hypothesis.

III.D Empirical Comparisons

The central purpose of this section is to present empirical evidence that supports our theoretical analysis and illustrates the improved performance afforded by SBL. As previously mentioned, others have established deterministic equivalence conditions,

dependent on D^*, whereby BP and OMP are guaranteed to find the unique $\boldsymbol{w}^*$. Unfortunately, the relevant theorems are of little value in assessing practical differences between algorithms. This is because, in the cases we have tested where BP/OMP equivalence is provably known to hold (e.g., via results in [17, 29, 95]), SBL always converges to $\boldsymbol{w}^*$ as well.

As such, we will focuss our attention on the insights provided by Sections III.B and III.C as well as probabilistic comparisons with [18]. Given a fixed distribution for the nonzero elements of $\boldsymbol{w}^*$, we will assess which algorithm is best (at least empirically) for most dictionaries relative to a uniform measure on the unit sphere as discussed.

To this effect, a number of monte-carlo simulations were conducted, each consisting of the following: First, a random, overcomplete $N \times M$ dictionary Φ is created whose columns are each drawn uniformly from the surface of an N-dimensional hypersphere. Next, sparse weight vectors $\boldsymbol{w}^*$ are randomly generated with D^* nonzero entries. Nonzero amplitudes $\widetilde{\boldsymbol{w}}^*$ are drawn iid from an experiment-dependent distribution. Response values are then computed as $\boldsymbol{t} = \Phi \boldsymbol{w}^*$. Each algorithm is presented with $\boldsymbol{t}$ and Φ and attempts to estimate $\boldsymbol{w}^*$. In all cases, we ran 1000 independent trials and compared the number of times each algorithm failed to recover $\boldsymbol{w}^*$. Under the specified conditions for the generation of Φ and $\boldsymbol{t}$, all other feasible solutions $\boldsymbol{w}$ almost surely have a diversity greater than D^*, so our synthetically generated $\boldsymbol{w}^*$ must be maximally sparse. Moreover, Φ will almost surely satisfy the URP.

With regard to particulars, there are essentially four variables with which to experiment: (i) the distribution of $\widetilde{\boldsymbol{w}}^*$, (ii) the diversity D^*, (iii) N, and (iv) M. In Figure

III.1, we display results from an array of testing conditions. In each *row* of the figure, $\widetilde{w}_i^*$ is drawn iid from a fixed distribution for all i; the first row uses $\widetilde{w}_i^* = 1$, the second has $\widetilde{w}_i^* \sim J(a = 0.001)$, and the third uses $\widetilde{w}_i^* \sim N(0, 1)$, i.e., a unit Gaussian. In all cases, the signs of the nonzero weights are irrelevant due to the randomness inherent in the basis vectors.

The *columns* of Figure III.1 are organized as follows: The first column is based on the values $N = 50$, $D^* = 16$, while M is varied from N to $5N$, testing the effects of an increasing level of dictionary redundancy, M/N. The second fixes $N = 50$ and $M = 100$ while D^* is varied from 10 to 30, exploring the ability of each algorithm to resolve an increasing number of nonzero weights. Finally, the third column fixes $M/N = 2$ and $D^*/N \approx 0.3$ while N, M, and D^* are increased proportionally. This demonstrates how performance scales with larger problem sizes.

The first row of plots essentially represents the worst-case scenario for SBL per our previous analysis, and yet performance is still consistently better than both BP and OMP. In contrast, the second row of plots approximates the best-case performance for SBL, where we see that SBL is almost infallible. The handful of failure events that do occur are because a is not sufficiently small and therefore, $J(a)$ was not sufficiently close to a true Jeffreys prior to achieve perfect equivalence (see center plot). Although OMP also does well here, the parameter a can generally never be adjusted such that OMP always succeeds. Finally, the last row of plots, based on Gaussian distributed weight amplitudes, reflects a balance between these two extremes. Nonetheless, SBL still holds a substantial advantage.

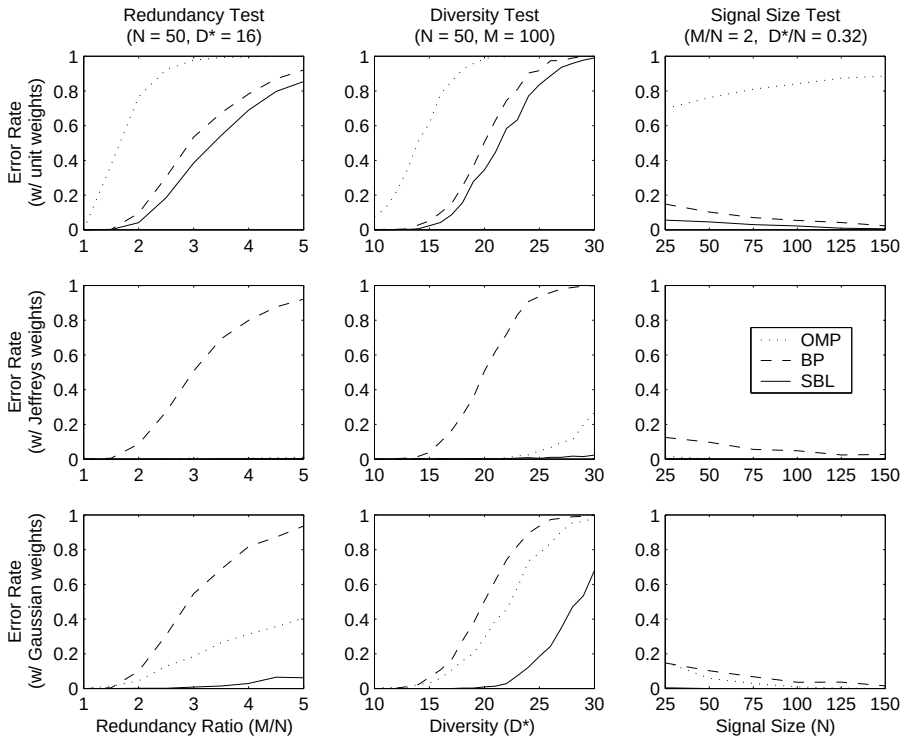

Figure III.1: Empirical results comparing the probability that OMP, BP, and SBL fail to find $\boldsymbol{w}^*$ under various testing conditions. Each data point is based on 1000 independent trials. The distribution of the nonzero weight amplitudes is labeled on the far left for each row, while the values for N, M, and D^* are included on the top of each column. Independent variables are labeled along the bottom of the figure.

In general, we observe that SBL is capable of handling more redundant dictionaries (column one) and resolving a larger number of nonzero weights (column two). Also, column three illustrates that both BP and SBL are able to resolve a number of weights that grows linearly in the signal dimension ($\approx 0.3N$), consistent with the analysis in [18] (which applies only to BP). In contrast, OMP performance begins to degrade in some cases (see the upper right plot), a potential limitation of this approach. Of course additional study is necessary to fully compare the relative performance of these methods on large-scale problems.

Finally, by comparing row one, two and three, we observe that the performance of BP is independent of the weight distribution (consistent with results from [59]), with performance slightly below the worst-case SBL performance. Like SBL, OMP results are highly dependent on the distribution; however, as the weight distribution approaches unity, performance is unsatisfactory. In summary, while the relative proficiency between OMP and BP is contingent on experimental particulars, SBL is uniformly superior in the cases we have tested (including examples not shown, e.g., results with other dictionary types).

III.E Conclusions

In this chapter, we have related the ability to find maximally sparse solutions using SBL to the particular distribution of amplitudes that compose the nonzero elements. At first glance, it may seem reasonable that the most difficult sparse inverse problems occur when some of the nonzero weights are extremely small, making them difficult to estimate. Perhaps surprisingly then, we have shown that the exact opposite is true with SBL: The more diverse the weight magnitudes, the better the chances we have of learning the optimal solution. In contrast, unit weights offer the most challenging task for SBL. Nonetheless, even in this worst-case scenario, we have shown that SBL outperforms the current state-of-the-art; the overall assumption here being that, if worst-case performance is superior, then it is likely to perform better in a variety of situations.

Unlike SBL, it has been shown that under very mild conditions BP performance is provably independent of the nonzero weight magnitudes [59]. While this inde-

pendence is compelling, it also ensures that there is no alternative distribution that can improve BP performance beyond what we have shown above, which is inferior to the worse-case SBL results in the situations we have tested thus far.

For a *fixed* dictionary and diversity D^*, successful recovery of unit weights does not absolutely guarantee that any alternative weighting scheme will necessarily be recovered as well. However, a weaker result does appear to be feasible: For fixed values of N, M, and D^*, if the success rate recovering unity weights approaches one for most dictionaries, where most is defined as in Section III.A, then the success rate recovering weights of any other distribution (assuming they are distributed independently of the dictionary) will also approach one. While a formal proof of this conjecture is beyond the scope of this paper, it seems to be a very reasonable result that is certainly born out by experimental evidence, geometric considerations, and the arguments presented in Section III.C. Nonetheless, this remains a fruitful area for further inquiry.

III.F Acknowledgements

This chapter, in part, is a reprint of material published as “Comparing the Effects of Different Weight Distributions on Finding Sparse Representations," *Advances in Neural Information Processing Systems 18* (2006). I was the primary researcher and B.D. Rao supervised the research.

III.G Appendix

III.G.1 Proof of Theorem 5 and Corollary 3

In the most general setting, the constants ν_i may all be unique, leading to the most flexible set of allowable weighting schemes. However, for simplicity we will assume that $\nu_1 = \nu_2 = \ldots = \nu_{M-1} = \epsilon$, where ϵ is a constant in the interval $(0, 1]$. The extension to the more general setting is straightforward.

Every local minimum of $\mathcal{L}(\boldsymbol{\gamma})$, the SBL cost function, is achieved at a basic feasible solution (BFS) (see Section II.C.2). By this we mean that every local minimum is achieved at a solution with $\gamma_i = w_i^2$ (for all i) such that,

$$\boldsymbol{w} \in \mathcal{W}^{\text{BFS}} \triangleq \{\boldsymbol{w} : \boldsymbol{t} = \Phi\boldsymbol{w}, \|\boldsymbol{w}\|_0 \leq N\}. \tag{III.13}$$

Interestingly, the converse is not true; that is, every element of $\mathcal{W}^{\text{BFS}}$ need not correspond with a minimum to $\mathcal{L}(\boldsymbol{\gamma})$. In fact, for a suitable selection of ϵ, we will show that this reduced set of minima naturally leads to a proof of Theorem 5.

We begin with a set of weights $\boldsymbol{w}'$ such that $w'_{(i+1)} \leq \epsilon w'_{(i)}$ and $\|\boldsymbol{w}'\|_0 \triangleq D' \in \{1, \ldots, M\}$. For convenience, we will also assume that $w'_{(i)} = |w'_i|$ for all i. In other words, the first element of $\boldsymbol{w}'$ has the largest magnitude, the second element has the second largest magnitude, and so on. To avoid any loss of generality, we incorporate an $M \times M$ permutation matrix P into our generative model, giving us the signal $\boldsymbol{t} = \Phi P\boldsymbol{w}' = \Phi'\boldsymbol{w}'$. Because $\Phi' \triangleq \Phi P$ is nothing more than Φ with reordered columns, it will necessarily satisfy the URP for all P given that Φ does.

We now examine the properties of an arbitrary BFS with nonzero weights $\widetilde{\boldsymbol{w}}$ and associated dictionary atoms $\widetilde{\Phi}$, i.e., $\boldsymbol{t} = \widetilde{\Phi}\widetilde{\boldsymbol{w}}$. There exist two possibilities for a candidate BFS:

- *Case I*: The columns of Φ' associated with the largest (in magnitude) $\min(N, D')$ nonzero weights of $\boldsymbol{w}'$ are contained in $\widetilde{\Phi}$. By virtue of the URP, no other basis vectors will be present even if $D' < N$, so we may conclude that $\widetilde{\Phi} = [\boldsymbol{\phi}'_1, \boldsymbol{\phi}'_2, \ldots, \boldsymbol{\phi}'_{\min(N,D')}]$.
- *Case II*: At least one of the columns associated with the largest $\min(N, D')$ weights is missing from $\widetilde{\Phi}$.

Given this distinction, we would like to determine when a candidate BFS, particularly a Case II BFS of which there are many, is a local minimum.

To accomplish this, we let $r \in \{1, \ldots, \min(N, D')\}$ denote the index of the of the largest weight for which the respective dictionary atom, $\boldsymbol{\phi}'_r$ is *not* in $\widetilde{\Phi}$. Therefore, by assumption the first $r-1$ columns of $\widetilde{\Phi}$ equal $[\boldsymbol{\phi}'_1, \boldsymbol{\phi}'_2, \ldots, \boldsymbol{\phi}'_{r-1}]$. The remaining columns of $\widetilde{\Phi}$ are arbitrary (provided of course that $\boldsymbol{\phi}'_r$ is not included). This allows us to express any Case II BFS as

$$\widetilde{\boldsymbol{w}} = \widetilde{\Phi}^{-1}\boldsymbol{t} = \widetilde{\Phi}^{-1}\Phi'\boldsymbol{w}' = \sum_{k=1}^{r-1} w'_k \mathbf{e}_k + \widetilde{\Phi}^{-1} \sum_{k=r}^{D'} w'_k \boldsymbol{\phi}'_k, \tag{III.14}$$

where $\mathbf{e}_k$ is a zero vector with a one in the k-th element and we have assumed that every Case II BFS utilizes exactly N columns of Φ' (i.e., $\widetilde{\Phi}$ is $N \times N$ and therefore invertible via the URP). This assumption is not restrictive provided we allow for zero-padding of

BFS with less than N nonzero weights (this implies that some elements of $\widetilde{\boldsymbol{w}}$ will be equal to zero if we have to add dummy columns to $\widetilde{\Phi}$).

Since SBL is invariant to the overall scaling of model weights, without loss of generality we will assume that $w'_r = 1$. We also define $\widetilde{\boldsymbol{v}} \triangleq \widetilde{\Phi}^{-1}\boldsymbol{\phi}'_r$, giving us

$$\widetilde{\boldsymbol{w}} = \widetilde{\Phi}^{-1}\boldsymbol{t} = \sum_{k=1}^{r-1} w'_k \mathbf{e}_k + \widetilde{\boldsymbol{v}} + \widetilde{\Phi}^{-1} \sum_{k=r+1}^{D'} w'_k \boldsymbol{\phi}'_k. \tag{III.15}$$

By virtue of the stipulated ϵ-dependent weight scaling, we know that

$$\widetilde{\Phi}^{-1} \sum_{k=r+1}^{D'} w'_k \boldsymbol{\phi}'_k = \sum_{k=r+1}^{D'} \mathrm{O}\left(\epsilon^{k-r}\right) \cdot \mathbf{1}_N \;\; = \;\; \mathrm{O}\left(\epsilon\right) \cdot \mathbf{1}_N, \tag{III.16}$$

where we have used the notation $f(x) = \mathrm{O}(g(x))$ to indicate that $|f(x)| < C_1|g(x)|$ for all $x < C_2$, with C_1 and C_2 constants independent of x. Also, $\mathrm{O}(x) \cdot \mathbf{1}_N$ refers to an N-dimensional vector with all elements of order $\mathrm{O}(x)$. Combining (III.15) and (III.16), we can express the i-th element of $\widetilde{\boldsymbol{w}}$ as

$$\widetilde{w}_i = w'_i \mathbf{I}\left[i < r\right] + \widetilde{v}_i + \mathrm{O}\left(\epsilon\right). \tag{III.17}$$

Provided ϵ is chosen suitably small, we can ensure that all $\widetilde{w}_i$ are necessarily nonzero (so in fact no zero-padding is ever necessary). When $i \geq r$, this occurs because all elements of $\widetilde{\boldsymbol{v}}$ must be strictly nonzero or we violate the URP assumption. For the $i < r$ case, a sufficiently small ϵ means that the w'_i term (which is of order $\mathrm{O}\left(1/\epsilon^{r-i}\right)$ by virtue of (III.4)) will dominate, leading to a nonzero $\widetilde{w}_i$. This allows us to apply Theorem 3, from

which we can conclude that a candidate BFS with N nonzero weights will represent a local minimum only if

$$\sum_{i \neq j} \frac{\widetilde{v}_i \widetilde{v}_j}{\widetilde{w}_i \widetilde{w}_j} \leq 0. \tag{III.18}$$

Substituting (III.17) into this criterion, we obtain

$$\sum_{i \neq j} \left(\frac{\widetilde{v}_i}{w_i' \mathbf{I}\left[i < r\right] + \widetilde{v}_i + \mathrm{O}(\epsilon)} \right) \left(\frac{\widetilde{v}_j}{w_j' \mathbf{I}\left[j < r\right] + \widetilde{v}_j + \mathrm{O}(\epsilon)} \right) = \tag{III.19}$$

$$\mathrm{O}(\epsilon) \quad + \quad \sum_{i \neq j;\, i,j \geq r} \left(\frac{\widetilde{v}_i}{\widetilde{v}_i + \mathrm{O}(\epsilon)} \right) \left(\frac{\widetilde{v}_j}{\widetilde{v}_j + \mathrm{O}(\epsilon)} \right).$$

If $D' < N$, then $r < N$ by definition and so there will always be at least one set of indices i and j that satisfy the above summation constraints. This then implies that

$$\sum_{i \neq j} \frac{\widetilde{v}_i \widetilde{v}_j}{\widetilde{w}_i \widetilde{w}_j} \approx \sum_{i \neq j;\, i,j \geq r} 1 \quad > \quad 0, \tag{III.20}$$

since each $\widetilde{v}_i$ is a nonzero constant independent of ϵ. So we have violated a necessary condition for the existence of a local minimum.

In contrast, If $D' \geq N$, then it will always be possible to choose $r = N$ such that there are no allowable terms that satisfy the index constraints, meaning that this Case II BFS could potentially satisfy (III.18) and therefore be a local minimum with N nonzero elements.

In summary, we have shown two properties of an arbitrary Case II BFS, provided that ϵ is small enough: We have shown that it will have exactly N nonzero elements and that it will not represent a local minimum to SBL if $D' < N$. The exact

value of this ϵ will depend on the particular BFS and permutation matrix P; however, we can simply choose the smallest ϵ across all possibilities. From these two properties, it follows that $D^* = \min(N, D')$, meaning that the maximally sparse solution can only have diversity less than D' if $D' > N$.

These results are sufficient to complete the proof of Theorem 5 as follows. Any stable fixed point of the SBL update rules (III.2) must necessarily correspond with a local minima to $\mathcal{L}(\boldsymbol{\gamma})$ and a particular BFS. If $D' \geq N$, then $D^* = N$ and so *any* BFS is a member of $\mathcal{W}^*$ (although all BFS need not be local minima of SBL). In contrast, if $D' < N$ then no Case II BFS are local minima. The unique minimum (and therefore stable fixed point) that remains is the Case I BFS which satisfies $D' = D^*$. This completes the proof. Also, because the maximally sparse solution is necessarily unique when $D' < N$, Corollary 3 naturally follows.

It is important to stress that these results specify *sufficient* conditions for finding maximally sparse representations via SBL, but these conditions are by no means *necessary*, and SBL performs quite well even when the weights are not highly scaled. This is desirable from a practical standpoint, especially since it is not generally feasible to determine the value of ϵ for an arbitrary dictionary of reasonable size. Moreover, even if ϵ where easily computable it will typically be prohibitively small in all but the most simple cases.

III.G.2 Proof of Theorem 6 and Corollary 4

Again, for convenience will assume that $\nu_1 = \nu_2 = \ldots = \nu_{M-1} = \epsilon$; extension to the more general case is straightforward. Theorem 5 is predicated on the existence of a sufficiently small $\epsilon > 0$ such that $\boldsymbol{w}^{\text{SBL}}$ is guaranteed to be in the set $\mathcal{W}^*$. The actual value of this ϵ is dependent on Φ. However, we will show that the $J(a)$ distribution is capable of producing weights that satisfy $w'_{(i+1)} \leq \epsilon w'_{(i)}$ with high probability no matter how small ϵ may be. Thus, we can fulfill the conditions of Theorem 5 with probability approaching one for any Φ.

The distribution of the ordered weight magnitudes is given by

$$p\left(w'_{(1)}, \ldots, w'_{(M)}\right) = \frac{M!}{(-2\log a)^M} \prod_{i=1}^{M} \frac{1}{w'_{(i)}} \quad \text{for} \quad a \leq w'_{(M)} \leq \ldots \leq w'_{(1)} \leq \frac{1}{a}. \tag{III.21}$$

However, we would like to calculate the probability mass contained within the restricted weight space

$$\Omega(\epsilon) \triangleq \{\boldsymbol{w}' : a \leq w'_{(M)} \leq \epsilon w'_{(M-1)} \leq \ldots \epsilon w'_{(1)} \leq 1/a\} \tag{III.22}$$

for an arbitrary ϵ. This is readily computed via the integral

$$\begin{aligned} P(\boldsymbol{w}' \in \Omega(\epsilon)) &= \int_{\Omega(\epsilon)} p(\boldsymbol{w}') d\boldsymbol{w}' \\ &= \int_a^{\epsilon w'_{(M-1)}} \cdots \int_a^{\epsilon w'_{(1)}} \int_a^{\frac{1}{a}} p\left(w'_{(1)}, \ldots, w'_{(M)}\right) dw'_{(M)} \cdots dw'_{(2)} dw'_{(1)} \\ &= \left(\frac{2\log a + (M-1)\log\epsilon}{2\log a}\right)^M - \mathrm{O}\left((\log a)^{-2}\right), \end{aligned} \tag{III.23}$$

where $\mathrm{O}(\cdot)$ is defined as before. For any fixed $\epsilon \in (0, 1]$, as $a \to 0$, the righthand term can be ignored while the lefthand term converges to one, giving us

$$\lim_{a \to 0} P(\boldsymbol{w}' \in \Omega(\epsilon)) = 1. \tag{III.24}$$

Therefore, as a becomes small, the probability that we satisfy the conditions of Theorem 5 approaches unity.

The proof of Corollary 4 follows directly from the arguments presented above.

Chapter IV

Perspectives on Sparse Bayesian Learning

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 SBL framework and MAP paradigms such as FOCUSS and BP. As a step in this direction, this chapter will demonstrate that SBL can be recast using duality theory, where the hyperparameters $\boldsymbol{\gamma}$ can be interpreted as a set of variational parameters. The end result of this analysis is a generalized evidence maximization procedure that is equivalent to the one originally formulated in [94]. 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 full Bayesian 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.

Unlike previous Chapters, here we take some consideration of the regression problem where, from a fully Bayesian perspective, the ultimate goal is accurately forming the predictive distribution $p(t_*|\boldsymbol{t})$, where t_* is an unknown response value not included in the training set $\boldsymbol{t}$. When $p(t_*|\boldsymbol{t})$ is not feasible to obtain, approximate methods are often used that, ideally should capture the mass in the full model [56]. The material contained in this Chapter quantifies exactly how SBL models the mass in the full predictive distribution, thus supporting heuristic claims made in [94].

IV.A Introduction

In an archetypical regression situation, we are presented with a collection of N regressor/target pairs $\{\boldsymbol{\phi}_i \in \mathbb{R}^M, t_i \in \mathbb{R}\}_{i=1}^N$ and the goal is to find a vector of weights $\boldsymbol{w}$ such that, in some sense,

$$t_i \approx \boldsymbol{\phi}_i^T \boldsymbol{w}, \ \forall i \quad \text{or} \quad \boldsymbol{t} \approx \Phi \boldsymbol{w}, \tag{IV.1}$$

where $\boldsymbol{t} \triangleq [t_1, \ldots, t_N]^T$ and $\Phi \triangleq [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N]^T \in \mathbb{R}^{N \times M}$. Ideally, we would like to learn this relationship such that, given a new training vector $\boldsymbol{\phi}_*$, we can make accurate predictions of $\boldsymbol{t}_*$, i.e., we would like to avoid overfitting. In practice, this requires some form of regularization, or a penalty on overly complex models.

The sparse Bayesian learning (SBL) framework was originally derived to find robust solutions to regression problems. When Φ is square and formed from a positive-definite kernel function, we obtain the relevance vector machine (RVM), a Bayesian competitor of SVMs with several significant advantages [23, 94].

IV.A.1 Sparse Bayesian Learning for Regression

Given a new regressor vector $\boldsymbol{\phi}_*$, the full Bayesian treatment of (IV.1) involves finding the predictive distribution $p(t_*|\boldsymbol{t})$.[1] We typically compute this distribution by marginalizing over the model weights, i.e.,

$$p(t_*|\boldsymbol{t}) = \frac{1}{p(\boldsymbol{t})} \int p(t_*|\boldsymbol{w}) p(\boldsymbol{w}, \boldsymbol{t}) d\boldsymbol{w}, \tag{IV.2}$$

where the joint density $p(\boldsymbol{w}, \boldsymbol{t}) = p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w})$ combines all relevant information from the training data (likelihood principle) with our prior beliefs about the model weights. The likelihood term $p(\boldsymbol{t}|\boldsymbol{w})$ is assumed to be Gaussian,

$$p(\boldsymbol{t}|\boldsymbol{w}) = (2\pi\lambda)^{-N/2} \exp\left(-\frac{1}{2\lambda}\|\boldsymbol{t} - \Phi\boldsymbol{w}\|^2\right), \tag{IV.3}$$

where for now we assume that the noise variance λ is known. For sparse priors $p(\boldsymbol{w})$ (possibly improper), the required integrations, including the computation of the normalizing term $p(\boldsymbol{t})$, are typically intractable, and we are forced to accept some form of approximation to $p(\boldsymbol{w}, \boldsymbol{t})$.

[1] For simplicity, we omit explicit conditioning on Φ and $\boldsymbol{\phi}_*$, i.e., $p(t_*|\boldsymbol{t}) \equiv p(t_*|\boldsymbol{t}, \Phi, \boldsymbol{\phi}_*)$.

