PERFORMANCE EVALUATION OF LATENT VARIABLE MODELS WITH SPARSE PRIORS

David Wipf, Jason Palmer, Bhaskar Rao, and Kenneth Kreutz-Delgado

Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093-0407 USA
e-mail: {dwipf,japalmer}@ucsd.edu, {brao,kreutz}@ece.ucsd.edu

ABSTRACT

A variety of Bayesian methods 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 paper 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. Later, all methods are shown to be performing MAP estimation using potentially non-factorable implicit priors, which suggests new sparse learning cost functions.

Index Terms— sparse representations, sparse priors, latent variable models, underdetermined inverse problems, Bayesian learning

1. INTRODUCTION

Here we will be concerned with the generative model

$$\boldsymbol{y} = \Phi \boldsymbol{x} + \boldsymbol{\epsilon}, \tag{1}$$

where $\Phi \in \mathbb{R}^{N \times M}$ is a dictionary of unit ℓ_2-norm basis vectors or features, $\boldsymbol{x}$ is a vector of unknown weights, $\boldsymbol{y}$ 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 $\text{rank}(\Phi) = N$. 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 non-negative 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

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

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

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

whose form is rooted in convex analysis and duality theory. As shown in [10], virtually all sparse priors of interest can be decomposed using both (2) and (3), including the popular Laplacian, Jeffreys, Student's t, and generalized Gaussian priors.[2] The key requirement is that $p(x_i)$ is *strongly supergaussian*, which requires that

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

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 $\hat{\boldsymbol{x}}$ is all that is required, rather than merely a good estimate of $p(\boldsymbol{y})$ or the conditional distribution of new data-points y^*, i.e., $p(y^*|\boldsymbol{y})$. 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 (2) and (3) offers a very convenient means of obtaining (local) MAP estimates of $\boldsymbol{x}$ using generalized EM procedures that iteratively solve

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}). \tag{5}$$

Henceforth referred to as *Type I methods*, common examples include minimum ℓ_p-quasi-norm approaches [6, 12], Jeffreys prior-based methods sometimes called FOCUSS [2, 3, 5], and algorithms for computing the basis pursuit (BP) or Lasso solution [3, 7, 12].

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

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

This work was supported by NSF grant DGE-0333451 and NSF grant IIS-0613595.

[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 (2).

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

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

where $\Gamma \triangleq \mathrm{diag}(\boldsymbol{\gamma})$. Relevant examples include sparse Bayesian learning (SBL) [15], automatic relevance determination (ARD) [9], evidence maximization [13], and methods for learning overcomplete dictionaries [4]. Perhaps surprisingly, even the popular variational mean-field approximations, which optimize a factorial posterior distribution such that $p(\boldsymbol{x},\boldsymbol{\gamma}|\boldsymbol{y}) \approx q(\boldsymbol{x}|\boldsymbol{y})q(\boldsymbol{\gamma}|\boldsymbol{y})$, are equivalent to the Type II methods in the context of strongly supergaussian priors [10]. A specific example of this can be found in [1].

In applying all of these methods in practice, the performance achieving sparse solutions is varied. Results can be highly dependent on the (subjective) parameterization used in forming the latent variables. This occurs because the decomposition of $p(\boldsymbol{x})$ 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 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 limiting 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 [16]. We conclude by showing that all Type II methods can be viewed as performing MAP estimation using non-separable (i.e., non-factorable) implicit priors. This elucidates connections between methods and suggests new sparse learning cost functions.

2. A UNIFIED COST FUNCTION

Given the significant discrepancies between the modelling assumptions of 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 related to a single underlying objective function. We start with two intermediate results before presenting the main idea.