Sparse Bayesian learning addresses this issue by introducing a set of hyperparameters into the specification of the problematic weight prior $p(\boldsymbol{w})$ before adopting a particular approximation. The key assumption is that $p(\boldsymbol{w})$ can be expressed as

$$p(\boldsymbol{w}) = \prod_{i=1}^{M} p(w_i) = \prod_{i=1}^{M} \int p(w_i|\gamma_i)p(\gamma_i)d\gamma_i, \tag{IV.4}$$

where $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_M]^T$ represents a vector of hyperparameters, (one for each weight). The implicit SBL derivation presented in [94] can then be reformulated as follows,

$$\begin{aligned} p(t_*|\boldsymbol{t}) &= \frac{1}{p(\boldsymbol{t})} \int p(t_*|\boldsymbol{w})p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w} \\ &= \frac{1}{p(\boldsymbol{t})} \int \int p(t_*|\boldsymbol{w})p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})d\boldsymbol{w}d\boldsymbol{\gamma}. \end{aligned} \tag{IV.5}$$

Proceeding further, by applying Bayes' rule to this expression, we can exploit the plugin rule [21] via,

$$\begin{aligned} p(t_*|\boldsymbol{t}) &= \int \int p(t_*|\boldsymbol{w})p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{\gamma})\frac{p(\boldsymbol{\gamma}|\boldsymbol{t})}{p(\boldsymbol{t}|\boldsymbol{\gamma})}d\boldsymbol{w}d\boldsymbol{\gamma} \\ &\approx \int \int p(t_*|\boldsymbol{w})p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{\gamma})\frac{\delta(\boldsymbol{\gamma}_{MAP})}{p(\boldsymbol{t}|\boldsymbol{\gamma})}d\boldsymbol{w}d\boldsymbol{\gamma} \\ &= \frac{1}{p(\boldsymbol{t};\boldsymbol{\gamma}_{MAP})} \int p(t_*|\boldsymbol{w})p(\boldsymbol{w},\boldsymbol{t};\boldsymbol{\gamma}_{MAP})d\boldsymbol{w}. \end{aligned} \tag{IV.6}$$

The essential difference from (IV.2) is that we have replaced $p(\boldsymbol{w},\boldsymbol{t})$ with the approximate distribution $p(\boldsymbol{w},\boldsymbol{t};\boldsymbol{\gamma}_{MAP}) = p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w};\boldsymbol{\gamma}_{MAP})$. Also, the normalizing term becomes $\int p(\boldsymbol{w},\boldsymbol{t};\boldsymbol{\gamma}_{MAP})d\boldsymbol{w}$ and we assume that all required integrations can now be handled in closed form. Of course the question remains, how do we structure this new

set of parameters $\boldsymbol{\gamma}$ to accomplish this goal? The answer is that the hyperparameters enter as weight prior variances of the form,

$$p(w_i|\gamma_i) = \mathcal{N}(0, \gamma_i). \tag{IV.7}$$

The hyperpriors are given by,

$$p(\gamma_i^{-1}) \propto \gamma_i^{1-a} \exp(-b/\gamma_i), \tag{IV.8}$$

where $a, b > 0$ are constants. The crux of the actual learning procedure presented in [94] is to find some MAP estimate of $\boldsymbol{\gamma}$ (or more accurately, a function of $\boldsymbol{\gamma}$). In practice, we find that many of the estimated γ_i's converge to zero, leading to sparse solutions since the corresponding weights, and therefore columns of Φ, can effectively be pruned from the model. The Gaussian assumptions, both on $p(\boldsymbol{t}|\boldsymbol{w})$ and $p(\boldsymbol{w};\boldsymbol{\gamma})$, then facilitate direct, analytic computation of (IV.6).

IV.A.2 Ambiguities in Current SBL Derivation

Modern Bayesian analysis is primarily concerned with finding distributions and locations of significant probability mass, not just modes of distributions, which can be very misleading in many cases [56]. With SBL, the justification for the additional level of sophistication (i.e., the inclusion of hyperparameters) is that the adoption of the plugin rule (i.e., the approximation $p(\boldsymbol{w}, \boldsymbol{t}) \approx p(\boldsymbol{w}, \boldsymbol{t}; \boldsymbol{\gamma}_{MAP})$) is reflective of the true mass, at least sufficiently so for predictive purposes. However, no rigorous motivation

for this particular claim is currently available nor is it immediately obvious exactly how the mass of this approximate distribution relates to the true mass.

A more subtle difficulty arises because MAP estimation, and hence the plugin rule, is not invariant under a change in parameterization. Specifically, for an invertible function $f(\cdot)$,

$$[f(\boldsymbol{\gamma})]_{MAP} \neq f(\boldsymbol{\gamma}_{MAP}). \tag{IV.9}$$

Different transformations lead to different modes and ultimately, different approximations to $p(\boldsymbol{w}, \boldsymbol{t})$ and therefore $p(t_*|\boldsymbol{t})$. So how do we decide which one to use? The canonical form of SBL, and the one that has displayed remarkable success in the literature, does not in fact find a mode of $p(\boldsymbol{\gamma}|\boldsymbol{t})$, but a mode of $p(-\log \boldsymbol{\gamma}|\boldsymbol{t})$. But again, why should this mode necessarily be more reflective of the desired mass than any other?

As already mentioned, SBL often leads to sparse results in practice, namely, the approximation $p(\boldsymbol{w}, \boldsymbol{t}; \boldsymbol{\gamma}_{MAP})$ is typically nonzero only on a small subspace of M-dimensional $\boldsymbol{w}$ space. The question remains, however, why should an approximation to the full Bayesian treatment necessarily lead to sparse results in practice?

To address all of these ambiguities, we will herein demonstrate that the sparse Bayesian learning procedure outlined above can be recast as the application of a rigorous variational approximation to the distribution $p(\boldsymbol{w}, \boldsymbol{t})$.[2] This will allow us to quantify the exact relationship between the true mass and the approximate mass of this distribution. In effect, we will demonstrate that SBL is attempting to directly capture significant por-

[2]We note that the analysis in this paper is different from [5], which provides an alternative SBL derivation using a factorial approximation to minimize a Kullback-Leibler divergence-based cost function. The connection between these two types of variational methods can be found in [70].

tions of the probability mass of $p(\boldsymbol{w}, \boldsymbol{t})$, while still allowing us to perform the required integrations. This framework also obviates the necessity of assuming any hyperprior $p(\boldsymbol{\gamma})$ and is independent of the (subjective) parameterization (e.g., $\boldsymbol{\gamma}$ or $-\log \boldsymbol{\gamma}$, etc.). Moreover, this perspective leads to natural, intuitive explanations of why sparsity is observed in practice and why, in general, this need not be the case. Chapter V will consider this issue in greater detail.

IV.B A Variational Interpretation of Sparse Bayesian Learning

To begin, we review that the ultimate goal of this analysis is to find a well-motivated approximation to the distribution

$$p(t_*|\boldsymbol{t}; \mathcal{H}) \propto \int p(t_*|\boldsymbol{w}) p(\boldsymbol{w}, \boldsymbol{t}; \mathcal{H}) d\boldsymbol{w} = \int p(t_*|\boldsymbol{w}) p(\boldsymbol{t}|\boldsymbol{w}) p(\boldsymbol{w}; \mathcal{H}) d\boldsymbol{w}, \qquad \text{(IV.10)}$$

where we have explicitly noted the hypothesis of a model with a sparsity inducing (possibly improper) weight prior by $\mathcal{H}$. As already mentioned, the integration required by this form is analytically intractable and we must resort to some form of approximation. To accomplish this, we appeal to variational methods to find a viable approximation to $p(\boldsymbol{w}, \boldsymbol{t}; \mathcal{H})$ [47]. We may then substitute this approximation into (IV.10), leading to tractable integrations and analytic posterior distributions. To find a class of suitable approximations, we first express $p(\boldsymbol{w}; \mathcal{H})$ in its dual form by introducing a set of variational parameters. This is similar to a procedure outlined in [31] in the context of independent component analysis.

IV.B.1 Dual Form Representation of $p(\boldsymbol{w};\mathcal{H})$

At the heart of this methodology is the ability to represent a convex function in its dual form. For example, given a convex function $f(y):\mathbb{R}\rightarrow\mathbb{R}$, the dual form is given by

$$f(y)=\sup_{\upsilon}\left[\upsilon y-f^{*}(\upsilon)\right], \tag{IV.11}$$

where $f^{*}(\upsilon)$ denotes the conjugate function. Geometrically, this can be interpreted as representing $f(y)$ as the upper envelope or supremum of a set of lines parameterized by υ. The selection of $f^{*}(\upsilon)$ as the intercept term ensures that each line is tangent to $f(y)$. If we drop the maximization in (IV.11), we obtain the bound

$$f(y)\geq\upsilon y-f^{*}(\upsilon). \tag{IV.12}$$

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

To apply this theory to the problem at hand, we specify the form for our sparse prior $p(\boldsymbol{w};\mathcal{H})=\prod_{i=1}^{M}p(w_i;\mathcal{H})$. Using (IV.7) and (IV.8), we obtain the prior

$$p(w_i;\mathcal{H})=\int p(w_i|\gamma_i)p(\gamma_i)d\gamma_i=C\left(b+\frac{w_i^2}{2}\right)^{-(a+1/2)}, \tag{IV.13}$$

which for $a,b>0$ is proportional to a Student-t density. The constant C is not chosen to enforce proper normalization; rather, it is chosen to facilitate the variational analysis below. Also, this density function can be seen to encourage sparsity since it has heavy

tails and a sharp peak at zero. Clearly $p(w_i;\mathcal{H})$ is not convex in w_i; however, if we let $y_i \triangleq w_i^2$ as suggested in [47] and define

$$f(y_i) \triangleq \log p(w_i;\mathcal{H}) = -(a+1/2)\log\left(b+\frac{y_i}{2}\right)+\log C, \tag{IV.14}$$

we see that we now have a convex function in y_i amenable to dual representation. By computing the conjugate function $f^*(y_i)$, constructing the dual, and then transforming back to $p(w_i;\mathcal{H})$, we obtain the representation (see Appendix for details)

$$p(w_i;\mathcal{H}) = \max_{\gamma_i \geq 0} \left[(2\pi\gamma_i)^{-1/2}\exp\left(-\frac{w_i^2}{2\gamma_i}\right)\exp\left(-\frac{b}{\gamma_i}\right)\gamma_i^{-a}\right]. \tag{IV.15}$$

As $a, b \rightarrow 0$, it is readily apparent from (IV.15) that what were straight lines in the y_i domain are now Gaussian functions with variance γ_i in the w_i domain. Figure IV.1 illustrates this connection. When we drop the maximization, we obtain a lower bound on $p(w_i;\mathcal{H})$ of the form

$$p(w_i;\mathcal{H}) \geq p(w_i;\hat{\mathcal{H}}) \triangleq (2\pi\gamma_i)^{-1/2}\exp\left(-\frac{w_i^2}{2\gamma_i}\right)\exp\left(-\frac{b}{\gamma_i}\right)\gamma_i^{-a}, \tag{IV.16}$$

which serves as our approximate prior to $p(w;\mathcal{H})$. From this relationship, we see that $p(w_i;\hat{\mathcal{H}})$ does not integrate to one, except in the special case when $a, b \rightarrow 0$. We will now incorporate these results into an algorithm for finding a good $\hat{\mathcal{H}}$, or more accurately $\hat{\mathcal{H}}(\boldsymbol{\gamma})$, since each candidate hypothesis is characterized by a different set of variational parameters.

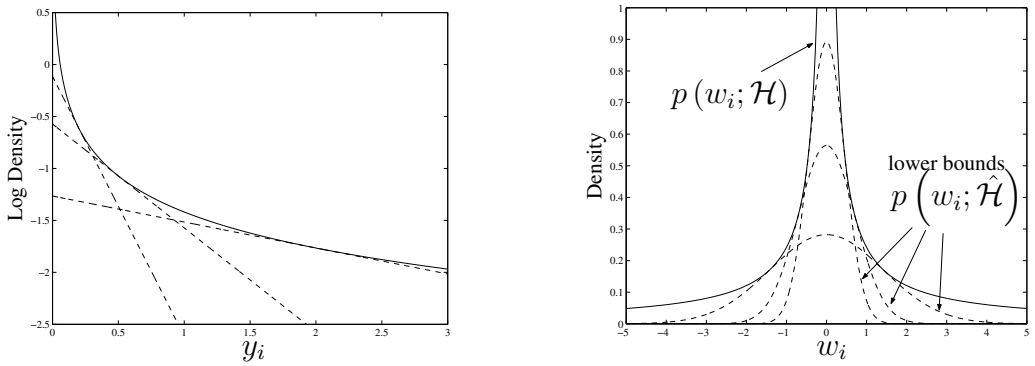

Figure IV.1: Variational approximation example in both y_i space and w_i space for $a, b \to 0$. *Left*: Dual forms in y_i space. The solid line represents the plot of $f(y_i)$ while the dotted lines represent variational lower bounds in the dual representation for three different values of υ_i. *Right*: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

IV.B.2 Variational Approximation to $p(\boldsymbol{w}, \boldsymbol{t}; \mathcal{H})$

So now that we have a variational approximation to the problematic weight prior, we must return to our original problem of estimating $p(t_*|\boldsymbol{t};\mathcal{H})$. Since the integration is intractable under model hypothesis $\mathcal{H}$, we will instead compute $p(t_*|\boldsymbol{t};\hat{\mathcal{H}})$ using $p(\boldsymbol{w},\boldsymbol{t};\hat{\mathcal{H}}) = p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w};\hat{\mathcal{H}})$, with $p(\boldsymbol{w};\hat{\mathcal{H}})$ defined as in (IV.16). How do we choose this approximate model? In other words, given that different $\hat{\mathcal{H}}$ are distinguished by a different set of variational parameters $\boldsymbol{\gamma}$, how do we choose the most appropriate $\boldsymbol{\gamma}$? Consistent with modern Bayesian analysis, we concern ourselves not with matching modes of distributions, but with aligning regions of significant probability mass. In choosing $p(\boldsymbol{w},\boldsymbol{t};\hat{\mathcal{H}})$, we would therefore like to match, where possible, significant regions of probability mass in the true model $p(\boldsymbol{w},\boldsymbol{t};\mathcal{H})$. For a given $\boldsymbol{t}$, an obvious way

to do this is to select $\hat{\mathcal{H}}$ by minimizing the sum of the misaligned mass, i.e.,

$$\begin{aligned} \hat{\mathcal{H}} &= \arg\min_{\hat{\mathcal{H}}} \int \left| p(\boldsymbol{w}, \boldsymbol{t}; \mathcal{H}) - p(\boldsymbol{w}, \boldsymbol{t}; \hat{\mathcal{H}}) \right| d\boldsymbol{w} \\ &= \arg\max_{\hat{\mathcal{H}}} \int p(\boldsymbol{t}|\boldsymbol{w}) p(\boldsymbol{w}; \hat{\mathcal{H}}) d\boldsymbol{w}, \end{aligned} \tag{IV.17}$$

where the variational assumptions have allowed us to remove the absolute value (since the argument must always be positive). Also, we note that (IV.17) is tantamount to selecting the variational approximation with maximal Bayesian evidence [56]. In other words, we are selecting the $\hat{\mathcal{H}}$, out of a class of variational approximations to $\mathcal{H}$, that most probably explains the training data $\boldsymbol{t}$, marginalized over the weights.

From an implementational standpoint, (IV.17) can be reexpressed using (IV.16) as,

$$\begin{aligned} \boldsymbol{\gamma} &= \arg\max_{\boldsymbol{\gamma}} \log \int p(\boldsymbol{t}|\boldsymbol{w}) \prod_{i=1}^{M} p\left(w_i; \hat{\mathcal{H}}(\gamma_i)\right) d\boldsymbol{w} \\ &= \arg\max_{\boldsymbol{\gamma}} -\frac{1}{2}\left[\log|\Sigma_t| + \boldsymbol{t}^T \Sigma_t^{-1} \boldsymbol{t}\right] + \sum_{i=1}^{M} \left(-\frac{b}{\gamma_i} - a \log \gamma_i\right), \end{aligned} \tag{IV.18}$$

where $\Sigma_t \triangleq \lambda I + \Phi \text{diag}(\boldsymbol{\gamma}) \Phi^T$. This is the same cost function as in [94] only without terms resulting from a prior on λ, which we will address later. Thus, the end result of this analysis is an evidence maximization procedure equivalent to the one in [94]. The difference is that, where before we were optimizing over a somewhat arbitrary model parameterization, now we see that it is actually optimization over the space of variational approximations to a model with a sparse, regularizing prior. Also, we know

from (IV.17) that this procedure is effectively matching, as much as possible, the mass of the full model $p(\boldsymbol{w}, \boldsymbol{t}; \hat{\mathcal{H}})$.

IV.C Analysis

While the variational perspective is interesting, two pertinent questions still remain:

1. Why should it be that approximating a sparse prior $p(\boldsymbol{w}; \mathcal{H})$ leads to sparse representations in practice?

2. How do we extend these results to handle an unknown, random variance λ?

We first treat *Question (1)*. In Figure IV.2 below, we have illustrated a $2D$ example of evidence maximization within the context of variational approximations to the sparse prior $p(\boldsymbol{w}; \mathcal{H})$. For now, we will assume $a, b \rightarrow 0$, which from (IV.13), implies that $p(w_i; \mathcal{H}) \propto 1/|w_i|$ for each i. On the left, the shaded area represents the region of $\boldsymbol{w}$ space where both $p(\boldsymbol{w}; \mathcal{H})$ and $p(\boldsymbol{t}|\boldsymbol{w})$ (and therefore $p(\boldsymbol{w}, \boldsymbol{t}; \mathcal{H})$) have significant probability mass. Maximization of (IV.17) involves finding an approximate distribution $p(\boldsymbol{w}, \boldsymbol{t}; \hat{\mathcal{H}})$ with a substantial percentage of its mass in this region.

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 the largest percentage of its mass (and therefore the mass of $p(\boldsymbol{w}, \boldsymbol{t}; \hat{\mathcal{H}}_a)$) in the shaded region. This

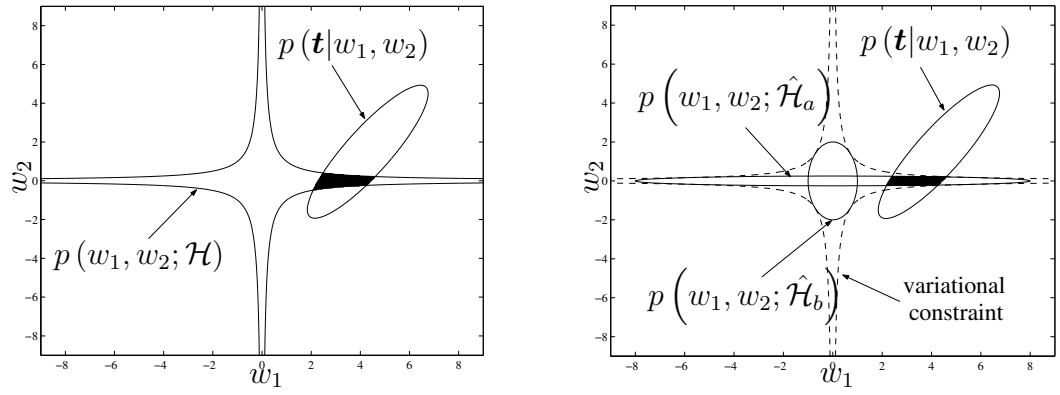

Figure IV.2: Comparison between full model and approximate models with $a, b \to 0$. *Left:* Contours of equiprobability density for $p(\boldsymbol{w}; \mathcal{H})$ and constant likelihood $p(\boldsymbol{t}|\boldsymbol{w})$; the prominent density and likelihood lie within each region respectively. The 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 $p(\boldsymbol{w}; \mathcal{H})$.

corresponds with a prior of

$$p(\boldsymbol{w}; \hat{\mathcal{H}}_a) = p(w_1, w_2; \gamma_1 \gg 0, \gamma_2 \approx 0). \tag{IV.19}$$

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 and b go to zero), 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 and the analysis in [23]. In higher dimensions, the algorithm only retains those weights associated with the prior spines 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.

In contrast, when $a, b > 0$, the situation is somewhat different. It is not difficult to show that, assuming a noise variance $\lambda > 0$, the variational approximation to $p(\boldsymbol{w}, \boldsymbol{t}; \mathcal{H})$ with maximal evidence cannot have any $\gamma_i = w_i = 0$. Intuitively, this occurs because the now *finite* spines of the prior $p(\boldsymbol{w}; \mathcal{H})$, which bound the variational approximation, do not allow us to place infinite prior density in any region of weight space (as occurred previously when any $\gamma_i \to 0$). Consequently, if any γ_i goes to zero with $a, b > 0$, the associated approximate prior mass, and therefore the approximate evidence, must also fall to zero by (IV.16). As such, *models with all non-zero weights will be now be favored when we form the variational approximation*. We therefore cannot assume an approximation to a sparse prior will necessarily give us sparse results in practice.

We now address *Question (2)*. Thus far, we have considered a known, fixed noise variance λ; however, what if λ is unknown? SBL assumes it is unknown and random with prior distribution $p(1/\lambda) \propto (\lambda)^{1-c} \exp(-d/\lambda)$, and $c, d > 0$. After integrating

out the unknown λ, we arrive at the implicit likelihood equation,

$$p(\boldsymbol{t}|\boldsymbol{w}) = \int p(\boldsymbol{t}|\boldsymbol{w},\lambda)p(\lambda)d\lambda \propto \left(d + \frac{1}{2}\|\boldsymbol{t} - \Phi\boldsymbol{w}\|^2\right)^{-(\bar{c}+1/2)}, \tag{IV.20}$$

where $\bar{c} \triangleq c + (N-1)/2$. We may then form a variational approximation to the likelihood in a similar manner as before (with w_i being replaced by $\|\boldsymbol{t} - \Phi\boldsymbol{w}\|$) giving us,

$$\begin{aligned} p(\boldsymbol{t}|\boldsymbol{w}) &\geq (2\pi)^{-N/2}(\lambda)^{-1/2}\exp\left(-\frac{1}{2\lambda}\|\boldsymbol{t} - \Phi\boldsymbol{w}\|^2\right)\exp\left(-\frac{d}{\lambda}\right)(\lambda)^{-\bar{c}} \\ &= (2\pi\lambda)^{-N/2}\exp\left(-\frac{1}{2\lambda}\|\boldsymbol{t} - \Phi\boldsymbol{w}\|^2\right)\exp\left(-\frac{d}{\lambda}\right)(\lambda)^{-c}, \end{aligned} \tag{IV.21}$$

where the second step follows by substituting back in for $\bar{c}$. By replacing $p(\boldsymbol{t}|\boldsymbol{w})$ with the lower bound from (IV.21), we then maximize over the variational parameters $\boldsymbol{\gamma}$ and λ via

$$\boldsymbol{\gamma}, \lambda = \arg\max_{\boldsymbol{\gamma},\lambda} -\frac{1}{2}\left[\log|\Sigma_t| + \boldsymbol{t}^T\Sigma_t^{-1}\boldsymbol{t}\right] + \sum_{i=1}^{M}\left(-\frac{b}{\gamma_i} - a\log\gamma_i\right) - \frac{d}{\lambda} - c\log\lambda, \tag{IV.22}$$

the exact SBL optimization procedure. Thus, we see that the entire SBL framework, including noise variance estimation, can be seen in variational terms. However, as discussed in Section VIII.A, care should be exercised when jointly estimating $\boldsymbol{\gamma}$ and λ.

IV.D Conclusions

The end result of this analysis is an evidence maximization procedure that is equivalent to the generalized SBL framework originally formulated in [94]. The difference is that, where before we were optimizing over a somewhat arbitrary model parameterization, we now see that SBL is actually searching a space of variational approximations to find an alternative distribution that captures the significant mass of the full model. Additionally, this formulation obviates the necessity of assuming any subjective hyperpriors and leads to natural, intuitive explanations of why sparsity is achieved in practice. This topic will be taken up in more detail in the next Chapter, where the general relationship between sparse priors and approximations to sparse priors is discussed.

IV.E Acknowledgements

This chapter, in part, is a reprint of material published as “Perspectives on Sparse Bayesian Learning," *Advances in Neural Information Processing Systems 16* (2004). J.A. Palmer and I were the primary authors and B.D. Rao supervised the research.

IV.F Appendix: Derivation of the Dual Form of $p(w_i; \mathcal{H})$

To accommodate the variational analysis of Sec. IV.B.1, we require the dual representation of $p(w_i; \mathcal{H})$. As an intermediate step, we must find the dual representation

of $f(y_i)$, where $y_i \triangleq w_i^2$ and

$$f(y_i) \triangleq \log p(w_i; \mathcal{H}) = \log \left[C \left(b + \frac{y_i}{2} \right)^{-(a+1/2)} \right]. \tag{IV.23}$$

To accomplish this, we find the conjugate function $f^*(\upsilon_i)$ using the duality relation

$$f^*(\upsilon_i) = \max_{y_i} \left[\upsilon_i y_i - f(y_i) \right] = \max_{y_i} \left[\upsilon_i y_i - \log C + \left(a + \frac{1}{2} \right) \log \left(b + \frac{y_i}{2} \right) \right] \tag{IV.24}$$

To find the maximizing y_i, we take the gradient of the left side and set it to zero, giving us,

$$y_i^{max} = -\frac{a}{\upsilon_i} - \frac{1}{2\upsilon_i} - 2b. \tag{IV.25}$$

Substituting this value into the expression for $f^*(\upsilon_i)$ and selecting

$$C = (2\pi)^{-1/2} \exp \left[- \left(a + \frac{1}{2} \right) \right] \left(a + \frac{1}{2} \right)^{(a+1/2)}, \tag{IV.26}$$

we arrive at

$$f^*(\upsilon_i) = \left(a + \frac{1}{2} \right) \log \left(\frac{-1}{2\upsilon_i} \right) + \frac{1}{2} \log 2\pi - 2b\upsilon_i. \tag{IV.27}$$

We are now ready to represent $f(y_i)$ in its dual form, observing first that we only need consider maximization over $\upsilon_i \leq 0$ since $f(y_i)$ is a monotonically decreasing function (i.e., all tangent lines will have negative slope). Proceeding forward, we have

$$f(y_i) = \max_{\upsilon_i \leq 0} \left[\upsilon_i y_i - f^*(\upsilon_i) \right] = \max_{\gamma_i \geq 0} \left[\frac{-y_i}{2\gamma_i} - \left(a + \frac{1}{2} \right) \log \gamma_i - \frac{1}{2} \log 2\pi - \frac{b}{\gamma_i} \right] \tag{IV.28}$$

where we have used the monotonically increasing transformation $\upsilon_i = -1/(2\gamma_i), \gamma_i \geq 0$. The attendant dual representation of $p(w_i; \mathcal{H})$ can then be obtained by exponentiating both sides of (IV.28) and substituting $y_i = w_i^2$,

$$p(w_i; \mathcal{H}) = \max_{\gamma_i \geq 0} \left[\frac{1}{\sqrt{2\pi\gamma_i}} \exp\left(-\frac{w_i^2}{2\gamma_i}\right) \exp\left(-\frac{b}{\gamma_i}\right) \gamma_i^{-a} \right]. \tag{IV.29}$$

Chapter V

A General Framework for Latent Variable Models with Sparse Priors

A variety of general Bayesian methods, some of which have been discussed in previous chapters, have recently been introduced for finding sparse representations from overcomplete dictionaries of candidate features. These methods often capitalize on latent structure inherent in sparse distributions to perform standard MAP estimation, variational Bayes, approximation using convex duality, or evidence maximization. Despite their reliance on sparsity-inducing priors however, these approaches may or may not actually lead to sparse representations in practice, and so it is a challenging task to determine which algorithm and sparse prior is appropriate. Rather than justifying prior selections and modelling assumptions based on the credibility of the full Bayesian model as is commonly done, this chapter bases evaluations on the actual cost functions that emerge from each method. Two minimal conditions are postulated that ideally any

sparse learning objective should satisfy. Out of all possible cost functions that can be obtained from the methods described above using (virtually) any sparse prior, a unique function is derived that satisfies these conditions. Both sparse Bayesian learning (SBL) and basis pursuit (BP) are special cases.

These results elucidate connections between methods and suggests new sparse learning cost functions. For example, we demonstrate that all of the above sparse learning procedures can be viewed as simple MAP estimation giving the appropriate prior. However, where as traditional MAP methods for sparse recovery (e.g., BP, LASSO, FOCUSS, etc.) employ a factorial (separable) prior, SBL and other empirical Bayesian methods do not.

V.A Introduction

Here we will again be concerned with the generative model

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

where $\Phi \in \mathbb{R}^{N \times M}$ is a dictionary of unit ℓ_2-norm basis vectors or features, $\boldsymbol{w}$ is a vector of unknown weights, $\boldsymbol{t}$ is the observation vector, and $\boldsymbol{\epsilon}$ is uncorrelated noise distributed as $\mathcal{N}(0, \lambda I)$. In many practical situations, this dictionary will be *overcomplete*, meaning $M > N$ and $\mathrm{rank}(\Phi) = N$. When large numbers of features are present relative to the signal dimension, the estimation problem is fundamentally ill-posed. A Bayesian framework is intuitively appealing for formulating these types of problems because prior

assumptions must be incorporated, whether explicitly or implicitly, to regularize the solution space.

Recently, there has been a growing interest in models that employ sparse priors to encourage solutions with mostly zero-valued coefficients. For purposes of optimization, approximation, and inference, these models can be conveniently framed in terms of a collection of latent variables $\boldsymbol{\gamma} \triangleq [\gamma_1, \ldots, \gamma_M]^T$. The latent variables dictate the structure of the sparse prior in one of two ways. First, in the integral-type representation, the prior is formed as a scale mixture of Gaussians via

$$p(\boldsymbol{w}) = \prod_{i=1}^{M} p(w_i), \qquad p(w_i) = \int \mathcal{N}(0, \gamma_i) p(\gamma_i) d\gamma_i. \tag{V.2}$$

In contrast, the convex-type representation takes the form[1]

$$p(w_i) = \sup_{\gamma_i \geq 0} \mathcal{N}(0, \gamma_i) p(\gamma_i), \tag{V.3}$$

whose form is rooted in convex analysis and duality theory. As shown in [70], virtually all sparse priors of interest can be decomposed using both (V.2) and (V.3), including the popular Laplacian, Jeffreys, Student's t, and generalized Gaussian priors.[2] The key requirement is that $p(w_i)$ is *strongly supergaussian*, which requires that

$$p(w_i) \propto \exp[-g(w_i^2)], \tag{V.4}$$

[1]Here we use a slight abuse of notation, in that $p(\gamma_i)$ need not be a proper probability distribution.
[2]The convex-type representation is slightly more general than (V.2).

with $g(\cdot)$ a concave and non-decreasing function.

In the context of regression and model selection, the fully Bayesian treatment would involve integration (or maximization for the convex representation) over both the latent variables and the unknown weights. With sparse priors, however, this is intractable. Moreover, in applications where sparsity is important, often a sparse point estimate $\hat{\boldsymbol{w}}$ is all that is required, rather than merely a good estimate of $p(\boldsymbol{t})$ or the conditional distribution of new data-points t^*, i.e., $p(t^*|\boldsymbol{t})$. As such, nearly all models with sparse priors are handled in one of two ways, both of which can be viewed as approximations to the full model.

First, the latent structure afforded by (V.2) and (V.3) offers a very convenient means of obtaining (local) MAP estimates of $\boldsymbol{w}$ using generalized EM procedures that iteratively solve

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w}). \tag{V.5}$$

Henceforth referred to as *Type I methods*, common examples include minimum ℓ_p-quasi-norm approaches [50, 79], Jeffreys prior-based methods sometimes called FOCUSS [24, 27, 34], and algorithms for computing the basis pursuit (BP) or Lasso solution [27, 54, 79].

Secondly, instead of integrating out (or maximizing out) the hyperparameters, *Type II methods* instead integrate out the unknown $\boldsymbol{w}$ and then solve

$$\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} p(\boldsymbol{\gamma}|\boldsymbol{t}) = \arg\max_{\boldsymbol{\gamma}} \int p(\boldsymbol{t}|\boldsymbol{w})\mathcal{N}(0,\boldsymbol{\gamma})p(\boldsymbol{\gamma})d\boldsymbol{w}. \tag{V.6}$$

Once $\hat{\boldsymbol{\gamma}}$ is obtained, a point estimate for $\boldsymbol{w}$ naturally emerges as

$$\hat{\boldsymbol{w}} = \mathrm{E}\left[\boldsymbol{w}|\boldsymbol{t};\hat{\boldsymbol{\gamma}}\right] = \hat{\Gamma}\Phi^T\left(\lambda I + \Phi\hat{\Gamma}\Phi^T\right)^{-1}\boldsymbol{t}, \tag{V.7}$$

where $\Gamma \triangleq \mathrm{diag}(\boldsymbol{\gamma})$. Relevant examples include sparse Bayesian learning (SBL) [94], automatic relevance determination (ARD) [66], evidence maximization [86], and methods for learning overcomplete dictionaries [31]. Perhaps surprisingly, even the popular variational mean-field approximations, which optimize a factorial posterior distribution such that $p(\boldsymbol{w},\boldsymbol{\gamma}|\boldsymbol{t}) \approx q(\boldsymbol{w}|\boldsymbol{t})q(\boldsymbol{\gamma}|\boldsymbol{t})$, are equivalent to the Type II methods in the context of strongly supergaussian priors [70]. A specific example of this can be found in [5].

In applying all of these methods in practice, the performance achieving sparse solutions can be highly varied. Results can be highly dependent on the (subjective) parameterization used in forming the latent variables. This occurs because the decomposition of $p(\boldsymbol{w})$ is generally not unique. In some cases, these methods lead to identical results, in others, they may perform poorly or even lead to provably non-sparse representations, despite their foundation on a sparse prior-based generative model. In still other cases, they may be very successful. As such, sorting out the meaningful differences between these methods remains an important issue.

In the past, sparse models have sometimes been justified solely based on their ostensible affiliation with a sparse prior. However, a more thorough means of evaluation involves looking at the actual cost function that results from various prior and modelling

assumptions. We would argue that models should be justified based on this lower level, not the plausibility of the full model, which may be irrelevant and/or non-unique.

In this paper, we will begin by examining the cost functions that emerge from all possible Type I and Type II methods, demonstrating that the former is actually a special case of the latter, with a common underlying set of objective functions uniting both methods. However, it still remains unclear how to reliably select from this class of algorithms when sparsity is the foremost concern. To this effect, we postulate two minimal conditions that ideally any sparse approximation cost function should satisfy. We then select, out of all the possible Type I and II methods discussed above, the unique function that satisfies these two conditions. Interestingly, both BP and SBL are special cases. In general, we would argue that these results significantly compress the space of 'useful' sparse algorithms and provides a more rigorous justification for using a particular method consistent with observed empirical results. We conclude by discussing an important distinguishing factor between candidate algorithms that suggests avenues for improvement.

V.B A Unified Cost Function

Given the significant discrepancies between the various latent variable sparse approximation methods, it would seem that the respective cost functions should be very different. However, this section demonstrates that they all can be viewed as special cases of a single underlying objective function. We start with two intermediate results before presenting the main idea.

Lemma 7. Given a sparse prior expressible using (V.2) or (V.3), the resulting posterior mode over $\boldsymbol{w}$ (as is sought by Type I methods) can be obtained by minimizing the cost function

$$\mathcal{L}_{(I)}(\boldsymbol{\gamma}; \lambda, f) \triangleq \boldsymbol{t}^T \Sigma_t^{-1} \boldsymbol{t} + \sum_{i=1}^{M} f(\gamma_i) \tag{V.8}$$

over the latent variables $\boldsymbol{\gamma}$, where $\Sigma_t \triangleq \lambda I + \Phi \Gamma \Phi^T$ and $f(\cdot)$ is a suitably chosen function on $[0, \infty)$.

Proof: From basic linear algebra, we have

$$\boldsymbol{t}^T \Sigma_t^{-1} \boldsymbol{t} = \min_{\boldsymbol{w}} \ \frac{1}{\lambda} \|\boldsymbol{t} - \Phi \boldsymbol{w}\|_2^2 + \boldsymbol{w}^T \Gamma^{-1} \boldsymbol{w}. \tag{V.9}$$

The minimizing $\boldsymbol{w}$ is given by (V.7). If we choose $f(\gamma_i) = -g^*\left(\gamma_i^{-1}\right)$, where $g^*(\cdot)$ denotes the concave conjugate of $g(\cdot)$, then the optimization problem becomes

$$\begin{aligned} &\min_{\boldsymbol{\gamma}} \mathcal{L}_{(I)}(\boldsymbol{\gamma}; \lambda, f) = \\ &\qquad \min_{\boldsymbol{\gamma}} \min_{\boldsymbol{w}} \frac{1}{\lambda} \|\boldsymbol{t} - \Phi \boldsymbol{w}\|_2^2 + \boldsymbol{w}^T \Gamma^{-1} \boldsymbol{w} + \sum_i -g^*(\gamma_i^{-1}). \end{aligned} \tag{V.10}$$

When we switch the order of minimization (allowable) and optimize over $\boldsymbol{\gamma}$ first, we get

$$\min_{\boldsymbol{\gamma}} \boldsymbol{w}^T \Gamma^{-1} \boldsymbol{w} + \sum_i -g^*(\gamma_i^{-1}) = \sum_{i=1}^{M} g(w_i^2), \tag{V.11}$$

which follows from the representation (V.3) and its assumption that $g(\cdot)$ is concave in

w_i^2 [70]. Since the posterior mode is given by the minimum of

$$\mathcal{L}_{(I)}(\boldsymbol{w};\lambda,f) \triangleq -\log p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w}) \equiv \|\boldsymbol{t}-\Phi\boldsymbol{w}\|_2^2 + \lambda\sum_{i=1}^{M} g(w_i^2), \tag{V.12}$$

this completes the proof. Additionally, local minima are preserved as well, meaning there is a one-to-one correspondence between local minima of (V.8) and local minima of (V.12). Note that this analysis is valid even for priors constructed via the integral representation (V.2), since such priors are a subset of those built upon (V.3). ■

Lemma 8. All of the Type II methods can be obtained by minimizing the cost function

$$\mathcal{L}_{(II)}(\boldsymbol{\gamma};\lambda,f) \triangleq \boldsymbol{t}^T\Sigma_t^{-1}\boldsymbol{t} + \log|\Sigma_t| + \sum_i f(\boldsymbol{\gamma}_i). \tag{V.13}$$

Proof: This result can be obtained by computing the integral in (V.6) and applying a $-\log(\cdot)$ transformation. The value of $f(\cdot)$ will depend on the prior representation that is assumed. ■

Theorem 7. Up to an inconsequential scaling factor, both the Type I and Type II cost functions can be reduced to (V.13) with the appropriate selection of $f(\cdot)$.

Proof: It only remains to show that (V.8) is a special case of (V.13). This is very straightforward because we can always reparameterize things such that the $\log|\Sigma_t|$ term

vanishes. Let

$$\bar{f}(\cdot) \triangleq \alpha f[\alpha(\cdot)], \quad \bar{\lambda} \triangleq \alpha^{-1}\lambda, \tag{V.14}$$

where $\alpha \geq 0$ is a constant. Note that $\bar{f}(\cdot)$ represents a valid selection for either Type I or Type II methods. Under these definitions, (V.13) can be expressed as

$$\begin{aligned}
\mathcal{L}_{(II)}(\boldsymbol{\gamma}; \bar{\lambda}, \bar{f}) &= \boldsymbol{t}^T \left(\bar{\lambda} I + \Phi\Gamma\Phi^T\right)^{-1} \boldsymbol{t} + \log\left|\bar{\lambda} I + \Phi\Gamma\Phi^T\right| + \sum_{i=1}^{M} \bar{f}(\gamma_i) \\
&= \boldsymbol{t}^T \left[\alpha^{-1}\left(\lambda I + \alpha\Phi\Gamma\Phi^T\right)\right]^{-1} \boldsymbol{t} + \log\left|\alpha^{-1}\left(\lambda I + \alpha\Phi\Gamma\Phi^T\right)\right| \\
&\quad + \sum_{i=1}^{M} \alpha f(\alpha\gamma_i) \\
&\equiv \boldsymbol{t}^T \left[\lambda I + \alpha\Phi\Gamma\Phi^T\right]^{-1} \boldsymbol{t} + \alpha^{-1}\log\left|\lambda I + \alpha\Phi\Gamma\Phi^T\right| + \sum_{i=1}^{M} f(\alpha\gamma_i)
\end{aligned} \tag{V.15}$$

As α becomes large, we have

$$\mathcal{L}_{(II)}(\boldsymbol{\gamma}; \bar{\lambda}, \bar{f}) \rightarrow \boldsymbol{t}\left[\lambda I + \Phi(\alpha\Gamma)\Phi^T\right]^{-1}\boldsymbol{t} + \sum_{i=1}^{M} f(\alpha\gamma_i). \tag{V.16}$$

This is equivalent to (V.8) with the exception of the scaling factor of α on $\boldsymbol{\gamma}$. However, this factor is irrelevant in that the weight estimates $\hat{\boldsymbol{w}}$ so obtained will be identical. To make this explicit, let $\boldsymbol{\gamma}_{(I)}$ denote a minimum to (V.8) while $\boldsymbol{\gamma}_{(II)}$ is a minimum of (V.16), where if follows that $\boldsymbol{\gamma}_{(I)} = \alpha\boldsymbol{\gamma}_{(II)}$. The corresponding weight estimates $\hat{\boldsymbol{w}}_{(I)}$ and $\hat{\boldsymbol{w}}_{(II)}$ will be identical since

$$\begin{aligned}
\hat{\boldsymbol{w}}_{(I)} &= \Gamma_{(I)}\Phi^T\left(\lambda I + \Phi\Gamma_{(I)}\Phi^T\right)^{-1}\boldsymbol{t} \\
&= \alpha\Gamma_{(II)}\Phi^T\left[\alpha\bar{\lambda} I + \Phi\left(\alpha\Gamma_{(II)}\right)\Phi^T\right]^{-1}\boldsymbol{t}
\end{aligned}$$

$$= \Gamma_{(II)}\Phi^T \left(\bar{\lambda} I + \Phi\Gamma_{(II)}\Phi^T\right)^{-1} \boldsymbol{t}$$

$$= \hat{\boldsymbol{w}}_{(II)}. \tag{V.17}$$

This completes the proof. ■

In summary, by choosing the appropriate sparse prior, and therefore the function $f(\cdot)$, any Type I method can be reduced to a special case of Type II. As will be discussed in Section V.E, the real distinction between the two is that Type I methods are restricted to separable (i.e., factorial) effective priors while Type II approaches are not. Additionally, we will drop explicit use of the subscripts $_{(I)}$ and $_{(II)}$, using $\mathcal{L}(\boldsymbol{\gamma}; \lambda, f)$ to denote the cost function for all methods.

V.C Minimal Performance Conditions

In the past, different methods have been justified based on the plausibility of the full model and the full prior $p(\boldsymbol{w})$, or in terms of how well a particular approximation resembles the full model. But this can be problematic since, as already mentioned, sparse priors need not lead to sparsity-promoting cost functions when using Type I or Type II methods, even when well-motivated priors are in service. As such, we base our evaluation solely on two minimal performance criteria that we would argue a cost function should ideally satisfy if sparsity is the overall objective. While certainly there are different notions of sparsity, here we are concerned with cost functions that encourage sparsity in the ℓ_0-norm sense, meaning most weights go to exactly zero, not merely

small values. This notion of sparsity is often crucial, because with large numbers of features, it is very desirable for a variety of reasons that many may be pruned from the model.

Condition 1. *Every local minimum is achieved at a solution with at most N nonzero elements.*

In the noiseless case, this requirement is equivalent to stating that every local minima is achieved at a basic feasible solution (BFS). Many of the MAP algorithms satisfy this condition (e.g., using a generalized Gaussian prior with $p \leq 1$ or a Jeffreys prior). This ensures that an algorithm is guaranteed to prune at least $M - N$ unnecessary coefficients, a minimal sparsity condition.

Condition 2. *Given that $\boldsymbol{t} = \omega \boldsymbol{\phi}_i$ for some $\omega \in \mathbb{R}$ and unique dictionary column $\boldsymbol{\phi}_i$, then there is a unique, minimizing solution characterized by $\hat{\boldsymbol{w}} = \omega \boldsymbol{e}_i$, where $\boldsymbol{e}_i$ is the canonical unit vector.*

This can be viewed as a minimal recoverability criteria: if a method maintains troublesome local minima even when only a single, nonzero element need be found, then serious difficulties may arise for more challenging problems. In the context of source localization for neuroimaging, this is sufficient to ensure zero localization bias [90].

V.D Performance Analysis

Rather than directly considering each possible sparse prior and its attendant latent variable structure, we can instead analyze the general cost function $\mathcal{L}(\boldsymbol{\gamma}; \lambda, f)$ that encompasses all possibilities. This leads to a much more straightforward means of assessing the different Type I and Type II methods. Here we will begin with the assumption that $f(\cdot)$ is an arbitrary differentiable function on $[0, \infty)$. Note that there is some indeterminacy between the specification of the prior and the cost function that results. In other words, a given prior $p(\boldsymbol{w})$ can be decomposed using multiple latent parameterizations, leading to different effective values of $f(\cdot)$.

We begin with some preliminary definitions and results. We will say that the function $f(\cdot)$ is *strictly convex* on some (possibly open) interval $[a, b]$ if $f\left(\varepsilon a + (1-\varepsilon)b\right) < \varepsilon f(a) + (1-\varepsilon) f(b)$ for all $\varepsilon \in (0, 1)$.[3] *Strict concavity* is defined in an analogous manner.

Lemma 9. To satisfy Condition 1, $f(\cdot)$ must be a nondecreasing function on $[0, \infty)$.

This result is very straightforward to show.

Lemma 10. Let $f(\cdot)$ be strictly convex in some (possibly open) interval. Then $\mathcal{L}(\boldsymbol{\gamma}; \lambda, f)$ violates Condition 1.

It is not difficult to create examples that illustrate this result. In general, if a large subset of hyperparameters maintain similar values in the specified convex region, then certain dictionaries with redundant means of achieving nearly the same covariance Σ_t will lead to locally minimizing solutions with more than N nonzero elements.

[3] Here we assume a slightly nonstandard (and weaker) definition of strict convexity.

Lemma 11. Let $f(\cdot)$ be a strictly concave function on $[0, \infty)$. Then $\mathcal{L}(\boldsymbol{\gamma}; \lambda, f)$ violates Condition 2.

The proof has been deferred to Appendix V.G.1. Only the class of non-decreasing affine functions satisfy the above three lemma, which constitute necessary conditions. For sufficiency we have the following result:

Lemma 12. $\mathcal{L}(\boldsymbol{\gamma}; \lambda, f)$ satisfies Conditions 1 and 2 if $f(z) \propto \alpha z$, where $\alpha \geq 0$.

See the Appendix V.G.2 for the proof. Combining all of the above, we arrive at the following conclusion:

Theorem 8. $\mathcal{L}(\boldsymbol{\gamma}; \lambda, f)$ satisfies Conditions 1 and 2 if and only if $f(z) \propto \alpha z$, where $\alpha \geq 0$.

A couple of things are worth noting with respect to this result. First, the implicit prior associated with $f(z) \propto \alpha z$ depends on which representation of the latent variables is assumed. For example, using the integral representation from (V.2) to perform MAP estimation of $\boldsymbol{\gamma}$, we find that $p(\boldsymbol{w})$ is Laplacian, but using the convex representation (or when using the equivalent variational Bayes formulation), $p(\boldsymbol{w})$ becomes a kind of Jeffreys prior-like distribution with an infinite peak at zero. Both lead to the exact same algorithm and cost function, but a very different interpretation of the prior. In contrast, if a Laplacian prior is decomposed using (V.3) as in done in [31], a provably non-sparse cost function results. This underscores the difficulty in choosing a model based on the plausibility of the starting prior rather than performance criteria directly linked to the actual cost function that ensues.

Secondly, both the SBL and BP cost functions can be viewed as special limiting cases of $\mathcal{L}(\boldsymbol{\gamma}; \lambda, f)$ when using $f(z) = \alpha z$. SBL is obtained with $\alpha \rightarrow 0$, while BP results from the assumption $\alpha \rightarrow \infty$, with $\lambda \rightarrow \lambda/\alpha^{1/2}$. The general case is easily implemented using EM updates, where the E-step involves computing the posterior moments

$$E\left[\boldsymbol{w}\boldsymbol{w}^T|\boldsymbol{t};\boldsymbol{\gamma}\right] = \Gamma\Phi^T\Sigma_t^{-1}\boldsymbol{t}\boldsymbol{t}^T\Sigma_t^{-1}\Phi\Gamma + \Gamma - \Gamma\Phi^T\Sigma_t^{-1}\Phi\Gamma, \tag{V.18}$$

while the M-step reduces to

$$\gamma_i = \frac{-1 + \left(1 + 4\alpha E\left[\boldsymbol{w}\boldsymbol{w}^T|\boldsymbol{t};\boldsymbol{\gamma}\right]_{ii}\right)^{1/2}}{2\alpha}. \tag{V.19}$$

Consistent with the above observations, when $\alpha \rightarrow 0$, these expressions reduce to the exact SBL updates (EM version), while the assumptions $\alpha \rightarrow \infty$, with $\lambda \rightarrow \lambda/\alpha^{1/2}$ produce an interior point method for computing the BP solution. For all other α, the algorithm is very effective in empirical tests (not shown), although the optimal value is likely application dependent.

V.E Discussion

Bayesian algorithms for promoting sparsity have been derived using a variety of assumptions, from standard MAP estimation, to variational Bayes, to convex lower-bounding, to evidence maximization, etc. These methods capitalize on latent structure inherent to sparse distributions in one of two ways, leading to the distinction between Type I and Type II methods, all of which can be optimized using a general EM frame-

work [70]. However, despite their reliance on a sparsity-inducing prior, these approaches may or may not actually lead to sparse representations in practice.

Rather than subjectively evaluating different methods based on the plausibility of the particular prior or approximation strategy that is used, in this paper we have chosen to take a step back and evaluate each model with respect to how well the underlying cost function encourages sparsity. To accomplish this, we have described a general class of objective functions that encompasses all Type I and II approaches using results from [70]. From this family, we then demonstrated that only a single function satisfies two broad criteria directly tied to performance in finding sparse representations. Both SBL and BP objectives are special cases of this function. Perhaps not coincidentally then, SBL and BP were respectively the first and second best Bayesian approaches to solving extremely large sparse inverse problems tied to neuroelectromagnetic source imaging using $400+$ times overcomplete dictionaries [75].

A final point is worth exploring regarding the difference between Type I and Type II approaches. In the past, Type I methods, being labelled as MAP estimates for $\boldsymbol{w}$, have been distinguished from Type II methods, which can be viewed as MAP estimates for the hyperparameters $\boldsymbol{\gamma}$. In specific cases, arguments have been made for the merits of one over the other based on intuition or heuristic arguments [58, 94]. But we would argue that this distinction is somewhat tenuous. In fact, all Type II methods can equivalently be viewed as standard MAP estimation in $\boldsymbol{w}$-space using the prior

$$p(\boldsymbol{w}) \propto \exp\left[-\frac{1}{2}\min_{\boldsymbol{\gamma}}\left(\boldsymbol{w}^T\Gamma^{-1}\boldsymbol{w} + \log|\Sigma_t| + \sum_i f(\gamma_i)\right)\right]. \tag{V.20}$$

Although not generally available in closed form, this prior is necessarily concave in $\boldsymbol{w}^2$ in the same sense as the priors (V.2) and (V.3). Unlike the previous prior expressions however, (V.20) is *non-separable*, meaning $p(\boldsymbol{w}) \neq \sum_i p(w_i)$. This we believe is the key distinction between Type I and Type II; both are finding MAP estimates of $\boldsymbol{w}$, but the former is restricted to factorial priors while the latter is not (this is consistent with the notion that Type I is a special case of Type II).

This distinction between factorial and non-factorial priors appears both in $\boldsymbol{w}$-space and in hyperparameter $\boldsymbol{\gamma}$-space and is readily illustrated by comparing SBL and FOCUSS in the latter. Using a determinant identity and results from Section V.B, the SBL cost can be expressed as

$$\begin{aligned} \mathcal{L}_{\text{SBL}}(\boldsymbol{\gamma};\lambda) &= \boldsymbol{t}^T \Sigma_t^{-1} \boldsymbol{t} + \log|\Gamma| + \log\left|\Gamma^{-1} + \lambda^{-1}\Phi^T\Phi\right| \\ &= \mathcal{L}_{\text{FOCUSS}}(\boldsymbol{\gamma};\lambda) + \log\left|\Gamma^{-1} + \lambda^{-1}\Phi^T\Phi\right|. \end{aligned} \tag{V.21}$$

Thus, the two cost functions differ only with respect to the non-separable log-determinant term. In fact, it is this term that allows SBL to satisfy Condition 2 while FOCUSS does not. Again, this reinforces the notion that cost-function-based evaluations can be more direct and meaningful than other critiques.

These issues raise a key question. If we do not limit ourselves to separable regularization terms (i.e., priors), then what is the optimal selection for $p(\boldsymbol{w})$? Perhaps there is a better choice that does not neatly fit into current frameworks that are linked to the Gaussian distribution. This remains an interesting area for further research.

V.F Acknowledgements

This chapter, in part, is a reprint of material that has been submitted under the title “Performance Analysis of Latent Variable Models with Sparse Priors" and is in preparation for submission under the title “A General Framework for Handling Latent Variable Models with Sparse Priors." In both cases, I was the primary author, J.A. Palmer contributed to the research, and K. Kreutz-Delgado and B.D. Rao supervised the research.

V.G Appendix

V.G.1 Proof of Lemma 11

Preliminaries

Assume a dictionary Φ with $M > N$ that satisfies the URP. Condition 2 applies to the scenario where one column of Φ is proportional to $\boldsymbol{t}$. We denote the hyperparameter associated with this column as γ_t. Let $\boldsymbol{\gamma}_*$ be a hyperparameter vector such that $\|\boldsymbol{\gamma}_*\|_0 = N$ and $\gamma_t = 0$, with $\boldsymbol{w}_* = \Gamma_*^{1/2}(\Phi\Gamma_*^{1/2})^\dagger$.