Lemma 1. Given a sparse prior expressible using (2) or (3), the resulting posterior mode over $\boldsymbol{x}$ (as is sought by Type I methods) can be obtained by minimizing the cost function

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

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

Proof: From basic linear algebra, we have

$$\boldsymbol{y}^T\Sigma_y^{-1}\boldsymbol{y} = \min_{\boldsymbol{x}} \;\; \frac{1}{\lambda}\left\|\boldsymbol{y}-\Phi\boldsymbol{x}\right\|_2^2 + \boldsymbol{x}^T\Gamma^{-1}\boldsymbol{x}. \tag{9}$$

The minimizing $\boldsymbol{x}$ is given by (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

$$\min_{\boldsymbol{\gamma}} \mathcal{L}_{(I)}(\boldsymbol{\gamma};\lambda,f) = \min_{\boldsymbol{\gamma}}\min_{\boldsymbol{x}} \frac{1}{\lambda}\left\|\boldsymbol{y}-\Phi\boldsymbol{x}\right\|_2^2 + \boldsymbol{x}^T\Gamma^{-1}\boldsymbol{x} + \sum_i -g^*(\gamma_i^{-1}). \tag{10}$$

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

$$\min_{\boldsymbol{\gamma}} \boldsymbol{x}^T\Gamma^{-1}\boldsymbol{x} + \sum_i -g^*(\gamma_i^{-1}) = \sum_i g(x_i^2), \tag{11}$$

which follows from the representation (3) and its assumption that $g(\cdot)$ is concave in x_i^2 [10]. Since the posterior mode is given by the minimum of

$$\mathcal{L}_{(I)}(\boldsymbol{x};\lambda,f) \triangleq -\log p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}) \equiv \left\|\boldsymbol{y}-\Phi\boldsymbol{x}\right\|_2^2 + \lambda\sum_i g(x_i^2), \tag{12}$$

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

Lemma 2. All of the Type II methods can be obtained by minimizing the cost function

$$\mathcal{L}_{(II)}(\boldsymbol{\gamma};\lambda,f) \triangleq \boldsymbol{y}^T\Sigma_y^{-1}\boldsymbol{y} + \log|\Sigma_y| + \sum_{i=1}^{M} f(\boldsymbol{\gamma}_i). \tag{13}$$

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

Theorem 1. Both the Type I and Type II cost functions can be reduced to (13) with the appropriate selection of $f(\cdot)$ and λ.

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

$$\bar{f}(\cdot) \triangleq \alpha f\left[\alpha(\cdot)\right], \qquad \bar{\lambda} \triangleq \alpha^{-1}\lambda, \tag{14}$$

where $\alpha > 0$ is a constant. Under these definitions, we have

$$\begin{aligned}
&\mathcal{L}_{(II)}(\boldsymbol{\gamma};\bar{\lambda},\bar{f}) \\
&= \boldsymbol{y}^T\left[\bar{\lambda}I + \Phi\Gamma\Phi^T\right]^{-1}\boldsymbol{y} + \log\left|\bar{\lambda}I + \Phi\Gamma\Phi^T\right| + \sum_i \bar{f}(\gamma_i) \\
&\equiv \boldsymbol{y}^T\Big[\lambda I + \alpha\Phi\Gamma\Phi^T\Big]^{-1}\boldsymbol{y} + \frac{1}{\alpha}\log\left|\lambda I + \alpha\Phi\Gamma\Phi^T\right| + \sum_i f(\alpha\gamma_i)
\end{aligned}$$

and so as α becomes large

$$\mathcal{L}_{(II)}(\boldsymbol{\gamma};\bar{\lambda},\bar{f}) \rightarrow \boldsymbol{y}^T\left[\lambda I + \Phi(\alpha\Gamma)\Phi^T\right]^{-1}\boldsymbol{y} + \sum_i f(\alpha\gamma_i). \tag{15}$$

This is equivalent to (8) with the exception of the scaling factor of α on $\boldsymbol{\gamma}$. However, this factor is irrelevant in that the weight estimate $\hat{\boldsymbol{x}}$

so obtained will be identical [16]. ■

In summary, by choosing the appropriate sparse prior, and therefore the function $f(\cdot)$, any Type I cost function can be reduced to a limiting case of Type II. The same can be shown for the associated update rules. As will be discussed in Section 5, 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.

3. MINIMAL PERFORMANCE CONDITIONS

In the past, different methods have been justified based on the plausibility of the full model and the full prior $p(\boldsymbol{x})$, 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 [12]). This ensures that an algorithm is guaranteed to prune at least $M - N$ unnecessary coefficients, a minimal sparsity condition.