Define $\widetilde{\boldsymbol{\gamma}}$ and $\widetilde{\boldsymbol{w}}$ to be the N nonzero elements in and $\boldsymbol{\gamma}_*$ and $\boldsymbol{w}_*$ respectively, and $\widetilde{\Phi}$ the corresponding columns of Φ. Note that this implies that $\widetilde{\boldsymbol{w}} = \widetilde{\Phi}^{-1}\boldsymbol{t}$. We will later make use of the simple inequality

$$\boldsymbol{t}^T\boldsymbol{t} \;=\; \widetilde{\boldsymbol{w}}^T\widetilde{\Phi}^T\widetilde{\Phi}\widetilde{\boldsymbol{w}} \;\leq\; |\widetilde{\boldsymbol{w}}|^T\mathbf{1}_{(N\times N)}|\widetilde{\boldsymbol{w}}| \;=\; \left(\sum_{i=1}^{N}|\widetilde{w}_i|\right)^2, \tag{V.22}$$

where $\mathbf{1}_{(N\times N)}$ denotes an $N \times N$ matrix of ones and the $|\cdot|$ is understood to apply element-wise. Note that equality can only be achieved when every column of $\widetilde{\Phi}$ is identical up to a factor of -1, which violates the URP assumption. However, we can get arbitrarily close to this bound, while still satisfying the URP, by adding a small perturbation to each column. Finally, we define

$$a_i \triangleq \left.\frac{\partial f(\gamma)}{\partial \gamma}\right|_{\gamma=\widetilde{\gamma}_i} \qquad a_0 \triangleq \left.\frac{\partial f(\gamma)}{\partial \gamma}\right|_{\gamma=0}. \tag{V.23}$$

Sufficient Conditions for Local Minima

With a little effort, it can be shown that the following two conditions are sufficient for $\boldsymbol{\gamma}_*$ to be a local minimum of $\mathcal{L}(\boldsymbol{\gamma}; \lambda = 0, f)$.

Condition (A): $\widetilde{\boldsymbol{\gamma}}$ is the unique minimizer of the reduced cost function

$$\begin{aligned}\mathcal{L}(\widetilde{\boldsymbol{\gamma}}; \lambda = 0, f) &\triangleq \log|\widetilde{\Gamma}| + \boldsymbol{t}^T\left(\widetilde{\Phi}\widetilde{\Gamma}\widetilde{\Phi}^T\right)^{-1}\boldsymbol{t} + \sum_{i=1}^{N} f(\widetilde{\gamma}_i) \\ &= \sum_{i=1}^{N}\left(\log\widetilde{\gamma}_i + \frac{\widetilde{w}_i^2}{\widetilde{\gamma}_i} + f(\widetilde{\gamma}_i)\right).\end{aligned} \tag{V.24}$$

Condition (B):

$$\left.\frac{\partial \mathcal{L}(\boldsymbol{\gamma}; \lambda = 0, f)}{\partial \gamma_t}\right|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_*} > 0. \tag{V.25}$$

This condition can be motivated as follows. If $\boldsymbol{\gamma}_*$ is a local minima to $\mathcal{L}(\boldsymbol{\gamma}; \lambda = 0, f)$, then the gradient with respect to all zero-valued hyperparameters cannot be negative

(as discussed in Section II.C.2 in the case where $f(\cdot)$ equals zero). Otherwise the cost function can be reduced by increasing a particular hyperparameter above zero. By the URP assumption, γ_t will be zero-valued when $\boldsymbol{\gamma} = \boldsymbol{\gamma}_*$, moreover, the gradient with respect to γ_t will always be less than the gradient with respect to any other zero-valued hyperparameter, so if **(B)** holds, no other gradients need be checked.

The proof which follows is a demonstration that these conditions, which together are sufficient for the existence of a local minimum, can always be made to hold for $f(\cdot)$ strictly concave and non-decreasing.

Satisfying Condition (A)

Using simple calculus and some algebraic manipulations, it can be shown that if each $\widetilde{\gamma}_i$ satisfies

$$\widetilde{\gamma}_i = \frac{-1 + \sqrt{1 + 4a_i \widetilde{w}_i^2}}{2a_i}, \tag{V.26}$$

then $\widetilde{\boldsymbol{\gamma}}$ is the unique minimizer of (V.24). Note that a_i is a function of $\widetilde{w}_i$. The $\widetilde{\gamma}_i$ that satisfies (V.26) will increase monotonically as $\widetilde{w}_i$ increases, while a_i will decrease monotonically from a_0 due to the assumed concavity of $f(\cdot)$.

Satisfying Condition (B)

(B) can be reduced to

$$\left.\frac{\partial \mathcal{L}(\gamma_t; \lambda = 0, f)}{\partial \gamma_t}\right|_{\gamma_t = 0} > 0, \tag{V.27}$$

where, excluding terms without γ_t, the relevant cost is

$$\mathcal{L}(\gamma_t; \lambda = 0, f) \triangleq \log\left(1 + \frac{\gamma_t}{\boldsymbol{t}^T\boldsymbol{t}}\beta\right) + \frac{\beta}{1 + \frac{\gamma_t}{\boldsymbol{t}^T\boldsymbol{t}}\beta} + f(\gamma_t), \tag{V.28}$$

where

$$\beta \triangleq \sum_{i=1}^{N} \frac{\widetilde{w}_i^2}{\widetilde{\gamma}_i} = \sum_{i=1}^{N} \frac{2a_i\widetilde{w}_i^2}{-1 + \sqrt{1 + 4a_i\widetilde{w}_i^2}}. \tag{V.29}$$

The later equality follows from satisfying **(A)**. The required gradient can be analytically computed leading to

$$\beta - \beta^2 + a_0\boldsymbol{t}^T\boldsymbol{t} > 0. \tag{V.30}$$

Substituting (V.22) gives the weaker sufficient condition

$$\beta - \beta^2 + a_0\left(\sum_{i=1}^{N} |\widetilde{w}_i|\right)^2 \geq 0. \tag{V.31}$$

To show that there will always exist cases where (V.31) holds, we allow $\boldsymbol{t}$, and therefore each $\widetilde{w}_i$, to grow arbitrarily large. This permits the reduction[4]

$$\beta = \sum_i a_i^{1/2}|\widetilde{w}_i| + \mathrm{O}(1), \tag{V.32}$$

which reduces (V.31) to

$$\left(\sum_{i=1}^{N} a_0^{1/2}|\widetilde{w}_i|\right)^2 - \left(\sum_i a_i^{1/2}|\widetilde{w}_i|\right)^2 + \mathrm{O}\left(\sum_i a_i^{1/2}|\widetilde{w}_i|\right) > 0. \tag{V.33}$$

[4]We assume here that $a_i > 0$; otherwise, the condition obviously holds.

Since $f(\cdot)$ is strictly concave and nondecreasing, there will always be some $\widetilde{\boldsymbol{w}}$ with elements sufficiently large such that $a_0 > a_i$ for all i. Consequently, we can ignore the lower-order terms and satisfy the sufficient condition for some $\widetilde{\boldsymbol{w}}$ sufficiently large, implying that there will always be cases where local minima exist using such an $f(\cdot)$.

V.G.2 Proof of Lemma 12

Assume $f(z) = \alpha z$, with $\alpha \geq 0$. Condition 1 is satisfied as a natural consequence of Theorem 2, which implicitly assumes $f(z) = 0$ but is easily extended to include any concave, nondecreasing function. So the only work is to show that it also fulfills Condition 2. For this purpose, we will assume Φ satisfies the URP; this assumption can be relaxed, but it makes the presentation more straightforward.

Using the above we can assume, without loss of generality, that any local minimum is achievable with a solution $\boldsymbol{\gamma}_*$ such that $\|\boldsymbol{\gamma}_*\|_0 \leq N$. We can be more specific; either $\|\boldsymbol{\gamma}_*\|_0 = 1$ if γ_t is the lone nonzero hyperparameter, or $\|\boldsymbol{\gamma}_*\|_0 = N$ (a non-degenerate BFS per the parlance of Section II.A). No intermediate solution is possible. This occurs as a consequence of the URP assumption and [34, Theorem 1]. So to satisfy Condition 2, we only need show that no solutions with ℓ_0 norm equal to N are local minima. The only remaining possibility will then represent the unique, global minimizer, i.e., $\boldsymbol{w}_* = \Gamma_*^{1/2}\left(\Phi\Gamma^{1/2}\right)^{\dagger} = \boldsymbol{w}_0$.

If we relax the strict inequality in (V.25) to allow for equality, then the sufficiency conditions from the previous proof become necessary conditions for a local minimum to occur at a BFS. Following the analysis from Section V.G.1 leads to the

necessary condition

$$\beta - \beta^2 + \alpha \left(\sum_{i=1}^{N} |\widetilde{w}_i| \right)^2 \geq 0, \tag{V.34}$$

where we note that $a_i = a_0 = \alpha$ given our assumptions on $f(\cdot)$. Using the definition

$$C_i(\alpha) \triangleq \frac{2|w_i|\alpha^{1/2}}{-1 + \sqrt{1 + 4\alpha w_i^2}}, \tag{V.35}$$

it follows that

$$\beta = \sum_i \alpha^{1/2} |w_i| C_i(\alpha) \tag{V.36}$$

and therefore (V.34) becomes

$$\sum_i \alpha^{1/2} |w_i| C_i(\alpha) - \left(\sum_i \alpha^{1/2} |w_i| C_i(\alpha) \right)^2 + \left(\sum_i \alpha^{1/2} |w_i| \right)^2 \geq 0. \tag{V.37}$$

To check if (V.37) holds, we note that

$$\sum_i \alpha^{1/2} |w_i| C_i(\alpha) - \sum_i \alpha w_i^2 C_i(\alpha)^2 + \sum_i \alpha |w_i|^2 = 0 \tag{V.38}$$

and that

$$-\sum_{i \neq j} \alpha |w_i||w_j| C_i(\alpha) C_i(\alpha) + \sum_{i \neq j} \alpha |w_i||w_j| \leq 0. \tag{V.39}$$

The later inequality follows because $C_i(\alpha) \geq 1$, with equality only in the limit as $\alpha \rightarrow \infty$. Together (V.38) and (V.39) dictate that (V.37) can never be true (except in the special case where $\alpha \rightarrow 0$, which will be discussed below). Since a necessary condition has been violated, no BFS with N nonzero elements can be a local minimum. That leaves

only the solution with $\|\boldsymbol{\gamma}_*\|_0 = 1$ as the unique global minimum.

Note that (V.37) cannot hold even in the limit as $\alpha \to 0$. Because Φ satisfies the URP, $\boldsymbol{t}^T\boldsymbol{t}$ will always be strictly less than $\left(\sum_{i=1}^{N} |\widetilde{w}_i|\right)^2$. This fact, when propagated through the various inequalities above, imply that (V.37) will even fail when α is unbounded.

Chapter VI

Solving the Simultaneous Sparse Approximation Problem

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 (ARD), this chapter 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.

VI.A Introduction

Previous chapters have focused on what we will refer to as the single response problem, meaning that estimation of the unknown weights $\boldsymbol{w}_{\text{gen}}$ is based on a single observed $\boldsymbol{t}$. But suppose instead 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 [33, 40, 73, 74, 75], communications [12, 25], signal processing [45, 92], and source localization [60]. Other examples that directly comply with this formulation include compressed sensing [8, 20, 102] and landmark point selection for sparse manifold learning [91]. 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 single response models of the standard form $\boldsymbol{t} = \Phi\boldsymbol{w} + \boldsymbol{\epsilon}$, the multiple response model with which we are concerned becomes

$$T = \Phi W + \mathcal{E}, \tag{VI.1}$$

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 j-th column of X while $\boldsymbol{x}_{i\cdot}$ represents the i-th row of X. Likewise, x_{ij} refers the i-th element in the j-th column of X. In the statistics literature, (VI.1) represents a multiple response model [46] or multiple output model [38]. 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.

VI.A.1 Problem Statement

To simplify matters, it is useful to introduce the notation

$$d(W) \triangleq \sum_{i=1}^{M} \mathcal{I}\left[\|\boldsymbol{w}_{i\cdot}\| > 0\right], \tag{VI.2}$$

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*.

To reiterate some definitions, we define the *spark* of a dictionary Φ as the smallest number of linearly dependent columns [17]. 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 [34], 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 [12, Lemma 1], if a matrix of generating weights W_{gen} satisfies

$$d(W_{\text{gen}}) \;<\; \left(\text{spark}(\Phi) + \text{rank}(T) - 1\right)/2 \;\leq\; \left(\text{spark}(\Phi) + \min(L, d(W_{\text{gen}})) - 1\right)/2, \tag{VI.3}$$

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

$$d(W_{\text{gen}}) < \text{spark}(\Phi) - 1, \tag{VI.4}$$

which follows from the analysis in Section II.B.2. Given that one or both of these conditions hold, then recovering W_{gen} is tantamount to solving

$$W_{\text{gen}} = W_0 \triangleq \arg\min_W \; d(W), \qquad \text{s.t. } T = \Phi W. \tag{VI.5}$$

In general, this problem is NP-hard so approximate procedures are in order. In Section VI.E.1 we will examine the solution of (VI.5) in further detail. The single response ($L = 1$) reduction of (VI.5) has been studied exhaustively [17, 29, 35, 95]. For the remainder of this paper, whenever $\mathcal{E} = 0$, we will assume that W_{gen} satisfies (VI.3) or (VI.4), 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

$$W_0(\lambda) \triangleq \arg\min_W \; \|T - \Phi W\|_{\mathcal{F}}^2 + \lambda d(W), \tag{VI.6}$$

where λ is a trade-off parameter balancing estimation quality with row sparsity. 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. Un-

fortunately, solving (VI.6) 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.[1]

From a conceptual standpoint, (VI.6) 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 $p(W) \propto \exp\left[-d(W)\right]$. 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 (VI.6) can equivalently be written as

$$W_0(\lambda) \equiv \arg\max_W p(T|W)p(W) = \arg\max_W \frac{p(T|W)p(W)}{p(T)} = \arg\max_W p(W|T) \tag{VI.7}$$

Therefore, (VI.6) can be viewed as a challenging MAP estimation task, with a posterior characterized by numerous locally optimal solutions.

VI.A.2 Summary

In Section VI.B, we discuss current methods for solving the simultaneous sparse approximation problem, all of which can be understood, either implicity or ex-

[1] Although not our focus, if the ultimate goal is compression of T, then the solution of (VI.6) may trump other concerns.

plicitly, 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) [57, 66]. Starting in Section VI.C, 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 [56]. 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 [94], 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. The per-iteration complexity relative to the other algorithms is also considered.

In Section VI.D, 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 [18] and the requirements of compressed sensing applications [102]. 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 [16].

In Section VI.E, 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 (sub-optimal) local minima are removed and we are guaranteed to solve (VI.5) 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 the variational perspective of Chapter IV that applies to the single response ($L = 1$) case. Finally, Section VI.F contains concluding remarks as well as a brief discussion of recent results applying M-SBL to large-scale neuroimaging applications.

VI.B 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 [9, 12, 60, 77, 81, 98, 97]. First, there are forward sequential selection methods based on some flavor of Matching Pursuit (MP) [61]. 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 (VI.6) [12]. A similar algorithm is analyzed in [97].

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 [10] or the LASSO [38]. Essentially, we construct a convex relaxation of (VI.6) and attempt to solve

$$W_{\text{M-BP}} = \arg\min_{W} \ \|T - \Phi W\|_{\mathcal{F}}^2 + \lambda \sum_{i=1}^{M} \|\boldsymbol{w}_{i\cdot}\|_2. \tag{VI.8}$$

This convex cost function can be globally minimized using a variety of standard optimization packages. In keeping with a Bayesian perspective, (VI.8) 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 (VI.8) as M-BP, con-

sistent with previous notation. The properties of the M-BP cost function and algorithms for its minimization have been explored in [12, 60]. Other variants involve replacing the row-wise ℓ_2 norm with the ℓ_∞ norm [98, 99] and the ℓ_1 norm [9]. 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.

Secondly, 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.[2] The optimization problem then becomes

$$W_{\text{M-Jeffreys}} = \arg\min_{W} \ \|T - \Phi W\|_{\mathcal{F}}^2 + \lambda \sum_{i=1}^{M} \log \|\boldsymbol{w}_{i\cdot}\|_2. \tag{VI.9}$$

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 $W_{\text{M-BP}}$. From an implementational standpoint, M-Jeffreys can be solved using natural, multiple response extensions of the algorithms derived in [27, 34].

Thirdly, we weigh in the M-FOCUSS algorithm derived in [12, 77, 81] based on the generalized FOCUSS algorithm of [79]. This approach employs an ℓ_p-norm-like diversity measure [14], 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

[2] The Jeffreys prior is an improper prior of the form $p(x) = 1/x$ [4].

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

$$W_{\text{M-FOCUSS}} = \arg\min_{W} \quad \|T - \Phi W\|_{\mathcal{F}}^2 + \lambda \sum_{i=1}^{M} (\|\boldsymbol{w}_{i\cdot}\|_2)^p \,. \tag{VI.10}$$

If $p \rightarrow 0$, the M-FOCUSS cost function approaches (VI.6). 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 Appendix VI.H.1. In practice, it is sometimes possible to jointly select values of p and λ such that the algorithm outperforms both 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 (VI.5).

VI.C An 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 *automatic relevance determination* (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 [57, 66]. These ideas have also been applied to Bayesian kernel machines [94]. 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 λ that is assumed to be known (the case where λ is not known is discussed briefly in Section

VIII.A). Thus, for each $\boldsymbol{t}_{\cdot j}$, $\boldsymbol{w}_{\cdot j}$ pair, we have,

$$p(\boldsymbol{t}_{\cdot j}|\boldsymbol{w}_{\cdot j}) = (2\pi\lambda)^{-N/2} \exp\left(-\frac{1}{2\lambda}\|\boldsymbol{t}_{\cdot j} - \Phi\boldsymbol{w}_{\cdot j}\|_2^2\right), \tag{VI.11}$$

which is consistent with the likelihood model implied by (VI.6) and previous Bayesian methods. Next, application of ARD involves assigning to the i-th row of W an L-dimensional Gaussian prior:

$$p(\boldsymbol{w}_{i\cdot}; \gamma_i) \triangleq \mathcal{N}(0, \gamma_i I), \tag{VI.12}$$

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

$$p(W; \boldsymbol{\gamma}) = \prod_{i=1}^{M} p(\boldsymbol{w}_{i\cdot}; \gamma_i), \tag{VI.13}$$

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 j-th column of W then becomes

$$p(\boldsymbol{w}_{\cdot j}|\boldsymbol{t}_{\cdot j}; \boldsymbol{\gamma}) = \frac{p(\boldsymbol{w}_{\cdot j}, \boldsymbol{t}_{\cdot j}; \boldsymbol{\gamma})}{\int p(\boldsymbol{w}_{\cdot j}, \boldsymbol{t}_{\cdot j}; \boldsymbol{\gamma}) d\boldsymbol{w}_{\cdot j}} = \mathcal{N}(\boldsymbol{\mu}_{\cdot j}, \Sigma), \tag{VI.14}$$

with mean and covariance given by

$$\Sigma \triangleq \text{Cov}[\boldsymbol{w}_{\cdot j}|\boldsymbol{t}_{\cdot j};\boldsymbol{\gamma}] = \Gamma - \Gamma\Phi^T\Sigma_t^{-1}\Phi\Gamma, \qquad \forall j = 1,\ldots,L,$$

$$\mathcal{M} = [\boldsymbol{\mu}_{\cdot 1},\ldots,\boldsymbol{\mu}_{\cdot L}] \triangleq \text{E}[W|T;\boldsymbol{\gamma}] = \Gamma\Phi^T\Sigma_t^{-1}T, \tag{VI.15}$$

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

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 $\text{Prob}(\boldsymbol{w}_{i\cdot} = \mathbf{0}|T;\gamma_i = 0) = 1$, ensuring that the posterior mean of the i-th 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 VI.D and VI.E provide empirical and analytical support for this claim.

VI.C.1 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 [56]. The marginal likelihood that

results is then maximized with respect to $\boldsymbol{\gamma}$, leading to the ARD-based cost function

$$\begin{aligned}\mathcal{L}(\boldsymbol{\gamma}) &\triangleq -2\log\int p(T|W)p(W;\boldsymbol{\gamma})dW \\ &= -2\log p\left(T;\boldsymbol{\gamma}\right) \\ &\equiv L\log|\Sigma_t| + \sum_{j=1}^{L}\boldsymbol{t}_{\cdot j}^T\Sigma_t^{-1}\boldsymbol{t}_{\cdot j},\end{aligned} \tag{VI.16}$$

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 [36, 37]. 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 [4, 56, 66].

There are (at least) two ways to minimize $\mathcal{L}(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$. (Section VII.B.1 briefly discusses additional possibilities.) 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 [13, 37] for covariance estimation. For the E-step, this requires computation of the posterior moments using (VI.15), while the M-step is expressed via the update rule

$$\gamma_i^{(\text{new})} = \frac{1}{L}\|\boldsymbol{\mu}_{i\cdot}\|_2^2 + \Sigma_{ii}, \qquad \forall i = 1,\ldots,M. \tag{VI.17}$$

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.

Secondly, at the expense of proven convergence, we may instead optimize (VI.16) by taking the derivative with respect to $\boldsymbol{\gamma}$, equating to zero, and forming a fixed-point equation that typically leads to faster convergence [56, 94]. Effectively, this involves replacing the M-step from above with

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

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 (VI.17). In the context of kernel regression using a complete dictionary (meaning $N = M$) and $L = 1$, use of (VI.18), along with a modified form of (VI.15),[3] 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) [94]. Similar update rules have also been effectively applied to an energy prediction competition under the guise of ARD [57]. 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.

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

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 $\lambda \to 0$ using the modified moments

$$\Sigma = \left[I - \Gamma^{1/2} \left(\Phi \Gamma^{1/2} \right)^{\dagger} \Phi \right] \Gamma, \qquad \mathcal{M} = \Gamma^{1/2} \left(\Phi \Gamma^{1/2} \right)^{\dagger} T, \tag{VI.19}$$

where $(\cdot)^{\dagger}$ denotes the Moore-Penrose pseudo-inverse. This is particularly useful if we wish to solve (VI.5).

VI.C.2 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 (VI.15), or in the noiseless case, using (VI.19).

3. Update $\boldsymbol{\gamma}$ using the EM rule (VI.17) or the faster fixed-point rule (VI.18).

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).

VI.C.3 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 [49]

$$p(\boldsymbol{t}_{\cdot j}|\boldsymbol{w}_{\cdot j}) = (\pi\lambda)^{-N} \exp\left(-\frac{1}{\lambda}\|\boldsymbol{t}_{\cdot j} - \Phi\boldsymbol{w}_{\cdot j}\|_2^2\right), \tag{VI.20}$$

where all quantities except λ 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 [60] 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.

VI.C.4 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 $\widetilde{T} \in \mathbb{R}^{N \times \mathrm{rank}(T)}$ such that $\widetilde{T}\widetilde{T}^T = TT^T$. Substituting $\widetilde{T}$ 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 [60] is $\mathrm{O}\left(M^3L^3\right)$ per iteration. While the effective value of L can be reduced (beyond what we described above) using various 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 neu-

roimaging 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.

VI.D Empirical Studies

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

VI.D.1 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 (VI.5)); this facilitates direct comparisons because discrepancies in results cannot be attributed to poor selection of trade-off 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 [18] as a useful benchmark. Additionally, it is exactly what is required in compressed sensing applications [102]. 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 $T = \Phi W_{\text{gen}}$. 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, spark$(\Phi) = N+1$ and (VI.4) is in force. Therefore, we can be sure that $W_{\text{gen}} = W_0$ 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 Figure VI.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 figures, 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}. Figure VI.1(*d*) displays sparse recovery results as the trade-off 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.[4] For

[4]Most of the theoretical study of approximate sparse representations in noise has focused on when a simpler

example, even an exhaustive search for the optimal solution to (VI.6) over all λ would likely exhibit similar performance to M-SBL in this situation.

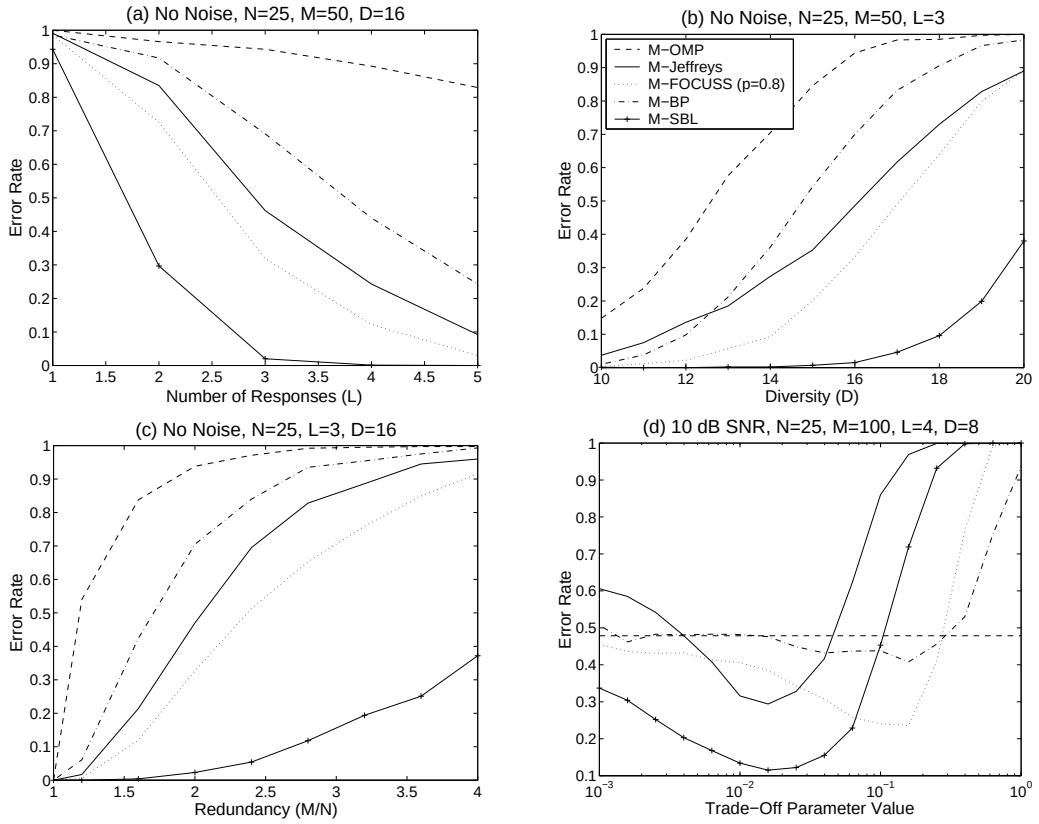

Figure VI.1: Results comparing the empirical probability (over 1000 trials) that each algorithm fails to find the sparse generating weights under various testing conditions. Plots (*a*), (*b*), and (*c*) display results as L, D and M are varied under noiseless conditions. Plot (*d*) shows results with 10dB AGWN for different values of the trade-off parameter λ.

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 Figure VI.1(*d*) with the inclusion of the the global solution to (VI.6) for different values of λ. The exhaustive search failed to locate the correct sparsity profile with an empirical probability similar

method, e.g., BP- or OMP-based, is guaranteed to provide a good solution to (VI.6), or at least exhibit a similar sparsity profile. Currently, we know of no work that examines rigorous conditions whereby the minimum of (VI.6) 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.

to M-SBL (about 0.10 using λ_{opt}), underscoring the overall difficulty of finding sparse generating weights in noisy environments.[5] Moreover, it demonstrates that, unlike in the noise-free case, the NP-hard optimization problem of (VI.6) is not necessarily guaranteed to be the most desirable solution even if computational resources are abundant.

VI.D.2 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 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-Loève 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 [17, 16].

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

[5] 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.

a Hadamard basis (other examples have been explored as well). Results are displayed in Figure VI.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 $d(\hat{W}) > d(W_{\text{gen}})$.

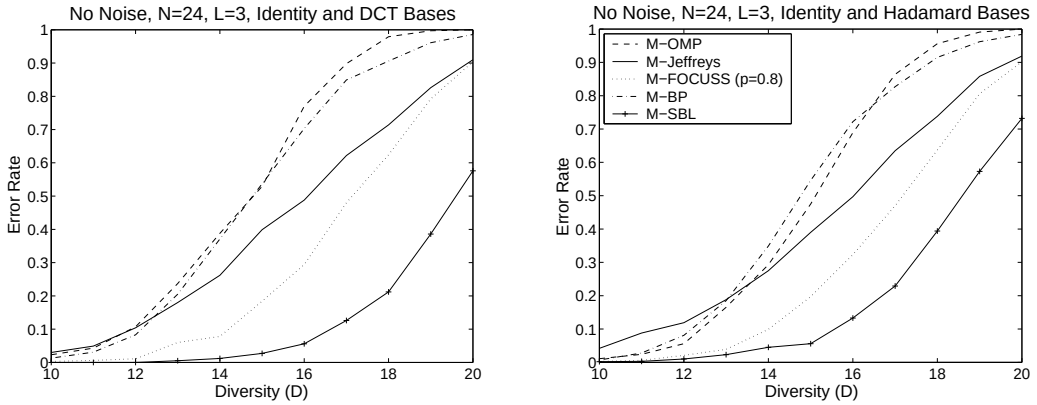

Figure VI.2: Results using pairs of orthobases with $L = 3$ and $N = 24$ while D is varied from 10 to 20. *Left*: Θ is an identity matrix and Ψ is an N-dimensional DCT. *Right*: Θ is again identity and Ψ is a Hadamard matrix.

VI.E Analysis

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

VI.E.1 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 VI.E.3. But 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 $\lambda \to 0$. This result follows from 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.

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 (VI.5).

To address this issue, we consider the fixed points of the M-SBL iterations using the modified moments from (VI.19). 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.[6] We now establish conditions whereby a unique stable fixed point exists that is also guaranteed to solve (VI.5).

Theorem 9. 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

$$\mathcal{M}^* = \mathrm{E}\left[W|T;\boldsymbol{\gamma}^*\right] = \Gamma^{*1/2}\left(\Phi\Gamma^{*1/2}\right)^{\dagger} T = W_0, \tag{VI.21}$$

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

See Appendix VI.H.2 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 (VI.5), and therefore find W_{gen}, assuming condition (VI.3) or (VI.4) 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 9.

Admittedly, the required conditions for Theorem 9 to apply are highly ideal-

[6] 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 VI.1: Verification of Theorem 9 with $N = 5$, $M = 50$, $D = L = 4$. Φ is generated as in Section VI.D.1, 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**

ized. 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 VI.B. 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 VI.D.1. Table VI.1 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 [18, 17, 29, 35, 95]. 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 non-convex 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 [57, 66, 94]. Here we have tried to make some progress in this regard.

And while Theorem 9 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 as shown in previous chapters, 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 VI.D.

VI.E.2 Geometric Interpretation

As discussed above, a significant utility of simultaneously incorporating multiple responses is the increased probability that we avoid suboptimal extrema, a benefit which exists in addition to any improvement in the effective SNR (see Section VI.E.3 below). The following simple example serves to illustrate how this happens geometrically from a Gaussian process perspective [82]. In [94], such a perspective is also considered, but only to heuristically argue why standard SBL may sometimes produce sparse representations in practice; there is no connection made to the geometry of locally minimizing solutions. In contrast, here the goal is to illustrate how a local minima

that exists when $L = 1$ can be removed when $L = 2$ or higher.

Suppose we have a single response vector $\boldsymbol{t} \in \mathbb{R}^3$ as well as a dictionary of five candidate basis vectors $\Phi = [\boldsymbol{\phi}_{\cdot 1}, \ldots, \boldsymbol{\phi}_{\cdot 5}]$. In minimizing the SBL cost function, we are linearly combining basis vectors to form a distribution that aligns itself with $\boldsymbol{t}$. As discussed in [94], SBL manipulates the covariance Σ_t of the Gaussian distribution $p(\boldsymbol{t}; \boldsymbol{\gamma})$ anchored at mean zero to maximize the likelihood of $\boldsymbol{t}$. In our simplified situation (and assuming $\lambda = 0$), we can express this covariance as

$$\Sigma_t = \Phi \Gamma \Phi^T = \sum_{j=1}^{5} \gamma_j \boldsymbol{\phi}_{\cdot j} \boldsymbol{\phi}_{\cdot j}^T, \tag{VI.22}$$

where increasing a particular γ_j causes the covariance to bulge out along the direction of the corresponding $\boldsymbol{\phi}_{\cdot j}$. Figure VI.3 depicts a scenario where the global minimum occurs with only $\gamma_4, \gamma_5 > 0$ whereas a suboptimal local minimum occurs with only $\gamma_1, \gamma_2, \gamma_3 > 0$. For convenience and ease of illustration, we have assumed that all vectors (basis and response) have been normalized to lie on the surface of a unit sphere in 3D and that there is no noise present. In (*a*), each dot labelled from 1 to 5 represents a single basis vector on the surface of this sphere while the star likewise represents $\boldsymbol{t}$. The ellipse represents a 95% confidence region for a hypothetical covariance Σ_t using only basis vectors 1, 2, and 3 (i.e., $\gamma_1, \gamma_2, \gamma_3 > 0$ while $\gamma_4 = \gamma_5 = 0$). Note that the smaller the ellipse, the higher the concentration of probability mass and the more probable any $\boldsymbol{t}$ found within.

To see why (*a*) necessarily represents a local minimum, consider slowly in-

creasing γ_4 and/or γ_5 while concurrently reducing γ_1, γ_2, and/or γ_3. This situation is represented in (*b*) where the confidence region is forced to expand, decreasing the probability density at $\boldsymbol{t}$. However, if we continue this process sufficiently far, we achieve the situation in (*c*), where we are close to the global minimum with only γ_4 and γ_5 significantly greater than zero. This latter solution places an extremely high density (in fact infinite) on $\boldsymbol{t}$ since $\boldsymbol{t}$ is essentially in the span of these two basis vectors alone. Intuitively, the local minimum occurs because we have a set of three basis vectors defining an ellipsoid with a sharp major axis that is roughly orthogonal to the plane defined by $\boldsymbol{\phi}_{\cdot 4}$ and $\boldsymbol{\phi}_{\cdot 5}$ (i.e., compare (*a*) and (*c*)).

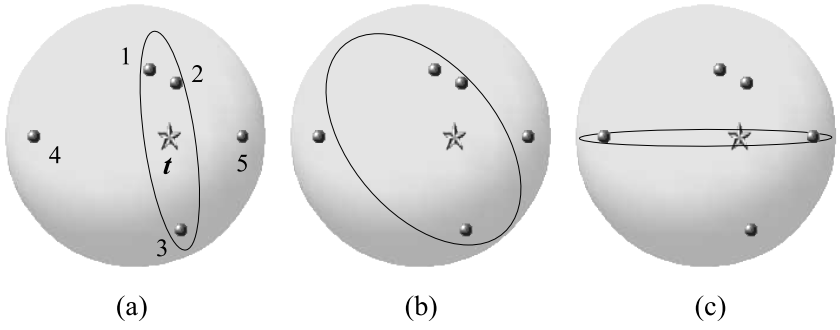

Figure VI.3: 3D example of local minimum occurring with a single response vector $\boldsymbol{t}$. (*a*): 95% confidence region for Σ_t using only basis vectors 1, 2, and 3 (i.e., there is a hypothesized 95% chance that $\boldsymbol{t}$ will lie within this region). (*b*): Expansion of confidence region as we allow contributions from basis vectors 4 and 5. (*c*): 95% confidence region for Σ_t using only basis vectors 4 and 5. The probability density at $\boldsymbol{t}$ is high in (*a*) and (*c*) but low in (*b*).

Figure VI.4 illustrates how the existence of multiple response vectors can reduce the possibility of such local minima. Here we have repeated the above analysis with the inclusion of two response vectors $\boldsymbol{t}_{\cdot 1}$ and $\boldsymbol{t}_{\cdot 2}$ that are both in the span of $\boldsymbol{\phi}_{\cdot 4}$ and

$\boldsymbol{\phi}_{\cdot 5}$ (we note that this is consistent with our assumption that each model should have the same sparsity profile). In assessing local minima, we must now consider the joint probability density of $T = [\boldsymbol{t}_{\cdot 1}, \boldsymbol{t}_{\cdot 2}]$, i.e., both $\boldsymbol{t}_{\cdot 1}$ and $\boldsymbol{t}_{\cdot 2}$ must reside in areas of significant density. Therefore in (*a*), although $\boldsymbol{t}_{\cdot 1}$ is in a region of significant density, $\boldsymbol{t}_{\cdot 2}$ is not and consequently, the likelihood of T increases from (*a*) to (*b*) and (*b*) to (*c*). In effect, the inclusion of the additional response has removed the local minimum that existed before.

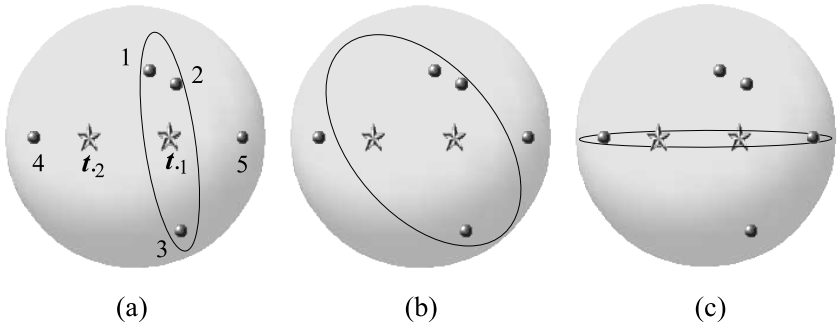

Figure VI.4: 3D example with two response vectors $\boldsymbol{t}_{\cdot 1}$ and $\boldsymbol{t}_{\cdot 2}$. (*a*): 95% confidence region for Σ_t using only basis vectors 1, 2, and 3. (*b*): Expansion of confidence region as we allow contributions from basis vectors 4 and 5. (*c*): 95% confidence region for Σ_t using only basis vectors 4 and 5. The probability of $T = [\boldsymbol{t}_{\cdot 1}, \boldsymbol{t}_{\cdot 2}]$ is very low in (*a*) since $\boldsymbol{t}_{\cdot 2}$ lies outside the ellipsoid but higher in (*b*) and highest in (*c*). Thus, configuration (*a*) no longer represents a local minimum.

VI.E.3 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 [93], 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

$$\mathcal{L}(\gamma_i) = L \log(\lambda + \gamma_i) + \frac{1}{\lambda + \gamma_i} \sum_{j=1}^{L} \left(w_{ij}^{\text{MN}}\right)^2, \tag{VI.23}$$

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

$$\gamma_i^* = \left(\frac{1}{L} \sum_{j=1}^{L} \left(w_{ij}^{\text{MN}}\right)^2 - \lambda \right)^+, \tag{VI.24}$$

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

$$\boldsymbol{\mu}_{i\cdot}^* = \boldsymbol{w}_{i\cdot}^{\text{MN}} \left(1 - \frac{L\lambda}{\|\boldsymbol{w}_{i\cdot}^{\text{MN}}\|_2^2} \right)^+, \tag{VI.25}$$

Interestingly, these weights represent a direct, multiple-response extension of those obtained using the nonnegative garrote estimator [7, 30, 93]. 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 (VI.25) 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 (non-global) 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

$$\min_{\boldsymbol{w}_{i\cdot}} \; \|\boldsymbol{w}_{i\cdot}^{\text{MN}}\|_2^2 - 2\boldsymbol{w}_{i\cdot}^T \boldsymbol{w}_{i\cdot}^{\text{MN}} + \|\boldsymbol{w}_{i\cdot}\|_2^2 + \lambda \log \|\boldsymbol{w}_{i\cdot}\|_2. \tag{VI.26}$$

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 (VI.26) to

$$\min_{\|\boldsymbol{w}_{i\cdot}\|_2} \; \|\boldsymbol{w}_{i\cdot}^{\text{MN}}\|_2^2 - 2\|\boldsymbol{w}_{i\cdot}\|_2\|\boldsymbol{w}_{i\cdot}^{\text{MN}}\|_2 + \|\boldsymbol{w}_{i\cdot}\|_2^2 + \lambda \log \|\boldsymbol{w}_{i\cdot}\|_2 \tag{VI.27}$$

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 $\|\boldsymbol{w}_{i\cdot}\|_2 = \frac{1}{2}\left(\|\boldsymbol{w}_{i\cdot}^{\text{MN}}\|_2 + \sqrt{\|\boldsymbol{w}_{i\cdot}^{\text{MN}}\|_2^2 - 2\lambda}\right)$. Unlike M-SBL, this ensures that the M-Jeffreys cost function will have $2^{\left(\sum_i \mathcal{I}[\|\boldsymbol{w}_{i\cdot}^{\text{MN}}\|_2^2 > 2\lambda]\right)}$ 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 $\|\boldsymbol{w}_{i\cdot}\|_2$. But the optimal solution does entail a threshold and asymptotic results are obtained (for the single response case) as $\|\boldsymbol{w}_{i\cdot}\|_2 \rightarrow \infty$ in [63]. Also, as $p \rightarrow 0$, 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 non-truncated 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.[7] When $p = 1$, we obtain an analogous soft-threshold operator, as discussed in [93] 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

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

problem. It also verifies that ARD methods will push unnecessary coefficients to exactly zero, as opposed to merely making them small.

VI.E.4 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 Appendix VI.H.1) 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 the ideas from Chapter IV.

To begin, we assume the standard likelihood model from (VI.11) and hypothesize a generalized sparse prior $\mathcal{H}$ that includes the M-Jeffreys prior as a special case. Specifically, for the i-th row of W we adopt the distribution:

$$p(\boldsymbol{w}_{i\cdot};\mathcal{H}) \triangleq C\left(b + \frac{\|\boldsymbol{w}_{i\cdot}\|_2^2}{2}\right)^{-(a+L/2)}, \tag{VI.28}$$

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 VI.B, 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. But 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 [56]. 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_{gen}. The goal is to model the problematic $p(W,T;\mathcal{H})$ with an approximating distribution $p(W,T;\hat{\mathcal{H}})$ that:

1. Captures the significant mass of the full posterior, which we assume reflects the region where the weights W_{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., $\mathrm{E}\left[W|T;\hat{\mathcal{H}}\right]$ can be analytically com-

puted 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.,

$$\min_{\hat{\mathcal{H}}} \int \left| p(W, T; \mathcal{H}) - p(W, T; \hat{\mathcal{H}}) \right| dW. \tag{VI.29}$$

The ultimate goal here is to chose 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 (VI.29), 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 [47]. 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 (VI.29). 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 [47]. We do not address the

alternative variational technique of forming a factorial approximation that minimizes a free-energy-based cost function [1, 3]. While these two strategies can be related in certain settings [70], 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 Chapter IV, we arrive at

$$p(\boldsymbol{w}_{i\cdot};\mathcal{H}) = \max_{\gamma_i \geq 0} \; \exp\left(-\frac{b}{\gamma_i}\right)\gamma_i^{-a}\prod_{j=1}^{L}(2\pi\gamma_i)^{-1/2}\exp\left(-\frac{w_{ij}^2}{2\gamma_i}\right). \tag{VI.30}$$

Details are contained in Appendix VI.H.3. When the maximization is dropped, we obtain the rigorous lower bound

$$p(\boldsymbol{w}_{i\cdot};\mathcal{H}) \geq p(\boldsymbol{w}_{i\cdot};\hat{\mathcal{H}}) \triangleq \exp\left(-\frac{b}{\gamma_i}\right)\gamma_i^{-a}\mathcal{N}\left(0,\gamma_i I\right), \tag{VI.31}$$

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 (VI.29) 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 (VI.31), this process conveniently allows us to remove the absolute value, leading to the simplification

$$\min_{\hat{\mathcal{H}}} \; \int p(T|W)\left|p(W;\mathcal{H}) - p(W;\hat{\mathcal{H}})\right| dW = \; \min_{\hat{\mathcal{H}}} - \int p(T|W)p(W;\hat{\mathcal{H}})dW, \tag{VI.32}$$

where each candidate hypothesis $\hat{\mathcal{H}}$ is characterized by a different $\boldsymbol{\gamma}$ vector. Using (VI.31) and (VI.11), the constituent integral of (VI.32) can be analytically evaluated as before, leading to the cost function

$$\mathcal{L}(\boldsymbol{\gamma}; a, b) \triangleq \mathcal{L}(\boldsymbol{\gamma}) + 2\sum_{i=1}^{M}\left(\frac{b}{\gamma_i} + a\log\gamma_i\right). \tag{VI.33}$$

For arbitrary $a, b > 0$, (VI.33) represents a multiple response extension of the generalized SBL cost function from [94] that, while appropriate for other circumstances, does not produce strictly sparse representations (see Chapter IV for more details). 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 trade-off 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 sensitiv-

ity 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 Appendix VI.H.1. Likewise, many other undesirable peaks exist with $d(W) > 0$. For example, M such peaks exist with $d(W) = 1$, $\binom{M}{2}$ 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 “non-global" 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 λ 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 Figure VI.1(d), where M-SBL performance is uniformly superior to M-Jeffreys for all values of λ.

VI.F Conclusions

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 [75, 76]. 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. This will be discussed further in the next Chapter.

VI.G Acknowledgements

Chapter 6, in part, is a reprint of material that has been accepted for publication as “An Empirical Bayesian Strategy for Solving the Simultaneous Sparse Approximation Problem," *IEEE Trans. Signal Processing*. I was the primary author and B.D. Rao supervised the research.

VI.H Appendix

VI.H.1 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

$$F_p(W) \triangleq \|T - \Phi W\|_{\mathcal{F}}^2 + \frac{\lambda'}{p} \sum_{i=1}^{M} \|\boldsymbol{w}_{i\cdot}\|_2^p - \frac{\lambda'}{p}, \tag{VI.34}$$

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

$$\lim_{p \to 0} \frac{1}{p} \left(\|\boldsymbol{w}_{i\cdot}\|_2^p - 1 \right) = \log \|\boldsymbol{w}_{i\cdot}\|_2, \tag{VI.35}$$

which we derive as follows. First, assume $\|\boldsymbol{w}_{i\cdot}\|_2 > 0$. Using L'Hôpitals rule, we arrive at

$$\frac{\frac{\partial \left(\|\boldsymbol{w}_{i\cdot}\|_2^p - 1 \right)}{\partial p}}{\frac{\partial p}{\partial p}} = \|\boldsymbol{w}_{i\cdot}\|_2^p \log \|\boldsymbol{w}_{i\cdot}\|_2 \longrightarrow \log \|\boldsymbol{w}_{i\cdot}\|_2. \tag{VI.36}$$

Likewise, when $\|\boldsymbol{w}_{i\cdot}\|_2 = 0$, we have

$$\frac{1}{p} \left(\|\boldsymbol{w}_{i\cdot}\|_2^p - 1 \right) = -\frac{1}{p} \longrightarrow \log \|\boldsymbol{w}_{i\cdot}\|_2 = -\infty. \tag{VI.37}$$

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

$$\lim_{p\to 0} F_p(W) = \|T - \Phi W\|_{\mathcal{F}}^2 + \lambda' \sum_{i=1}^{M} \log \|\boldsymbol{w}_{i\cdot}\|_2, \tag{VI.38}$$

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 (VI.34) has grown arbitrarily large. This discounts the quality-of-fit component, leading to the globally optimal, yet degenerate solution $W = 0$. But 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 non-global, 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 [84].

VI.H.2 Proof of Theorem 9

We say that a vector of hyperparameters is *feasible* iff

$$T = \Phi\mathcal{M} = \Phi\Gamma^{1/2}\left(\Phi\Gamma^{1/2}\right)^{\dagger} T. \tag{VI.39}$$

It is not difficult to show that any stable fixed point (SFP) of (VI.17) and (VI.19), denoted $\boldsymbol{\gamma}^*$, must be feasible. Conversely, if a fixed point is not feasible, it is unstable. Additionally, if $\boldsymbol{\gamma}$ is feasible (whether a fixed point or not), then under the stipulated conditions

$$N \geq \text{rank}\left(\Phi\Gamma\Phi^T\right) \geq \min\left(\text{spark}(\Phi) - 1, \text{rank}(\Gamma)\right) \geq D_0, \tag{VI.40}$$

where $\Phi\Gamma\Phi^T$ is the limiting value of Σ_t as $\lambda \to 0$ and $D_0 \triangleq d(W_0)$.

Now suppose we have converged to a stable fixed point $\boldsymbol{\gamma}^*$ that satisfies the condition $\text{rank}(\Phi\Gamma^*\Phi^T) = N$ (later we will address the case where $\text{rank}(\Phi\Gamma^*\Phi^T) < N$). By virtue of the convergence properties of the EM algorithm, this solution must necessarily represent a local minimum to the limiting cost function

$$\mathcal{L}(\boldsymbol{\gamma}) = L\log\left|\Phi\Gamma\Phi^T\right| + \sum_{j=1}^{L} \boldsymbol{t}_{\cdot j}^T\left(\Phi\Gamma\Phi^T\right)^{-1}\boldsymbol{t}_{\cdot j}\,. \tag{VI.41}$$

Otherwise $\boldsymbol{\gamma}^*$ will be an unstable fixed point. We will now show that no local minima, and therefore no SFPs, can exist with $\text{rank}(\Phi\Gamma^*\Phi^T) = N$.

Since we have specified that there exists a solution with D_0 nonzero rows, we

know that T is in the span of some subset of D_0 columns of Φ, denoted Φ_{D_0}. Therefore $T = \Phi_{D_0} W_{D_0}$, where W_{D_0} is the $D_0 \times L$ matrix of weigths associated with Φ_{D_0}. For convenience, let S_{D_0} be a diagonal matrix whose ii-th element equals the ℓ_2 norm of the i-th row of W_{D_0} and let $U \triangleq \Phi_{D_0} S_{D_0}$. It follows that $T = U S_{D_0}^{-1} W_{D_0}$.

Our goal will be to show that adding a contribution from these D_0 columns (by increasing the associated hyperparameters) will necessarily reduce $\mathcal{L}(\boldsymbol{\gamma})$, indicating that we cannot be at a local minimum. With this consideration in mind, we can express the cost function in the neighborhood of $\boldsymbol{\gamma}^*$ as,

$$\mathcal{L}(\alpha, \beta) \;=\; L \log |\alpha \Sigma_t^* + \beta U U^T| + \sum_{j=1}^{L} \boldsymbol{t}_{\cdot j}^T \left(\alpha \Sigma_t^* + \beta U U^T\right)^{-1} \boldsymbol{t}_{\cdot j}, \quad \text{(VI.42)}$$

where $\Sigma_t^* = \Phi \Gamma^* \Phi^T$ and α and β are parameters allowing us to balance contributions from Σ_t^* and U to the overall covariance. When $\beta = 0$, we achieve the presumed local minimum, whereas for $\beta > 0$, we are effectively adding a uniform contribution from U. Also, the second term of this expression can be simplified via

$$\begin{aligned}
\sum_{j=1}^{L} \boldsymbol{t}_{\cdot j}^T \left(\alpha \Sigma_t^* + \beta U U^T\right)^{-1} \boldsymbol{t}_{\cdot j} &= \text{tr}\left[T^T \left(\alpha \Sigma_t^* + \beta U U^T\right)^{-1} T\right] \\
&= \text{tr}\left[W_{D_0}^T S_{D_0}^{-1} U^T \left(\alpha \Sigma_t^* + \beta U U^T\right)^{-1} U S_{D_0}^{-1} W_{D_0}\right] \\
&= \text{tr}\left[U^T \left(\alpha \Sigma_t^* + \beta U U^T\right)^{-1} U S_{D_0}^{-1} S_{D_0}^2 S_{D_0}^{-1}\right] \\
&= \text{tr}\left[U^T \left(\alpha \Sigma_t^* + \beta U U^T\right)^{-1} U\right]. \qquad \text{(VI.43)}
\end{aligned}$$

where we have used the fact that $W_{D_0} W_{D_0}^T = S_{D_0}^2$ which follows from the stated orthog-

onality condition.

At any true local minimum, the following conditions must hold:

$$\left.\frac{\partial \mathcal{L}(\alpha,\beta)}{\partial \alpha}\right|_{\alpha=1,\beta=0} = 0 \qquad \left.\frac{\partial \mathcal{L}(\alpha,\beta)}{\partial \beta}\right|_{\alpha=1,\beta=0} \geq 0, \tag{VI.44}$$

where we note that the gradient with respect to β need not equal zero since β must be greater than or equal to zero. This is a reflection of the fact that all γ_i's must be greater than or equal to zero. To satisfy the first condition, it is easily shown that at the point $\alpha = 1$, $\beta = 0$,

$$\text{tr}\left[U^T \left(\Sigma_t^*\right)^{-1} U\right] = LN. \tag{VI.45}$$

With regard to the second condition, after a series of manipulations, we arrive at

$$\left.\frac{\partial \mathcal{L}(\alpha,\beta)}{\partial \beta}\right|_{\alpha=1,\beta=0} = \sum_{i=1}^{D_0} \left(L\lambda_i - \lambda_i^2\right), \tag{VI.46}$$

where λ_i is the i-th eigenvalue of $U^T \left(\Sigma_t^*\right)^{-1} U$. Because

$$\sum_{i=1}^{D_0} \lambda_i = \text{tr}(U^T(\Sigma_t^*)^{-1}U) = LN, \tag{VI.47}$$

then we have

$$\left.\frac{\partial \mathcal{L}(\alpha,\beta)}{\partial \beta}\right|_{\alpha=1,\beta=0} = L^2N - \sum_{i=1}^{D_0} \lambda_i^2 \leq L^2N - \sum_{i=1}^{D_0} \bar{\lambda}^2 = L^2N - \frac{(LN)^2}{D_0}, \tag{VI.48}$$

where $\bar{\lambda} \triangleq 1/D_0 \sum_{i=1}^{D_0} \lambda_i = LN/D_0$. Since we have assumed that $D_0 < N$, this

gradient must be negative, contradicting our assumption that we are at a local minima. Therefore, no local minima, and therefore no SFPs, can exist with rank$(\Phi\Gamma^*\Phi^T) = N$.

Now we assume that rank $\left(\Phi\Gamma^*\Phi^T\right)$ is equal to some integer N' in the interval $D_0 \leq N' < N$. Because $\boldsymbol{\gamma}^*$ must be feasible, this implies that each column of T must lie in a N'-dimensional subspace. Likewise, the D^* columns of Φ associated with nonzero elements in $\boldsymbol{\gamma}^*$, as well as the D_0 columns associated with nonzero rows of W_0 must also lie within this subspace (although there may exist some redundancy between these two sets). In general, there will be $K \geq D^*$ columns of Φ in this subspace.

As both the M-SBL cost function and update rules are rotationally invariant, we can replace T and Φ by QT and $Q\Phi$ where Q is an arbitrary orthonormal matrix. Therefore, at any fixed point $\boldsymbol{\gamma}^*$, we can always transform the original sparse recovery problem to a more restricted one based on a data matrix $T' \in \mathbb{R}^{N' \times L}$ and dictionary $\Phi' \in \mathbb{R}^{N' \times K}$. The columns of Φ not in this subspace have been pruned.

More importantly, we have a useful isomorphism in the following sense: If $\boldsymbol{\gamma}^*$ is a SFP with respect to our original problem, then the K hyperparameters associated with Φ' must comprise a SFP with respect to the reduced problem based on Φ' and T'. Therefore, any SFP with rank$(\Phi\Gamma^*\Phi^T) = N'$ must be a local minimum to the transformed problem

$$\mathcal{L}'(\boldsymbol{\gamma}) = L \log \left|\Phi'\Gamma\Phi'^T\right| + \sum_{j=1}^{L} (\boldsymbol{t}'^T_{\cdot j} \left(\Phi'\Gamma\Phi'^T\right)^{-1} \boldsymbol{t}'_{\cdot j} \,. \tag{VI.49}$$

When testing the local minimum condition, we get an analogous result as before, with

N' replacing N. In general, as long as N' is greater than D_0, we cannot have a SFP. However, when $N' = D_0$, then (VI.48) is ambiguous since $L^2N' - \frac{(LN')^2}{D_0} = 0$. In this unique situation, Φ' must be a square (i.e., $K = N' = D_0$), otherwise we violate the assumption $D_0 < \text{spark}(\Phi) - 1$. The reduced cost function (VI.49) simplifies to

$$\mathcal{L}'(\boldsymbol{\gamma}) \equiv L \log |\Gamma| + \sum_{j=1}^{L} \boldsymbol{t}'^{T}_{\cdot j} \left(\Phi'^{-T} \Gamma^{-1} \Phi'^{-1} \right) \boldsymbol{t}'_{\cdot j} \,. \tag{VI.50}$$

This expression has a single minimum at the solution $\gamma_i = 1/L \|(\boldsymbol{w}_0)_{i\cdot}\|_2^2$ for all $i = 1, \ldots, D_0$. By embedding this $\boldsymbol{\gamma}$ in the appropriate vector of zeros, we obtain the unique M-SBL stable fixed point.

VI.H.3 Derivation of the Dual Form of $p(\boldsymbol{w}_{i\cdot}; \mathcal{H})$

Any convex function $f(y) : \mathbb{R} \rightarrow \mathbb{R}$ can be represented in the dual form

$$f(y) = \sup_{\lambda} \left[\lambda y - f^*(\lambda) \right], \tag{VI.51}$$

where $f^*(\lambda)$ denotes the conjugate function [85]. 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 (VI.51), we obtain a rigorous lower bound on $f(y)$, parameterized by λ. We may then optimize over λ to find the optimal or tightest bound in a region of interest.

To accommodate the model development of Section VI.E.4, we require the

dual representation of $p(\boldsymbol{w}_{i\cdot};\mathcal{H})$. Clearly this density is not convex in $\boldsymbol{w}_{i\cdot}$; however, if we let $y_i \triangleq \|\boldsymbol{w}_{i\cdot}\|_2^2$ and define

$$f(y_i) \triangleq \log p(\boldsymbol{w}_{i\cdot};\mathcal{H}) = -(a+L/2)\log\left(b+\frac{y_i}{2}\right)+\log C, \tag{VI.52}$$

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 below.

We can find the conjugate function $f^*(\lambda_i)$ using the duality relation

$$f^*(\lambda_i) = \max_{y_i}\left[\lambda_i y_i - f(y_i)\right] = \max_{y_i}\left[\lambda_i y_i + \left(a+\frac{L}{2}\right)\log\left(b+\frac{y_i}{2}\right) - \log C\right] \tag{VI.53}$$

To find the maximizing y_i, we take the gradient of the quantity on the left and set it to zero, giving us,

$$y_i^{(\max)} = -\frac{a}{\lambda_i} - \frac{L}{2\lambda_i} - 2b. \tag{VI.54}$$

Substituting this into the expression for $f^*(\lambda_i)$ and selecting

$$C = (2\pi)^{-L/2}\exp\left[-\left(a+\frac{L}{2}\right)\right]\left(a+\frac{L}{2}\right)^{(a+L/2)}, \tag{VI.55}$$

we arrive at

$$f^*(\lambda_i) = \left(a+\frac{L}{2}\right)\log\left(\frac{-1}{2\lambda_i}\right)+\frac{L}{2}\log 2\pi - 2b\lambda_i. \tag{VI.56}$$

We are now ready to represent $f(y_i)$ in its dual form, observing first that we only need

consider maximization over $\lambda_i \leq 0$ since $f(y_i)$ is a monotonically decreasing function (i.e., all tangent lines will have negative slope). Proceeding forward, we have

$$\begin{aligned} f(y_i) &= \max_{\lambda_i \leq 0} \left[\lambda_i y_i - f^*(\lambda_i)\right] \\ &= \max_{\lambda_i \leq 0} \left[\lambda_i y_i - \left(a + \frac{L}{2}\right) \log\left(\frac{-1}{2\lambda_i}\right) - \frac{L}{2}\log 2\pi + 2b\lambda_i\right] \\ &= \max_{\gamma_i \geq 0} \left[\frac{-y_i}{2\gamma_i} - \left(a + \frac{L}{2}\right)\log\gamma_i - \frac{L}{2}\log 2\pi - \frac{b}{\gamma_i}\right], \end{aligned} \tag{VI.57}$$

where we have used the monotonically increasing transformation $\lambda_i = -1/(2\gamma_i), \gamma_i \geq 0$. The attendant dual representation of $p(\boldsymbol{w}_{i\cdot}; \mathcal{H})$ can then be obtained by exponentiating both sides of (VI.57) and substituting $y_i = \|\boldsymbol{w}_{i\cdot}\|_2^2$, giving us

$$\begin{aligned} p(\boldsymbol{w}_{i\cdot}; \mathcal{H}) &= \max_{\gamma_i \geq 0} \; (2\pi)^{-L/2} \exp\left(-\frac{\|\boldsymbol{w}_{i\cdot}\|_2^2}{2\gamma_i}\right) \exp\left(-\frac{b}{\gamma_i}\right) \gamma_i^{-(a+L/2)} \\ &= \max_{\gamma_i \geq 0} \; (2\pi\gamma_i)^{-L/2} \exp\left(-\frac{\sum_{j=1}^{L} w_{ij}^2}{2\gamma_i}\right) \exp\left(-\frac{b}{\gamma_i}\right) \gamma_i^{-a} \\ &= \max_{\gamma_i \geq 0} \; \exp\left(-\frac{b}{\gamma_i}\right) \gamma_i^{-a} \prod_{j=1}^{L} (2\pi\gamma_i)^{-1/2} \exp\left(-\frac{w_{ij}^2}{2\gamma_i}\right). \end{aligned} \tag{VI.58}$$

Chapter VII

Covariance Component Estimation with Application to Neuroelectromagnetic Source Imaging

The purpose of this chapter is twofold. First, we discuss the subject of covariance component estimation, which extends the sparsity results from previous chapters to the case where dictionary columns can be arbitrarily grouped together to compose basis matrices, each with an associated hyperparameter. A sparse collection of these basis matrices is learned to estimate the sample (data) covariance. While most of the discussion will revolve around the application to MEG/EEG source imaging, the results are actually quite general and can be applied in many other situations.

The ill-posed nature of the MEG/EEG source localization problem requires the incorporation of prior assumptions when choosing an appropriate solution out of an

infinite set of candidates. Bayesian methods are useful in this capacity because they allow these assumptions to be explicitly quantified. Recently, a number of empirical Bayesian approaches have been proposed that attempt a form of model selection by using the data to guide the search for an appropriate prior. While seemingly quite different in many respects, we apply a unifying framework based on covariance component estimation and automatic relevance determination (ARD) that elucidates various attributes of these methods and suggests directions for improvement. We also derive theoretical properties of this methodology related to convergence, local minima, and localization bias and explore connections with established algorithms.

VII.A Introduction

Magnetoencephalography (MEG) and electroencephalography (EEG) use an array of sensors to take EM field measurements from on or near the scalp surface with excellent temporal resolution. In both cases, the observed field is generated by the same synchronous, compact current sources located within the brain. Because the mapping from source activity configuration to sensor measurement is many to one, accurately determining the spatial locations of these unknown sources is extremely difficult. The relevant localization problem can be posed as follows: The measured EM signal is $B \in \Re^{d_b \times n}$, where d_b equals the number of sensors and n is the number of time points at which measurements are made. The unknown sources $S \in \Re^{d_s \times n}$ are the (discretized) current values at d_s candidate locations distributed throughout the cortical surface. These candidate locations are obtained by segmenting a structural MR scan of

a human subject and tesselating the gray matter surface with a set of vertices. B and S are related by the generative model

$$B = \mathbb{L}S + \mathcal{E}, \tag{VII.1}$$

where $\mathbb{L}$ is the so-called lead-field matrix, the i-th column of which represents the signal vector that would be observed at the scalp given a unit current source at the i-th vertex with a fixed orientation (flexible orientations can be incorporated by including three columns per location, one for each directional component). Multiple methods based on the physical properties of the brain and Maxwell's equations are available for this computation. Finally, $\mathcal{E}$ is a noise term with columns drawn independently from $\mathcal{N}(0, \Sigma_\epsilon)$.

To obtain reasonable spatial resolution, the number of candidate source locations will necessarily be much larger than the number of sensors ($d_s \gg d_b$). The salient inverse problem then becomes the ill-posed estimation of these activity or source regions, which are reflected by the nonzero rows of the source estimate matrix $\hat{S}$. Because the inverse model is underdetermined, all efforts at source reconstruction are heavily dependent on prior assumptions, which in a Bayesian framework are embedded in the distribution $p(S)$. Such a prior is often considered to be fixed and known, as in the case of minimum ℓ_2-norm approaches, minimum current estimation (MCE) [42, 100],[1] FOCUSS [12, 33], and sLORETA [71]. Alternatively, a number of empirical Bayesian

[1] MCE is another name for BP applied to the neuroelectromagnetic source localization problem.

approaches have been proposed that attempt a form of model selection by using the data to guide the search for an appropriate prior. Examples include variational Bayesian methods [87, 89], hierarchial covariance component models [28, 62, 73], and automatic relevance determination (ARD) [56, 66, 75, 76, 94]. While seemingly quite different in some respects, we present a generalized framework that encompasses many of these methods and points to connections between algorithms. We also analyze several theoretical properties of this framework related to computational/convergence issues, local minima, and localization bias. Overall, we envision that by providing a unifying perspective on these approaches, neuroelectromagnetic imaging practitioners will be better able to assess the relative strengths with respect to a particular application. This process also points to several promising directions for future research.

VII.B A Generalized Bayesian Framework for Source Localization

In this section, we present a general-purpose Bayesian framework for source localization. In doing so, we focus on the common ground between many of the methods discussed above. While derived using different assumptions and methodology, they can be related via the notion of automatic relevance determination [66] and evidence maximization [56].

To begin we involve the noise model from (VII.1), which fully defines the assumed likelihood $p(B|S)$. While the unknown noise covariance can also be parameterized and estimated from the data, for simplicity we assume that Σ_ϵ is known and

fixed. Next we adopt the following source prior for S:

$$p\left(S;\Sigma_s\right)=\mathcal{N}\left(0,\Sigma_s\right), \quad \Sigma_s=\sum_{i=1}^{d_\gamma}\gamma_i C_i, \tag{VII.2}$$

where the distribution is understood to apply independently to each column of S. Here $\boldsymbol{\gamma}=[\gamma_1,\ldots,\gamma_{d_\gamma}]^T$ is a vector of d_γ nonnegative hyperparameters that control the relative contribution of each covariance basis matrix C_i, all of which we assume are fixed and known. The unknown hyperparameters can be estimated from the data by first integrating out the unknown sources S giving

$$p(B;\Sigma_b)=\int p\left(B|S\right)p\left(S;\Sigma_s\right)dS=\mathcal{N}(0,\Sigma_b), \tag{VII.3}$$

where $\Sigma_b=\Sigma_\epsilon+\mathbb{L}\Sigma_s\mathbb{L}^T$. A hyperprior $p(\boldsymbol{\gamma})$ can also be included if desired. This expression is then maximized with respect to the unknown hyperparameters, a process referred to as type-II maximum likelihood or evidence maximization [56, 66] or restricted maximum likelihood [28]. Thus the optimization problem shifts from finding the maximum a posteriori sources given a fixed prior to finding the optimal hyperparameters of a parameterized prior. Once these estimates are obtained (computational issues will be discussed in Section VII.B.1), a tractable posterior distribution $p(S|B;\hat{\Sigma}_s)$ exists in closed form, where $\hat{\Sigma}_s=\sum_i\hat{\gamma}_i C_i$. To the extent that the 'learned' prior $p(S;\hat{\Sigma}_s)$ is realistic, this posterior quantifies regions of significant current density and point estimates

for the unknown sources can be obtained by evaluating the posterior mean

$$\hat{S} \triangleq \mathrm{E}\left[S|B;\hat{\Sigma}_s\right] = \hat{\Sigma}_s \mathbb{L}^T \left(\Sigma_\epsilon + \mathbb{L}\hat{\Sigma}_s \mathbb{L}^T\right)^{-1} B. \tag{VII.4}$$

The specific choice of the C_i's is crucial and can be used to reflect any assumptions about the possible distribution of current sources. It is this selection, rather than the adoption of a covariance component model per se, that primarily differentiates the many different empirical Bayesian approaches and points to novel algorithms for future study. The optimization strategy adopted for computing $\hat{\boldsymbol{\gamma}}$, as well as the particular choice of hyperprior $p(\boldsymbol{\gamma})$, if any, can also be distinguishing factors.

In the simplest case, use of the single component $\Sigma_s = \gamma_1 C_1 = \gamma_1 I$ leads to a regularized minimum-ℓ_2-norm solution. More interesting covariance component terms have been used to effect spatial smoothness, depth bias compensation, and candidate locations of likely activity [62, 73]. With regard to the latter, it has been suggested that prior information about a source location can be codified by including a C_i term with all zeros except a patch of 1's along the diagonal signifying a location of probable source activity, perhaps based on fMRI data [73]. An associated hyperparameter γ_i is then estimated to determine the appropriate contribution of this component to the overall prior covariance. The limitation of this approach is that we generally do not know, a priori, the regions where activity is occurring with both high spatial and temporal resolution. Therefore, we cannot reliably known how to choose an appropriate location-prior term in many situations.

The empirical Bayesian solution to this dilemma, which amounts to a form of model selection, is to try out many different (or even all possible) combinations of location priors, and determine which one has the highest Bayesian evidence, i.e., maximizes $p(B;\Sigma_b)$ [56]. For example, if we assume the underlying currents are formed from a collection of dipolar point sources located at each vertex of the lead-field grid, then we may choose $\Sigma_s = \sum_{i=1}^{d_s} \gamma_i \mathbf{e}_i \mathbf{e}_i^T$, where each $\mathbf{e}_i$ is a standard indexing vector of zeros with a '1' for the i-th element (and so $C_i = \mathbf{e}_i \mathbf{e}_i^T$ encodes a prior preference for a single dipolar source at location i).[2] This specification for the prior involves the counterintuitive addition of an unknown hyperparameter for every candidate source location which, on casual analysis may seem prone to severe overfitting (in contrast to [73], which uses only one or two fixed location priors). However, the process of marginalization, or the integrating out of the unknown sources S, provides an extremely powerful regularizing effect, driving most of the unknown γ_i to zero during the evidence maximization stage (more on this in Section VII.C). This ameliorates the overfitting problem and effectively reduces the space of possible active source locations by choosing a small relevant subset of location priors that optimizes the Bayesian evidence (hence ARD). With this 'learned' prior in place, a once ill-posed inverse problem is no longer untenable, with the posterior mean providing a good estimate of source activity. Such a procedure has been empirically successful in the context of neural networks [66], kernel machines [94], and multiple dipole fitting for MEG [75], a significant benefit to the latter being that the optimal number of dipoles need not be known a priori.

[2]Here we assume dipoles with orientations constrained to be orthogonal to the cortical surface; however, the method is easily extended to handle unconstrained dipoles.

In contrast, to model sources with some spatial extent, we can choose $C_i = \boldsymbol{\psi}_i \boldsymbol{\psi}_i^T$, where each ψ_i represents, for example, an $d_s \times 1$ geodesic neural basis vector that specifies an *a priori* weight location *and* activity extent. In this scenario, the number of hyperparameters satisfies $d_\gamma = v d_s$, where v is the number of scales we wish to examine in a multi-resolution decomposition, and can be quite large ($d_\gamma \approx 10^6$). As mentioned above, the ARD framework tests many priors corresponding to many hypotheses or beliefs regarding the locations and scales of the nonzero current activity within the brain, ultimately choosing the one with the highest evidence. The net result of this formulation is a source prior composed of a mixture of Gaussian kernels of varying scales. The number of mixture components, or the number of nonzero γ_i's, is learned from the data and is naturally forced to be small (sparse). In general, the methodology is quite flexible and other prior specifications can be included as well, such as temporal and spectral constraints. But the essential ingredient of ARD, that marginalization and subsequent evidence maximization leads to a pruning of unsupported hypotheses, remains unchanged.

We turn now to empirical Bayesian procedures that incorporate variational methods. In [89], a plausible hierarchical prior is adopted that, unfortunately, leads to intractable integrations when computing the desired source posterior. This motivates the inclusion of a variational approximation that models the true posterior as a factored distribution over parameters at two levels of the prior hierarchy. While seemingly quite different, drawing on results from [5], we can show that the resulting cost function is

exactly equivalent to standard ARD assuming Σ_s is parameterized as

$$\Sigma_s = \sum_{i=1}^{d_s} \gamma_i \mathbf{e}_i \mathbf{e}_i + \sum_{j=1}^{d_s} \gamma_{(d_s+j)} \boldsymbol{\psi}_j \boldsymbol{\psi}_j^T, \tag{VII.5}$$

and so $d_\gamma = 2d_s$. When fMRI data is available, it is incorporated into a particular inverse Gamma hyperprior on $\boldsymbol{\gamma}$, as is also commonly done with ARD methods [5]. Optimization is then performed using simple EM update rules.

In summary then, the general methods of [28, 62, 73] and [75, 76, 94] as well as the variational method of [89] are all identical with respect to their ARD-based cost functions; they differ only in which covariance components (and possibly hyperpriors) are used and in how optimization is performed as will be discussed below. In contrast, the variational model from [87] introduces an additional hierarchy to the ARD framework to explicitly model correlations between sources which may be spatially separated.[3] Here it is assumed that S can be decomposed with respect to d_z *pre-sources* via

$$S = WZ, \quad p(W; \Sigma_w) = \mathcal{N}(0, \Sigma_w), \;\; p(Z) = \mathcal{N}(0, I), \tag{VII.6}$$

where $Z \in \Re^{d_z \times n}$ represents the pre-source matrix and Σ_w is analogous to Σ_s. As stated in [87], direct application of ARD would involve integration over W and Z to find the hyperparameters $\boldsymbol{\gamma}$ that maximize $p(B; \Sigma_b)$. While such a procedure is not analytically tractable, it remains insightful to explore the characteristics of this method were we able

[3] Standard ARD can directly handle locally correlated sources as discussed above, but is not easily extended to explicitly address correlated sources which are spatially separated.

to perform the necessary computation. This allows us to relate the full model of [87] to standard ARD.

Interestingly, it can be shown that the first and second order statistics of the full prior (VII.6) and the standard ARD prior (VII.2) are equivalent (up to a constant factor), although higher-order moments will be different. However, as the number of pre-sources d_z becomes large, multivariate central-limit-theorem arguments can be used to explicitly show that the distribution of S converges to an identical Gaussian prior as ARD. So exact evaluation of the full model, which is espoused as the ideal objective were it feasible, approaches regular ARD when the number of pre-sources grows large. In practice, because the full model is intractable, a variational approximation is adopted similar to that proposed in [89]. In fact, if we assume the appropriate hyperprior on $\boldsymbol{\gamma}$, then this correlated source method is essentially the same as the procedure from [89] but with an additional level in the approximate posterior factorization for handling the decomposition (VII.6). This produces approximate posteriors on W and Z but the result cannot be integrated to form the posterior on S. However, the posterior mean of W, $\hat{W}$, is used as an estimate of the source correlation matrix (using $\hat{W}\hat{W}^T$) to substantially improve beamforming results that were errantly based on uncorrelated source models. Note however that this procedure implicitly uses the somewhat non-standard criteria of combining the posterior mean of W with the prior on Z to form an estimate of the distribution of S.

VII.B.1 Computational Issues

The primary objective of ARD is to maximize the evidence $p(B;\Sigma_b)$ with respect to $\boldsymbol{\gamma}$ or equivalently, to minimize

$$\mathcal{L}(\boldsymbol{\gamma}) \triangleq -\log p(B;\Sigma_b) \equiv n\log|\Sigma_b| + \operatorname{trace}\left[B^T\Sigma_b^{-1}B\right]. \tag{VII.7}$$

In [28], a restricted maximum likelihood (ReML) approach is proposed for this optimization, which utilizes what amounts to EM-based updates. This method typically requires a nonlinear search for each M-step and does not guarantee that the estimated covariance is positive definite. While shown to be successful in estimating a handful of hyperparameters in [62, 73], this could potentially be problematic when very large numbers of hyperparameters are present. For example, in several toy problems (with $d_{\boldsymbol{\gamma}}$ large) we have found that a fraction of the hyperparameters obtained can be negative-valued, inconsistent with our initial premise.

As such, we present three alternative optimization procedures that extend the methods from [56, 75, 89, 94] to the arbitrary covariance model discussed above and guarantee that $\gamma_i \geq 0$ for all i. Because of the flexibility this allows in constructing Σ_s, and therefore Σ_b, some additional notation is required to proceed. A new decomposition of Σ_b is defined as

$$\Sigma_b = \Sigma_\epsilon + \mathbb{L}\left(\sum_{i=1}^{d_\gamma}\gamma_i C_i\right)\mathbb{L}^T = \Sigma_\epsilon + \sum_{i=1}^{d_\gamma}\gamma_i\widetilde{\mathbb{L}}_i\widetilde{\mathbb{L}}_i^T, \tag{VII.8}$$

where $\widetilde{\mathbb{L}}_i\widetilde{\mathbb{L}}_i^T \triangleq \mathbb{L}C_i\mathbb{L}^T$ with $r_i \triangleq \text{rank}(\widetilde{\mathbb{L}}_i\widetilde{\mathbb{L}}_i^T) \leq d_b$. Also, using commutative properties of the trace operator, $\mathcal{L}(\boldsymbol{\gamma})$ only depends on the data B through the $d_b \times d_b$ sample correlation matrix BB^T. Therefore, to reduce the computational burden, we replace B with a matrix $\widetilde{B} \in \Re^{d_b \times \text{rank}(B)}$ such that $\widetilde{B}\widetilde{B}^T = BB^T$. This removes any per-iteration dependency on n, which can potentially be large, without altering that actual cost function.

By treating the unknown sources as hidden data, an update can be derived for the $(k+1)$-th iteration

$$\gamma_i^{(k+1)} = \frac{1}{nr_i}\left\|\gamma_i^{(k)}\widetilde{\mathbb{L}}_i^T\left(\Sigma_b^{(k)}\right)^{-1}\widetilde{B}\right\|_{\mathcal{F}}^2 + \frac{1}{r_i}\text{trace}\left[\gamma_i^{(k)}I - \gamma_i^{(k)}\widetilde{\mathbb{L}}_i^T\left(\Sigma_b^{(k)}\right)^{-1}\widetilde{\mathbb{L}}_i\gamma_i^{(k)}\right], \tag{VII.9}$$

which reduces to the algorithm from [89] given the appropriate simplifying assumptions on the form of Σ_s and some additional algebraic manipulations. It is also equivalent to ReML with a different effective computation for the M-step. By casting the update rules in this way and noting that off-diagonal elements of the second term need not be computed, the per-iteration cost is at most $\text{O}\left(d_b^2\sum_{i=1}^{d_\gamma} r_i\right) \leq \text{O}\left(d_b^3 d_\gamma\right)$. This expense can be significantly reduced still further in cases where different pseudo lead-field components, e.g., some $\widetilde{\mathbb{L}}_i$ and $\widetilde{\mathbb{L}}_j$, contain one or more columns in common. This situation occurs if we desire to use the geodesic basis functions with flexible orientation constraints, as opposed to the fixed orientations assumed above. In general, the linear dependence on d_γ is one of the attractive aspects of this method, effectively allowing for extremely large numbers of hyperparameters and covariance components.

The problem then with (VII.9) is not the per-iteration complexity but the convergence rate, which we have observed to be prohibitively slow in practical situations with high-resolution lead-field matrices and large numbers of hyperparameters. The only reported localization results using this type of EM algorithm are from [89], where a relatively low resolution lead-field matrix is used in conjunction with a simplifying heuristic that constrains some of the hyperparameter values. However, to avoid these types of constraints, which can potentially degrade the quality of source estimates, a faster update rule is needed. To this end, we modified the procedure of [56], which involves taking the gradient of $\mathcal{L}(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$, rearranging terms, and forming the fixed-point update

$$\gamma_i^{(k+1)} = \frac{\gamma_i^{(k)}}{n} \left\| \widetilde{\mathbb{L}}_i^T \left(\Sigma_b^{(k)} \right)^{-1} \widetilde{B} \right\|_{\mathcal{F}}^2 \left(\text{trace} \left[\widetilde{\mathbb{L}}_i^T \left(\Sigma_b^{(k)} \right)^{-1} \widetilde{\mathbb{L}}_i \right] \right)^{-1}. \tag{VII.10}$$

The complexity of each iteration is the same as before, only now the convergence rate can be orders of magnitude faster.[4] For example, given $d_b = 275$ sensors, $n = 1000$ observation vectors, and using a pseudo lead-field with 120,000 unique columns and an equal number of hyperparameters, requires approximately 5-10 mins. runtime using Matlab code on a PC to completely converge. The EM update does not converge after 24 hours. Example localization results using (VII.10) demonstrate the ability to recover very complex source configurations with variable spatial extent [76].

Unlike the EM method, one criticism of (VII.10) is that there currently exists

[4]Note that the slower EM iterations and the faster update (VII.10) need not converge to the same fixed point even when initialized at the same location. In some situations, the EM variant may be preferred since it may be more likely to reach the global minimum of $\mathcal{L}(\boldsymbol{\gamma})$, time permitting.

no proof that it represents a descent function, although we have never observed it to increase (VII.7) in practice. While we can show that (VII.10) is equivalent to iteratively solving a particular min-max problem in search of a saddle point, provable convergence is still suspect. However, a similar update rule can be derived that is both significantly faster than EM *and* is proven to produce $\boldsymbol{\gamma}$ vectors such that $\mathcal{L}\left(\boldsymbol{\gamma}^{(k+1)}\right) \leq \mathcal{L}\left(\boldsymbol{\gamma}^{(k)}\right)$ for every iteration k. Using a dual-form representation of $\mathcal{L}(\boldsymbol{\gamma})$ that leads to a more tractable auxiliary cost function, this update is given by

$$\gamma_i^{(k+1)} = \frac{\gamma_i^{(k)}}{\sqrt{n}} \left\| \widetilde{\mathbb{L}}_i^T \left(\Sigma_b^{(k)}\right)^{-1} \tilde{B} \right\|_{\mathcal{F}} \left(\text{trace} \left[\widetilde{\mathbb{L}}_i^T \left(\Sigma_b^{(k)}\right)^{-1} \widetilde{\mathbb{L}}_i \right] \right)^{-1/2}. \qquad \text{(VII.11)}$$

Details of the derivation can be found in Appendix VII.F.1.

Finally, the correlated source method from [87] can be incorporated into the general ARD framework as well using update rules related to the above; however, because all off-diagonal terms are required by this method, the iterations now scale as $\left(\sum_i r_i\right)^2$ in the general case. This quadratic dependence can be prohibitive in applications with large numbers of covariance components.

VII.B.2 Relationship with Other Bayesian Methods

As a point of comparison, we now describe how ARD can be related to alternative Bayesian-inspired approaches such as the sLORETA paradigm [71] and the iterative FOCUSS source localization algorithm [33]. The connection is most transparent when we substitute the prior covariance $\Sigma_s = \sum_{i=1}^{d_s} \gamma_i \mathbf{e}_i \mathbf{e}_i^T = \Gamma$ into (VII.10), giving

the modified update

$$\gamma_i^{(k+1)} = \left\| \gamma_i^{(k)} \ell_i^T \left(\Sigma_\epsilon + \mathbb{L}\Gamma^{(k)}\mathbb{L}^T \right)^{-1} B \right\|_2^2 \left(n R_{ii}^{(k)} \right)^{-1}, \tag{VII.12}$$

where $\Gamma \triangleq \mathrm{diag}[\boldsymbol{\gamma}]$, ℓ_i is the i-th column of $\mathbb{L}$, and

$$R^{(k)} \triangleq \Gamma^{(k)}\mathbb{L}^T \left(\Sigma_\epsilon + \mathbb{L}\Gamma^{(k)}\mathbb{L}^T \right)^{-1} \mathbb{L} \tag{VII.13}$$

is the effective resolution matrix given the hyperparameters at the k-th iteration. The j-th column of R (called a point-spread function) equals the source estimate obtained using (VII.4) when the true source is a unit dipole at location j [90].

Continuing, if we assume that initialization of ARD occurs with $\boldsymbol{\gamma}^{(0)} = \mathbf{1}$ (as is customary), then the hyperparameters produced after a *single* iteration of ARD are equivalent to computing the sLORETA estimate for standardized current density power [71] (this assumes fixed orientation constraints). In this context, the inclusion of R as a normalization factor helps to compensate for depth bias, which is the propensity for deep current sources within the brain to be underrepresented at the scalp surface [71, 75]. So ARD can be interpreted as a recursive refinement of what amounts to the non-adaptive, linear sLORETA estimate.

As a further avenue for comparison, if we assume that $R = I$ for all iterations, then the update (VII.12) is nearly the same as the FOCUSS iterations modified to simultaneously handle multiple observation vectors [12]. The only difference is the factor of n in the denominator in the case of ARD, but this can be offset by an appropri-

ate rescaling of the FOCUSS λ trade-off parameter. Therefore, ARD can be viewed in some sense as taking the recursive FOCUSS update rules and including the sLORETA normalization that, among other things, allows for depth bias compensation.

Thus far, we have focused on similarities in update rules between the ARD formulation (restricted to the case where $\Sigma_s = \Gamma$) and sLORETA and FOCUSS. We now switch gears and examine how the general ARD cost function relates to that of FOCUSS and MCE and suggests a useful generalization of both approaches. Recall that the evidence maximization procedure upon which ARD is based involves integrating out the unknown *sources* before optimizing the hyperparameters $\boldsymbol{\gamma}$. However, if some $p(\boldsymbol{\gamma})$ is assumed for $\boldsymbol{\gamma}$, then we could just as easily do the opposite: namely, we can integrate out the *hyperparameters* and then maximize S directly, thus solving the MAP estimation problem

$$\max_S \int p\left(B|S\right) p\left(S;\Sigma_s\right) p(\boldsymbol{\gamma}) d\boldsymbol{\gamma} \quad \equiv \quad \min_{\{S:S=\sum_i A_i \widetilde{S}_i\}} \|B - \mathbb{L}S\|^2_{\Sigma_\epsilon^{-1}} + \sum_{i=1}^{d_{\boldsymbol{\gamma}}} g\left(\|\widetilde{S}_i\|_{\mathcal{F}}\right), \tag{VII.14}$$

where each A_i is derived from the i-th covariance component such that $C_i = A_i A_i^T$, and $g(\cdot)$ is a function dependent on $p(\boldsymbol{\gamma})$. For example, when $p(\boldsymbol{\gamma})$ is a noninformative Jeffreys prior, then $g(x) = \log x$ and (VII.14) becomes a generalized form of the FOCUSS cost function (and reduces to the exact FOCUSS cost when $A_i = \mathbf{e}_i \mathbf{e}_i^T$ for all i). Likewise, when an exponential prior chosen, then $g(x) = x$ and we obtain a generalized version of MCE. In both cases, multiple simultaneous constraints (e.g., flexible dipole orientations, spatial smoothing, etc.) can be naturally handled and, if desired, the noise

covariance Σ_ϵ can be seamlessly estimated as well (see [27] for a special case of the latter in the context of kernel regression). This addresses many of the concerns raised in [62] pertaining to existing MAP methods. Additionally, as with ARD, source components that are not sufficiently important in representing the observed data are pruned; however, the undesirable discontinuities in standard FOCUSS or MCE source estimates across time, which previously have required smoothing using heuristic measures [42], do not occur when using (VII.14). This is because sparsity is only encouraged *between* components due to the concavity of $g(\cdot)$, but not *within* components where the Frobenius norm operator promotes smooth solutions (see [12] as well as the issues discussed in Chapter VI).

Presumably, there are a variety of ways to optimize (VII.14). One particularly straightforward and convenient method involves a simple merger of the ARD rules from Section VII.B.1 with the FOCUSS EM-framework discussed in Section I.D.1. This leads to the

$$\gamma_i^{(k+1)} = \frac{1}{nr_i} \left\| \gamma_i^{(k)} \widetilde{\mathbb{L}}_i^T \left(\Sigma_b^{(k)} \right)^{-1} \widetilde{B} \right\|_{\mathcal{F}}^{2-p}, \tag{VII.15}$$

where $p \in [0, 1]$. Upon convergence to some fixed point $\boldsymbol{\gamma}^*$, which is guaranteed, the source estimate is computed using (VII.4) as with ARD. When $p = 1$, we get generalized MCE; $p = 0$ leads to generalized FOCUSS. Any p in between maintains a balance between the two.

VII.C General Properties of ARD Methods

ARD methods maintain several attributes that make them desirable candidates for source localization. For example, unlike most MAP procedures, the ARD cost function is often invariant to lead-field column normalizations, which only affect the implicit initialization that is used or potentially the selection of the C_i's. In contrast, MCE produces a different globally minimizing solution for every normalization scheme. As such, ARD is considerably more robust to the particular heuristic used for this task and can readily handle deep current sources.

Previously, we have claimed that the ARD process naturally forces excessive/irrelevant hyperparameters to converge to zero, thereby reducing model complexity. While this observation has been verified empirically by ourselves and others in various application settings, there has been relatively little corroborating theoretical evidence, largely because of the difficulty in analyzing the potentially multimodal, non-convex ARD cost function. As such, we provide the following result:

Theorem 10. Every local minimum of the generalized ARD cost function (VII.7) can be achieved at a solution with at most $\mathrm{rank}(B)d_b \leq d_b^2$ nonzero hyperparameters. Consequently, the use of all covariance components is often not necessary to locally minimize the cost function.

The proof is based on results in Section II.C.2. Theorem 10 comprises a worst-case bound that is only tight in very nuanced situations. In practice, for any reasonable value of Σ_ϵ, the number of nonzero hyperparameters is typically much smaller than

d_b. The bound holds for all Σ_ϵ, including $\Sigma_\epsilon = 0$, indicating that some measure of hyperparameter pruning, and therefore covariance component pruning, is built into the ARD framework irrespective of the noise-based regularization. Moreover, the number of nonzero hyperparameters decreases monotonically to zero as Σ_ϵ is increased. And so there is always some $\Sigma_\epsilon = \Sigma'_\epsilon$ sufficiently large such that all hyperparameters converge to exactly zero. Therefore, we can be reasonable confident that the pruning mechanism of ARD is not merely an empirical phenomena. Nor is it dependent on a particular sparse hyperprior, since the ARD cost from (VII.7) implicitly assumes a flat (uniform) hyperprior.

The number of observation vectors n also plays an important role in shaping ARD solutions. Increasing n has two primary benefits: (i) it facilitates convergence to the global minimum (as opposed to getting stuck in a suboptimal extrema) and (ii), it improves the quality of this minimum by mitigating the effects of noise (Section VI.E discusses these issues in more detail). With perfectly correlated (spatially separated) sources, primarily only the later benefit is in effect. For example, with low noise and perfectly correlated sources, the estimation problem reduces to an equivalent problem with $n = 1$, so the local minima profile of the cost function does not improve with increasing n. Of course standard ARD can still be very effective in this scenario [76]. In contrast, geometric arguments can be made to show that uncorrelated sources with large n offer the best opportunity for local minima avoidance. However, when strong correlations are present as well as high noise levels, the method of [87] (which explicitly attempts to model correlations) could offer a worthwhile alternative, albeit at a high

computational cost.

Further theoretical support for ARD is possible in the context of localization bias assuming simple source configurations. For example, substantial import has been devoted to quantifying localization bias when estimating a single dipolar source. Recently it has been shown, both empirically [71] and theoretically [90], that sLORETA has zero location bias under this condition at high SNR. Viewed then as an iterative enhancement of sLORETA as described in Section VII.B.2, the question naturally arises whether ARD methods retain this desirable property. In fact, it can be shown that this is indeed the case in two general situations. We assume that the lead-field matrix $\mathbb{L}$ represents a sufficiently high sampling of the source space such that any active dipole aligns with some lead-field column. Unbiasedness results can also be shown in the continuous case for both sLORETA and ARD, but the discrete scenario is more straightforward and of course more relevant to any practical task.

Theorem 11. Assume that Σ_s includes (among others) d_s covariance components of the form $C_i = \mathbf{e}_i \mathbf{e}_i^T$. Then in the absence of noise (high SNR), ARD has provably zero localization bias when estimating a single dipolar source, regardless of the value of n.

Theorem 12. Let Σ_s be constructed as above and assume the noise covariance matrix Σ_ϵ is known up to a scale factor. Then given a single dipolar source, in the limit as n becomes large the ARD cost function is unimodal, and a source estimate with zero localization bias achieves the global minimum. Additionally, for certain reasonable lead-field matrices and covariance components, this global minimum is unique.

We focus here on Theorem 12; the argument for Theorem 11 emerges as a special case. It also easily follows from both Theorem 5 and Theorem 9, which are also relevant to the analysis in this section and can possibly be generalized in this context. To begin, we require an intermediate lemma which is proven in Appendix VII.F.2.

Lemma 13. If the outerproduct BB^T can be expressed as some non-negative linear combination of the available covariance components $\Sigma_\epsilon, \widetilde{\mathbb{L}}_1\widetilde{\mathbb{L}}_1^T, \ldots, \widetilde{\mathbb{L}}_{d_\gamma}\widetilde{\mathbb{L}}_{d_\gamma}^T$, then the ARD cost function is unimodal and $\Sigma_b = n^{-1}BB^T$ at any minimizing solution.

As n becomes large, the conditions of Theorem 12 stipulate that $n^{-1}BB^T$ will converge to $\beta\ell_a\ell_a^T + \Sigma_\epsilon$, where ℓ_a denotes the column of $\mathbb{L}$ associated with the active dipole and $\beta > 0$ is some constant. Because we are assuming that Σ_ϵ and $\ell_a\ell_a^T$ are available covariance components, then the above lemma implies that at any minimum $\Sigma_b = \beta\ell_a\ell_a^T + \Sigma_\epsilon$. The hyperparameter vector $\boldsymbol{\gamma}^*$ characterized by all zeroes except for a value of β in the element corresponding to $\ell_a\ell_a^T$ achieves this result (there will also be a nonzero hyperparameter associated with the Σ_ϵ component). When we then proceed to compute $\hat{S}$ via (VII.4), all elements will be zero except for the row corresponding with the active dipole, hence zero localization bias.

Additionally, for certain reasonable lead-field and covariance components, $\boldsymbol{\gamma}^*$ will be the unique hyperparameter vector such that $\Sigma_b = \beta\ell_a\ell_a^T + \Sigma_\epsilon$, essentially guaranteeing that ARD will produce an unbiased estimate provided a proper descent algorithm is used. For example, if $\Sigma_s = \sum_i \mathbf{e}_i\mathbf{e}_i^T$, $\Sigma_\epsilon \propto I$, and $d_s < (d_b + 1)\, d_b/2$ (i.e., the number of degrees of freedom in a $d_b \times d_b$ covariance matrix), then $\boldsymbol{\gamma}^*$ will be the unique mini-

mizer.[5] Also, given the very particular ill-conditioned structure of $\mathbb{L}$, and therefore any derived $\widetilde{\mathbb{L}}_i$, it is very likely that much looser restrictions will lead to uniqueness as well. This is because it is very difficult for any combination of lead-field columns to exactly match the contributions of both $\Sigma_\epsilon \propto I$ and $\ell_a \ell_a^T$ to the overall covariance.

While theoretical results of this kind are admittedly very limited, other iterative Bayesian schemes in fact fail to exhibit similar performance. For example, all of the MAP-based focal algorithms we are aware of, including FOCUSS and MCE methods, provably maintain a localization bias in the general setting, although in particular cases they may not exhibit one. (Also, because of the additional complexity involved, it is still unclear whether the correlated source method of [87] satisfies a similar result.) When we move to more complex source configurations (e.g., multiple dipoles), theoretical results are not available; however, empirical tests provide a useful means of comparison. For example, given a $275 \times 40,000$ lead-field matrix constructed from an MR scan and assuming fixed orientation constraints and a spherical head model, ARD using $\Sigma_s = \mathrm{diag}(\boldsymbol{\gamma})$ and $n = 1$ consistently maintains zero empirical localization bias when estimating up to 15-20 dipoles, while sLORETA starts to show a bias with only a few.

MCE (or BP in the parlance of previous chapters) and FOCUSS have been compared with ARD as well; however, in both cases they are able to resolve fewer than half the dipoles that ARD is capable of [75]. With FOCUSS (and p small), this is because of a greater tendency to converge to local minima. With MCE, this is because

[5] This assumes a very minor technical condition on $\mathbb{L}$ to circumvent some very contrived situations.

the global solution is often not sufficiently sparse. This latter result is not surprising since the lead-field matrix $\mathbb{L}$ is well known to have many columns that are almost perfectly correlated, which can make it very difficult for MCE to be effective (e.g., see the discussion in Appendix II.F.2). Additionally, the prevalence of deep sulci implies that current sources with opposing dipole moments may exist that will exhibit relatively high ℓ_1 norm. Consequently, the MCE solution may not resemble the true sparse distribution of dipoles.

VII.D Discussion

The efficacy of modern empirical Bayesian techniques and variational approximations make them attractive candidates for source localization. However, it is not always transparent how these methods relate nor which should be expected to perform best in various situations. By developing a general framework around the notion of ARD, deriving several theoretical properties, and showing connections between algorithms, we hope to bring an insightful perspective to these techniques.

VII.E Acknowledgements

This chapter, in part, is a reprint of material that will be published as "Analysis of Empirical Bayesian Methods for Neuroelectromagnetic Source Localization," *Advances in Neural Information Processing Systems 19* (2007). I was the primary author, R.R. Ramírez and J.A. Palmer contributed to the research, and S. Makeig and B.D.

Rao supervised the research.

VII.F Appendix

VII.F.1 Derivation of Alternative Update Rule

In this section, we reexpress the ARD-based cost function $\mathcal{L}(\boldsymbol{\gamma})$ in a more convenient form leading to the update rule (VII.11) and a proof that $\mathcal{L}\left(\boldsymbol{\gamma}^{(k+1)}\right) \leq \mathcal{L}\left(\boldsymbol{\gamma}^{(k)}\right)$ at each iteration. In fact, a wide variety of alternative, convergent update rules can be developed by decoupling $\mathcal{L}(\boldsymbol{\gamma})$ using auxiliary functions and an additional set of parameters that can be easily optimized, along with γ, using coordinate descent. While applicable in the general covariance component setting discussed in this chapter, these results also lead to useful algorithms for finding sparse representations in the context of previous chapters.

To begin, the data fit term can be expressed as

$$\operatorname{trace}\left[\widetilde{B}^T \Sigma_b^{-1} \widetilde{B}\right] = \min_X \frac{1}{\lambda} \left\| \widetilde{B} - \sum_{i=1}^{d_\gamma} \widetilde{\mathbb{L}}_i X_i \right\|_{\mathcal{F}}^2 + \sum_{i=1}^{d_\gamma} \gamma_i^{-1} \|X_i\|_{\mathcal{F}}^2, \tag{VII.16}$$

where $X = \left[X_1^T, \ldots, X_{d_\gamma}^T\right]^T$. Likewise, because the log-determinant term of $\mathcal{L}(\boldsymbol{\gamma})$ is concave in $\boldsymbol{\gamma}$ (see Lemma 3), it can be expressed as an minimum over upper-bounding hyperplanes via

$$n \log |\Sigma_b| = \min_{\boldsymbol{z}} \boldsymbol{z}^T \boldsymbol{\gamma} - g^*(\boldsymbol{z}), \tag{VII.17}$$

where $g^*(\boldsymbol{z})$ is the concave conjugate of $\log |\Sigma_b|$. For our purposes below, we will never

actually have to compute $g^*(\boldsymbol{z})$.

Dropping the minimizations and combining terms from (VII.16) and (VII.17) leads to the modified cost function

$$\begin{aligned}\mathcal{L}(\boldsymbol{\gamma}, X, \boldsymbol{z}) &= \frac{1}{\lambda}\left\|\widetilde{B} - \sum_{i=1}^{d_\gamma} \widetilde{\mathbb{L}}_i X_i\right\|_{\mathcal{F}}^2 + \sum_{i=1}^{d_\gamma} \gamma_i^{-1}\|X_i\|_{\mathcal{F}}^2 + \boldsymbol{z}^T\boldsymbol{\gamma} - g^*(\boldsymbol{z}) \\ &= \frac{1}{\lambda}\left\|\widetilde{B} - \sum_{i=1}^{d_\gamma} \widetilde{\mathbb{L}}_i X_i\right\|_{\mathcal{F}}^2 + \sum_{i=1}^{d_\gamma} \left[\gamma_i^{-1}\|X_i\|_{\mathcal{F}}^2 + z_i\gamma_i\right] - g^*(\boldsymbol{z}), \quad \text{(VII.18)}\end{aligned}$$

where by construction

$$\mathcal{L}(\boldsymbol{\gamma}) = \min_X \min_{\boldsymbol{z}} \mathcal{L}(\boldsymbol{\gamma}, X, \boldsymbol{z}). \quad \text{(VII.19)}$$

It is straightforward to show that if $\{\boldsymbol{\gamma}^*, X^*, \boldsymbol{z}^*\}$ is a local minimum to $\mathcal{L}(\boldsymbol{\gamma}, X, \boldsymbol{z})$, then $\boldsymbol{\gamma}^*$ is a local minimum to $\mathcal{L}(\boldsymbol{\gamma})$. Likewise, if $\{\boldsymbol{\gamma}^*, X^*, \boldsymbol{z}^*\}$ is a global minimum of $\mathcal{L}(\boldsymbol{\gamma}, X, \boldsymbol{z})$, then $\boldsymbol{\gamma}^*$ globally minimizes $\mathcal{L}(\boldsymbol{\gamma})$.

Since direct optimization of $\mathcal{L}(\boldsymbol{\gamma})$ may be difficult, we can instead iteratively optimize $\mathcal{L}(\boldsymbol{\gamma}, X, \boldsymbol{z})$ via coordinate descent over $\boldsymbol{\gamma}$, X, and $\boldsymbol{z}$. In each case, when two are held fixed, the third can be globally minimized in closed form. (In the case of $\boldsymbol{\gamma}$ this occurs because each γ_i can be optimized independently given fixed values for X and $\boldsymbol{z}$.) This ensures that each cycle will reduce $\mathcal{L}(\boldsymbol{\gamma}, X, \boldsymbol{z})$, but more importantly, will reduce $\mathcal{L}(\boldsymbol{\gamma})$ (or leave it unchanged if a fixed-point or limit cycle is reached). The associated update rules from this process are as follows.

With $\boldsymbol{z}$ and X fixed, the minimizing $\boldsymbol{\gamma}$ is obtained by solving

$$\nabla_{\boldsymbol{\gamma}} \mathcal{L}(\boldsymbol{\gamma}, X, \boldsymbol{z}) = 0. \tag{VII.20}$$

This leads to the update

$$\gamma_i^{\text{new}} = \frac{\|X_i\|_{\mathcal{F}}}{\sqrt{z_i}}. \tag{VII.21}$$

The optimal X (with $\boldsymbol{\gamma}$ and $\boldsymbol{z}$ fixed) is just the standard weighted minimum-norm solution given by

$$X_i^{\text{new}} = \gamma_i \widetilde{\mathbb{L}}_i^T \Sigma_b^{-1} \widetilde{B} \tag{VII.22}$$

for each i. Finally, the minimizing $\boldsymbol{z}$ equals the slope at the current $\boldsymbol{\gamma}$ of $n \log |\Sigma_b|$. As such, we have

$$z_i^{\text{new}} = \nabla_{\gamma_i} n \log |\Sigma_b| = n \text{trace} \left[\widetilde{\mathbb{L}}_i^T \Sigma_b^{-1} \widetilde{\mathbb{L}}_i \right]. \tag{VII.23}$$

By merging these three rules into a single $\boldsymbol{\gamma}$ update, we arrive at the exact ARD iteration given by (VII.11). Moreover, by using a slightly different set of auxiliary functions, other updates (e.g., the standard EM rule), can be easily derived. Also, this process can be used to show that the fixed-point update (VII.10) is iteratively solving a particular min-max problem in search of a saddle point. Unfortunately though, proving convergence in this context is more difficult.

VII.F.2 Proof of Section VII.C Lemma

To facilitate the analysis below, we define a $d_b \times \text{rank}(B)$ matrix $\widetilde{B}$ such that $\widetilde{B}\widetilde{B}^T = n^{-1}BB^T$. Now suppose we are at some local minimum of $\mathcal{L}(\boldsymbol{\gamma})$ characterized by the covariance Σ_b^*. In the neighborhood of Σ_b^*, the ARD cost function can be written as

$$\mathcal{L}(\alpha,\beta) = \log\left|\alpha\widetilde{B}\widetilde{B}^T + \beta\Sigma_b^*\right| + \text{trace}\left[\widetilde{B}\widetilde{B}^T\left(\alpha\widetilde{B}\widetilde{B}^T + \beta\Sigma_b^*\right)^{-1}\right], \tag{VII.24}$$

where at the presumed local minimum, $\alpha = 0$ and $\beta = 1$. In contrast, by increasing α, we allow a contribution from $\widetilde{B}\widetilde{B}^T$ to the overall covariance. That such a term exists is possible by the assumption that $n^{-1}BB^T$, and therefore $\widetilde{B}\widetilde{B}^T$, can be represented via a nonnegative linear combination of available covariance components. Note that for simplicity, we will henceforth assume that the sample covariance $n^{-1}BB^T$ is full rank, and therefore any Σ_b^* must be too. However, the general case can be handled as well with a little extra effort.

If Σ_b^* is a true local minimum of the original cost $\mathcal{L}(\boldsymbol{\gamma})$, then it must also locally minimize $\mathcal{L}(\alpha,\beta)$, necessary conditions for which are

$$\left.\frac{\partial\mathcal{L}(\alpha,\beta)}{\partial\alpha}\right|_{\alpha=1,\beta=0} = 0 \qquad \left.\frac{\partial\mathcal{L}(\alpha,\beta)}{\partial\beta}\right|_{\alpha=1,\beta=0} \geq 0, \tag{VII.25}$$

where the gradient with respect to α need not actually equal zero since α must be greater than or equal to zero. After some manipulations, the first condition is equivalent to the

requirement

$$\text{trace}\left[\widetilde{B}\widetilde{B}^T\left(\Sigma_b^*\right)^{-1}\right] = d_b. \tag{VII.26}$$

Likewise, the second condition is tantamount to the inequality

$$\text{trace}\left[\widetilde{B}\widetilde{B}^T\left(\Sigma_b^*\right)^{-1}\right] - \text{trace}\left[\widetilde{B}\widetilde{B}^T\left(\Sigma_b^*\right)^{-1}\widetilde{B}\widetilde{B}^T\left(\Sigma_b^*\right)^{-1}\right] \geq 0. \tag{VII.27}$$

Using the eigendecomposition $\widetilde{B}^T\left(\Sigma_b^*\right)^{-1}\widetilde{B} = V\Lambda V^T$, this expression reduces to

$$\sum_{i=1}^{d_b}\lambda_i \geq \sum_{i=1}^{d_b}\lambda_i^2, \tag{VII.28}$$

where the summation is over the d_b eigenvalues defined above. Also, because

$$\text{trace}\left[\widetilde{B}\widetilde{B}^T\left(\Sigma_b^*\right)^{-1}\right] = \sum_{i=1}^{d_b}\lambda_i, \tag{VII.29}$$

the lefthand side of (VII.28) equals d_b. The only way then to satisfy this inequality is if $\lambda_i = 1$ for all $i = 1, \ldots, d_b$. This is why we chose to reparameterize via $\widetilde{B}$, thus forcing the number of eigenvalues to equal their sum. Furthermore, this implies that

$$\widetilde{B}^T\left(\Sigma_b^*\right)^{-1}\widetilde{B} = VV^T = I. \tag{VII.30}$$

Solving (VII.30) gives $\Sigma_b^* = \widetilde{B}\widetilde{B}^T = n^{-1}BB^T$, completing the proof.

Chapter VIII

Practical Issues and Extensions

This chapter discusses performance issues related to determining the trade-off parameter λ as well as convergence. It concludes by deriving a fast means of learning the dictionary Φ under the assumption that it is orthonormal. When combined with a pre-whitening step, this can be used to implement a robust, noisy version of independent component analysis (ICA).

VIII.A Estimating the Trade-Off Parameter λ

If we already have access to some reliable estimate for λ, then it can naturally be incorporated into any of the update rules described in this thesis. When no such luxury exists, it would be desirable to have some alternative at our disposal. As one option, λ estimation can be incorporated into the empirical Bayesian framework as originally discussed in [56, 94]. This involves replacing the M-step with a joint maximization over λ and the hyperparameters $\boldsymbol{\gamma}$. Because of decoupling, the $\boldsymbol{\gamma}$ update remains unchanged,

while we must include, e.g., for the fast version of the multiple response SBL algorithm from Section VI.C, the λ update

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

This equation generalizes (or reduces) to other SBL-based algorithms.

A word of caution is in order with respect to λ 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 $M \geq N$, λ 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 $\mathcal{L}(\boldsymbol{\gamma}, \lambda)$ can be minimized with $\lambda = 0$. For example, if we choose the dictionary $\Phi' = [\Phi\ I]$, then λ 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 λ 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 biased-low estimates of λ. 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.

VIII.B Implementational and Convergence Issues

Per-iteration complexity of various algorithms has been addressed in Sections VI.C.4 and VII.B.1, but several important outstanding issues warrant further discussion. First, while standard EM implementations exist for many of the Bayesian methods discussed in this thesis, not all have been provably shown to satisfy the Global Convergence Theorem (GCT) of Zangwill [104]. While all are proven descent functions in the sense that every iteration is guaranteed to reduce (or leave unchanged) the associated cost,[1] there is no assurance that the fixed points that ensue will be locally minimizing solutions (or even saddle points) of the underlying cost function.

There are exceptions. The FOCUSS algorithm using $p < 1$ has been explicitly proven to satisfy all the GCT conditions [78]. However, the $p = 1$ case has not been addressed, although this can be handled using standard alternatives like linear programming, so the issue is less relevant. Likewise, SBL has also not been analyzed in this sense of global convergence. The difficulty in doing so arises because the SBL cost function and associated EM update rules do not satisfy (at least they have not been proven to satisfy thus far) certain important properties that have been used in the past to guarantee EM convergence to local minima (or in rare cases, a saddle point). For example, if the likelihood (or posterior for MAP estimation) is not differentiable, or if the solution does not lie in the interior of the parameter space, the GCT conditions for EM algorithms do not seem to be covered by existing proofs [6, 103]. While we have not observed any problems in practice, this is an issue to consider.

[1] There exist other, stricter definitions of descent functions.

The heuristically-derived fast version of SBL proposed by Tipping [94] (and MacKay earlier) is a different story. Here it has not even been proven that each iteration will always reduce or leave unchanged the SBL cost, although we have never observed an exception in empirical studies. We can show that these updates are equivalent to alternating a min-max procedure to find a saddle point of a particular auxiliary function, but this perspective has not yet led to any performance guarantees. A more serious problem, perhaps, is that it does appear that fast SBL can sometimes converge to fixed points that are not local minima (or even saddle points) of the SBL cost. Apparently, some hyperparameters are pushed to zero too fast during application of the update rules. Once a hyperparameter hits zero, or close enough relative to machine precision or some other thresholding criteria, it will remain fixed forever unless some heuristic is developed to reintroduce non-zero values. But even this may not help if there exists undue pressure to push hyperparameters to zero at inopportune times.

Interestingly, this problem seems to be most pronounced (in the cases we have tested) in a noiseless setting when some nonzero elements of $\boldsymbol{w}_0$ are small and random dictionaries are used. (The EM version of SBL works much better in this case.) However, on large MEG or EEG leadfield dictionaries (e.g., 275 rows $\times 120,000$ columns) this issue does not seem to arise.

We have derived other fast versions of SBL using convex analysis that are guaranteed to reduce the cost at every iteration unlike the fast Tipping algorithm (see Section VII.B.1). But these methods, while appealing as descent methods, can still sometimes converge to fixed points that do not minimize the SBL cost. Regardless, this

is an area that warrants further study.

VIII.C Learning Orthogonal Transforms for Promoting Sparsity

Given a set of L data vectors, the goal is to find an orthonormal matrix Φ that promotes sparse representations in the transform domain. Such a procedure is useful in many applications such as sparse coding, image denoising, and compressed sensing, where the orthonormality restriction is essential to avoid inflating noise or inconsequential components. In practice, it is customary to either use a fixed, wavelet-based transform [15] or to run a general ICA algorithm to convergence followed by a heuristic orthonormalization step [44]. In contrast, we derive a novel algorithm that, like ICA-based methods, adaptively learns a sparsity-inducing transformation; however, with our approach the orthonormality constraint is embedded in the actual cost function and enforced at each iteration. The resulting update rules are provably convergent and computationally very efficient. This method compares well with wavelet and sparse code shrinkage methods in an image denoising application. Additionally, if we first whiten the data T, then this method reduces to a robust means of performing noisy ICA assuming super-Gaussian sources.

The generative model for this problem is

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

where $\Phi \in \mathbb{R}^{N \times N}$ satisfies $\Phi^T \Phi = I$ but is otherwise unknown, $W \in \mathbb{R}^{N \times L}$ is the

unknown sources which are assumed to be sparse in some sense (e.g., super-Gaussian), T is the observed mixtures, and $\mathcal{E}$ is unknown corrupting noise.

The actual optimization problem we propose to solve is

$$\min_{X,\Phi} \sum_{i=1}^{N} \sum_{j=1}^{L} f\left(x_{ij}; \lambda\right) \quad \text{s.t. } \Phi^T \Phi = I, \;\; X = \Phi^T T, \tag{VIII.3}$$

where

$$f(z;\lambda) = \begin{cases} z^2/\lambda + \log \lambda, & z^2 \in [0, \lambda] \\ 2\log|z| + 1, & z^2 \in [\lambda, \infty). \end{cases} \tag{VIII.4}$$

A plot of $f(z)$ is provided in Figure VIII.1.

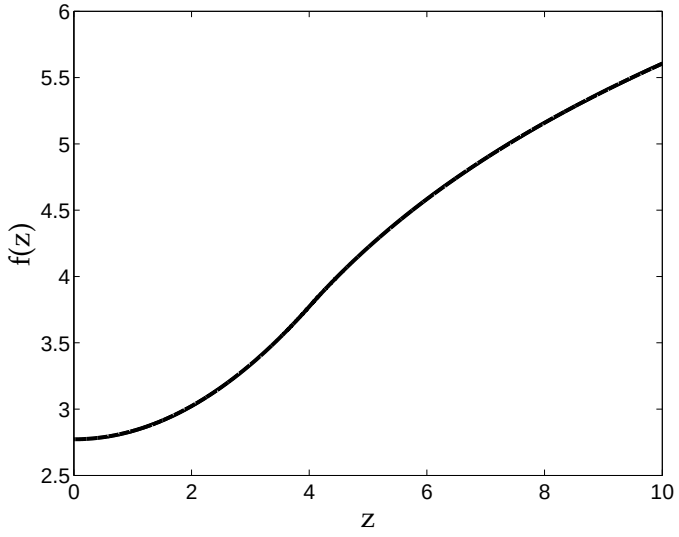

Figure VIII.1: Plot of $f(z; \lambda = 16)$. The inflection point occurs at $z = \sqrt{\lambda} = 4$.

This function encourages many of the elements of X to go below $\sqrt{\lambda}$, where $f(\cdot)$ is reduced quadratically. However, the constraint $\Phi^T \Phi = I$ will necessarily force some elements above $\sqrt{\lambda}$, but this only incurs a logarithmic penalty. The net result is many

small values below $\sqrt{\lambda}$ and a few large values above it as desired.

To form an estimate of W, for denoising or coding purposes, etc., a thresholding operator can be applied to X. For example, from Section VI.E.3 we know that the optimal SBL threshold in this case would be

$$\hat{w}_{ij} = x_{ij}\left(1 - \frac{\lambda}{x_{ij}^2}\right)^+, \tag{VIII.5}$$

which equals the non-negative garrote estimator [7]. This operator has been endorsed for wavelet denoising [26, 30].

The update rules for minimizing (VIII.3), which can be obtained using the EM algorithm in an empirical Bayesian framework, are surprisingly simple. First, a suitable initialization is chosen for the dictionary, $\hat{\Phi} := \Phi'$ which gives $\hat{X} = \hat{\Phi}^T T$. Then $\hat{W}$ is computed using (VIII.5). For the dictionary update, we have

$$\hat{\Phi} = UV^T, \tag{VIII.6}$$

where USV^T is the SVD of $T\hat{W}^T$. This value of $\hat{\Phi}$ solves the constrained optimization problem[2]

$$\min_{\hat{\Phi}} \|T - \hat{\Phi}\hat{W}\|_{\mathcal{F}} \quad \text{s.t. } \hat{\Phi}^T\hat{\Phi} = I. \tag{VIII.7}$$

This process is iterated until convergence. Note that these update rules are guaranteed to reduce (VIII.3) at each step.

[2] See [32] for the proof.

Preliminary results using this method on image data are quite promising. Moreover, the algorithm is quite fast, with each iteration (which uses all of the data unlike some ICA methods) incurring only a $O(N^2L)$ complexity cost, which is linear in the number of samples.

VIII.D Acknowledgements

The material in VIII.C was briefly addressed in an article published as “Finding Sparse Representations in Multiple Response Models via Bayesian Learning," *Workshop on Signal Processing with Adaptive Sparse Structured Representations* (2005). I was the primary researcher and B.D. Rao supervised the research.

Chapter IX

Conclusion

Applications of sparsity continue to grow in signal and image processing, functional brain imaging, neural modelling, and machine learning. While a diverse set of Bayesian tools exist for finding sparse representations from overcomplete feature dictionaries, the most common and well-understood methods involve simple MAP estimation using a fixed, sparsity-inducing prior (e.g., OMP, BP, and FOCUSS). In contrast, the relatively under-utilized empirical Bayesian approaches, which adopt a flexible, parameterized prior to encourage sparsity, show tremendous promise but lag behind in terms of solid theoretical justification and rigorous analysis in the context of sparse estimation problems. Nor have all the connections between various families of Bayesian algorithms been adequately fleshed out.

We have addressed these issues on a variety of fronts, particularly with respect to sparse Bayesian learning (SBL), an empirical Bayesian framework built upon the notion of automatic relevance determination (ARD). First, we have proven several

results about the associated SBL cost function that elucidate its general behavior and provide solid theoretical justification for using it to find maximally sparse representations. Specifically, we show that the global SBL minimum is always achieved at the maximally sparse solution, unlike the BP cost function, while often possessing a more limited constellation of local minima than comparable MAP methods which share this property. We also derive conditions, dependent on the distribution of the nonzero model weights embedded in the optimal representation, such that SBL has no local minima. Finally, we demonstrate how a generalized form of SBL, out of a large class of latent-variable Bayesian models (which includes both MAP and empirical Bayesian algorithms), uniquely satisfies two minimal performance criteria directly linked to sparsity. These results lead to a deeper understanding of the connections between various Bayesian-inspired strategies and suggest new sparse learning algorithms.

We have also extended these methodologies to handle more general problems relevant to compressed sensing, source localization and the analysis of neural data. In this context, modifications of SBL were considered for handling sparse representations that arise in spatio-temporal settings and in the context of covariance component estimation. Here we assume that a small set of common features underly the observed data collected over multiple instances. The theoretical properties of these SBL-based cost functions were examined and evaluated in the context of existing methods. The resulting algorithms display excellent performance on extremely large, ill-posed, and ill-conditioned problems in neuroimaging, suggesting a strong potential for impacting this field and others.

Bibliography

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

[2] I. Barrodale and F. Roberts, "Applications of mathematical programming to ℓ_p approximation," *Symposium on Nonlinear Programming*, pp. 447–464, May 1970.

[3] M. Beal and Z. Ghahrarmani, "The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures," *Bayesian Statistics 7*, Oxford University Press 2002.

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

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

[6] R. Boyles, "On the convergence properties of the EM algorithm," *J. Royal Statistical Society B*, vol. 44, pp. 47–50, 1983.

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

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

[9] J. Chen and X. Huo, "Sparse representations for multiple measurement vectors (MMV) in an overcomplete dictionary," *IEEE Int. Conf. Accoustics, Speech, and Signal Processing*, March 2005.

[10] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1999.

[11] S. Cotter and B. Rao, "Sparse channel estimation via matching pursuit with application to equalization," *IEEE Trans. Communications*, vol. 50, pp. 374–377, March 2002.

[12] S. Cotter, B. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Processing*, vol. 53, pp. 2477–2488, April 2005.

[13] A. Dempster, D. Rubin, and R. Tsutakawa, "Estimation in covariance components models," *Journal of the American Statistical Association*, vol. 76, pp. 341–353, June 1981.

[14] D. Donoho, *Wavelets: Theory, Algorithms, and Applications*, ch. On Minimum Entropy Segmentation, pp. 233–269. New York: Academic Press, 1994.

[15] D. Donoho and I. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

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

[17] D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proc. National Academy of Sciences*, vol. 100, pp. 2197–2202, March 2003.

[18] D. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution," *Stanford University Technical Report*, Sept. 2004.

[19] D. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Information Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.

[20] D. Donoho, "Compressed sensing," *IEEE Trans. Information Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[21] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York: John Wiley and Sons, Inc., 2nd ed., 2001.

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

[23] A. Faul and M. Tipping, "Analysis of sparse Bayesian learning," *Advances in Neural Information Processing Systems 14*, pp. 383–389, 2002.

[24] C. Févotte and S. Godsill, "Blind separation of sparse sources using JeffreyŠs inverse prior and the EM algorithm," *6th Int. Conf. Independent Component Analysis and Blind Source Separation*, March 2006.

[25] I. Fevrier, S. Gelfand, and M. Fitz, "Reduced complexity decision feedback equalization for multipath channels with large delay spreads," *IEEE Trans. Communications*, vol. 47, pp. 927–937, June 1999.

[26] M. Figueiredo and R. Nowak, "Wavelet-based image estimation: An empirical bayes approach using Jeffrey's noninformative prior," *IEEE Trans. Image Processing*, vol. 10, pp. 1322–1331, Sept. 2001.

[27] M. Figueiredo, "Adaptive sparseness using Jeffreys prior," *Advances in Neural Information Processing Systems 14*, pp. 697–704, 2002.

[28] K. Friston, W. Penny, C. Phillips, S. Kiebel, G. Hinton, and J. Ashburner, "Classical and Bayesian inference in neuroimaging: Theory," *NeuroImage*, vol. 16, pp. 465–483, 2002.

[29] J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Trans. Information Theory*, vol. 50, pp. 1341–1344, June 2004.

[30] H. Gao, "Wavelet shrinkage denoising using the nonnegative garrote," *Journal of Computational and Graphical Statistics*, vol. 7, no. 4, pp. 469–488, 1998.

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

[32] G. Golub and C. V. Loan, *Matrix Computations*. Baltimore, MD: Johns Hopkins University Press, 1989.

[33] I. Gorodnitsky, J. George, and B. Rao, "Neuromagnetic source imaging with FOCUSS: A recursive weighted minimum norm algorithm," *Journal Electroencephalography and Clinical Neurophysiology*, vol. 95, pp. 231–251, Oct. 1995.

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

[35] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Information Theory*, vol. 49, pp. 3320–3325, Dec. 2003.

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

[37] D. Harville, "Maximum likelihood approaches to variance component estimation and to related problems," *J. American Statistical Assoc.*, vol. 72, pp. 320–338, June 1977.

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

[39] R. Herbrich, *Learning Kernel Classifiers*. Cambridge, Mass.: MIT Press, 2002.

[40] 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," *12th Int. Conf. Biomagnetism*, pp. 753–756, 2000.

[41] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge University Press, 1990.

[42] M. Huang, A. Dale, T. Song, E. Halgren, D. Harrington, I. Podgorny, J. Canive, S. Lewis, and R. Lee, "Vector-based spatial-temporal minimum ℓ_1-norm solution for MEG," *NeuroImage*, vol. 31, no. 3, pp. 1025–1037, July 2006.

[43] A. Hyv$\ddot{a}$rinen, "Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation," *Neural Computation*, vol. 11, no. 7, pp. 1739–1768, 1999.

[44] A. Hyv$\ddot{a}$rinen, P. Hoyer, and E. Oja, *Intelligent Signal Processing*, ch. Image Denoising by Sparse Code Shrinkage. IEEE Press, 2001.

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

[46] R. Johnson and D. Wichern, *Applied Multivariate Statistical Analysis*. New Jersey: Prentice-Hall, 3rd ed., 1992.

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

[48] T. Jung, S. Makeig, M. McKeown, A. Bell, T. Lee, and T. Sejnowski, "Imaging brain dynamics using independent component analysis," *Proc. of the IEEE*, vol. 89, no. 7, pp. 1107–22, 2001.

[49] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. New Jersey: Prentice Hall, 1993.

[50] K. Kreutz-Delgado, J. F. Murray, B. Rao, K. Engan, T.-W. Lee, and T. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Computation*, vol. 15, pp. 349–396, Feb. 2003.

[51] R. Leahy and B. Jeffs, "On the design of maximally sparse beamforming arrays," *IEEE Transactions on Antennas and Propagation*, vol. 39, pp. 1178–1187, Aug. 1991.

[52] M. Lewicki and T. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, no. 2, pp. 337–365, 2000.

[53] Y. Li, S.-I. Amari, A. Cichocki, D. Ho, and S. Xie, "Underdetermined blind source separation based on sparse representation," *IEEE Trans. Signal Processing*, vol. 54, no. 2, pp. 423–437, 2006.

[54] Y. Lin and D. Lee, "Bayesian ℓ_1-norm sparse learning," *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, May 2006.

[55] G. Luenberger, *Linear and Nonlinear Programming*. Reading, Massachusetts: Addison–Wesley, second ed., 1984.

[56] D. MacKay, "Bayesian interpolation," *Neural Computation*, vol. 4, no. 3, pp. 415–447, 1992.

[57] D. MacKay, "Bayesian non-linear modelling for the energy prediction competition," *ASHRAE Transactions*, vol. 100, no. 2, pp. 1053–1062, 1994.

[58] D. MacKay, "Comparison of approximate methods for handling hyperparameters," *Neural Computation*, vol. 11, no. 5, pp. 1035–1068, 1999.

[59] D. Malioutov, M. Çetin, and A. Willsky, "Optimal sparse representations in general overcomplete bases," *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, pp. II–793–796, May 2004.

[60] D. Malioutov, M. Çetin, and A. Willsky, "Sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Processing*, vol. 53, pp. 3010–3022, Aug. 2005.

[61] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41, pp. 3397–3415, Dec. 1993.

[62] J. Mattout, C. Phillips, W. Penny, M. Rugg, and K. Friston, "MEG source localization under multiple constraints: An extended Bayesian framework," *NeuroImage*, vol. 30, pp. 753–767, 2006.

[63] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using generalized gaussian and complexity priors," *IEEE Trans. Information Theory*, vol. 45, April 1999.

[64] J. Murray and K. Kreutz-Delgado, "Sparse image coding using learned overcomplete dictionaries," *IEEE Int. Workshop on Machine Learning for Signal Processing*, Sept. 2004.

[65] J. Murray and K. Kreutz-Delgado, "Learning sparse overcomplete codes for images," *submitted*, February 2005.

[66] R. Neal, *Bayesian Learning for Neural Networks*. New York: Springer-Verlag, 1996.

[67] B. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, June 1996.

[68] B. Olshausen and D. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vision Research*, vol. 37, pp. 3311–3325, 1997.

[69] B. O. D. Field, "Sparse coding of sensory inputs," *Current Opinion in Neurobiology*, vol. 14, pp. 481–487, 2004.

[70] J. Palmer, D. Wipf, K. Kreutz-Delgado, and B. Rao, "Variational EM algorithms for non-Gaussian latent variable models," *Advances in Neural Information Processing Systems 18*, pp. 1059–1066, 2006.

[71] R. Pascual-Marqui, "Standardized low resolution brain electromagnetic tomography (sLORETA): Technical details," *Methods and Findings in Experimental and Clinical Pharmacology*, vol. 24, no. Suppl D, pp. 5–12, 2002.

[72] J. Perez-Orive, O. Mazor, G. Turner, S. Cassenaer, R. Wilson, and G. Laurent, "Oscillations and sparsening of odor representations in the mushroom body," *Science*, vol. 297, pp. 359–365, 2002.

[73] C. Phillips, J. Mattout, M. Rugg, P. Maquet, and K. Friston, "An empirical Bayesian solution to the source reconstruction problem in EEG," *NeuroImage*, vol. 24, pp. 997–1011, Jan. 2005.

[74] J. Phillips, R. Leahy, and J. Mosher, "MEG-based imaging of focal neuronal current sources," *IEEE Trans. Medical Imaging*, vol. 16, no. 3, pp. 338–348, 1997.

[75] R. Ramírez, *Neuromagnetic Source Imaging of Spontaneous and Evoked Human Brain Dynamics*. Phd Thesis, New York University, May 2005.

[76] R. Ramírez and S. Makeig, "Neuroelectromagnetic source imaging using multiscale geodesic neural bases and sparse Bayesian learning," *Human Brain Mapping, 12th Annual Meeting*, vol. Florence, Italy, June 2006.

[77] B. Rao and K. Kreutz-Delgado, "Basis selection in the presence of noise," *32nd Asilomar Conf. Signals, Systems and Computers*, vol. 1, pp. 752–756, Nov. 1998.

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

[79] B. 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, March 2003.

[80] B. Rao and B. Song, "Adaptive filtering algorithms for promoting sparsity," *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 6, pp. 361–364, April 2003.

[81] B. Rao, S. Cotter, and K. Engan, "Diversity measure minimization based method for computing sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, May 2004.

[82] C. Rasmussen, *Evaluation of Gaussian Processes and Other Methods for Non-Linear Regression*. PhD Thesis, Graduate Department of Computer Science, University of Toronto, 1996.

[83] S. Resnick, *A Probability Path*. Birkhauser, 1999.

[84] B. Ripley, *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

[85] R. Rockafellar, *Convex Analysis*. Princeton University Press, 1970.

[86] M. Sahani and J. Linden, "Evidence optimization techniques for estimating stimulus-response functions," *Advances in Neural Information Processing Systems 15*, pp. 301–308, 2003.

[87] M. Sahani and S. Nagarajan, "Reconstructing MEG sources with unknown correlations," *Advances in Neural Information Processing Systems 16*, 2004.

[88] J. Sarvas, "Basic methematical and electromagnetic concepts of the biomagnetic inverse problem," *Phys. Med. Biol.*, vol. 32, pp. 11–22, 1987.

[89] M. Sato, T. Yoshioka, S. Kajihara, K. Toyama, N. Goda, K. Doya, and M. Kawato, "Hierarchical Bayesian estimation for MEG inverse problem," *NeuroImage*, vol. 23, pp. 806–826, 2004.

[90] K. Sekihara, M. Sahani, and S. Nagarajan, "Localization bias and spatial resolution of adaptive and non-adaptive spatial filters for MEG source reconstruction," *NeuroImage*, vol. 25, pp. 1056–1067, 2005.

[91] J. Silva, J. Marques, and J. Lemos, "Selecting landmark points for sparse manifold learning," *Advances in Neural Information Processing Systems 18*, pp. 1241–1248, 2006.

[92] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Upper Saddle River, New Jersey: Prentice Hall, 1997.

[93] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Royal Statistical Society*, vol. 58, no. 1, pp. 267–288, 1996.

[94] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[95] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Information Theory*, vol. 50, pp. 2231–2242, Oct. 2004.

[96] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals," *IEEE Trans. Information Theory*, vol. 52, pp. 1030–1051, March 2006.

[97] J. Tropp, A. Gilbert, and M. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Processing*, vol. 86, pp. 572–588, April 2006.

[98] J. Tropp, "Algorithms for simultaneous sparse approximation. Part II: Convex relaxation," *Signal Processing*, vol. 86, pp. 589–602, April 2006.

[99] B. Turlach, W. Venables, and S. Wright, "Simultaneous variable selection," *Technometrics*, vol. 47, pp. 349–363, Aug. 2005.

[100] K. Uutela, M. Hamalainen, and E. Somersalo, "Visualization of magnetoencephalographic data using minimum current estimates," *NeuroImage*, vol. 10, pp. 173–180, 1999.

[101] W. Vinje and J. Gallant, "Sparse coding and decorrelation in primary visual cortex during natural vision," *Science*, vol. 287, pp. 1273–1276, 2000.

[102] M. Wakin, M. Duarte, S. Sarvotham, D. Baron, and R. Baraniuk, "Recovery of jointly sparse signals from a few random projections," *Advances in Neural Information Processing Systems 18*, pp. 1433–1440, 2006.

[103] C. Wu, "On the convergence properties of the EM algorithm," *The Annals of Statistics*, vol. 11, pp. 95–103, 1983.

[104] W. Zangwill, *Nonlinear Programming: A Unified Approach*. New Jersey: Prentice Hall, 1969.