Condition 2. *If $\boldsymbol{y} = \omega \boldsymbol{\phi}_i$ for some $\omega \in \mathbb{R}$ and unique dictionary column $\boldsymbol{\phi}_i$, then there is a unique, minimizing solution characterized by $\hat{\boldsymbol{x}} = \omega \boldsymbol{e}_i$, 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 [14].

4. 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 $p(\boldsymbol{x})$ can be decomposed using multiple latent parameterizations, leading to different effective values of $f(\cdot)$.

We first give some preliminary results before presenting the main theorem. Also, the analysis assumes that each column of Φ has unit ℓ_2 norm.

Lemma 3. To satisfy Condition 1, $f(\cdot)$ must be a nondecreasing function on $[0, \infty)$.

This result is very straightforward to show.

Lemma 4. Let $f(\cdot)$ be convex and nonlinear 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_y will lead to locally minimizing solutions with more than N nonzero elements [16].

Lemma 5. Let $f(\cdot)$ be concave and nonlinear on $[0, \infty)$. Then $\mathcal{L}(\boldsymbol{\gamma}; \lambda, f)$ violates Condition 2.

The proof has been deferred to [16]. 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 6. $\mathcal{L}(\boldsymbol{\gamma}; \lambda, f)$ satisfies Conditions 1 and 2 if $f(z) \propto \alpha z$, where $\alpha \geq 0$.

See [16] for the proof. Combining all of the above, we arrive at the following conclusion:

Theorem 2. $\mathcal{L}(\boldsymbol{\gamma}; \lambda, f)$ 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 (2) to perform MAP estimation of $\boldsymbol{\gamma}$, we find that $p(\boldsymbol{x})$ is Laplacian, but using the convex representation (or when using the equivalent variational Bayes formulation), $p(\boldsymbol{x})$ 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 (3) as in done in [4], 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 limiting cases of $\mathcal{L}(\boldsymbol{\gamma}; \lambda, f)$ when using $f(z) = \alpha z$. SBL is obtained with $\alpha \to 0$, while BP results from the assumption $\alpha \to \infty$, with $\lambda \to \lambda / \alpha^{1/2}$. The general case is easily implemented using EM updates, where the E-step involves computing the posterior moments

$$E\left[\boldsymbol{x}\boldsymbol{x}^T | \boldsymbol{y}; \boldsymbol{\gamma}\right] = \Gamma \Phi^T \Sigma_y^{-1} \boldsymbol{y}\boldsymbol{y}^T \Sigma_y^{-1} \Phi \Gamma + \Gamma - \Gamma \Phi^T \Sigma_y^{-1} \Phi \Gamma, \quad (16)$$

while the M-step reduces to

$$\gamma_i = \frac{-1 + \left(1 + 4\alpha E\left[\boldsymbol{x}\boldsymbol{x}^T | \boldsymbol{y}; \boldsymbol{\gamma}\right]_{ii}\right)^{1/2}}{2\alpha}. \quad (17)$$

Consistent with the above observations, when $\alpha \to 0$, these expressions reduce to the exact SBL updates (EM version), while the assumptions $\alpha \to \infty$, with $\lambda \to \lambda / \alpha^{1/2}$ produce an interior point method for computing the BP solution. For all other α, the algorithm is very effective in empirical tests [16], although the optimal value is likely application dependent.

5. 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 framework [10]. 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 [10]. 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 [11].

A final point is worth exploring regarding the difference between Type I and Type II approaches. By convention, Type I methods, being labelled as MAP estimates for $\boldsymbol{x}$, 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 [8, 15]. 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 $\boldsymbol{x}$-space using the prior

$$p(\boldsymbol{x}) \propto \exp\left[-\frac{1}{2}\min_{\boldsymbol{\gamma}}\left(\boldsymbol{x}^T\Gamma^{-1}\boldsymbol{x} + \log|\Sigma_y| + \sum_i f(\gamma_i)\right)\right]. \tag{18}$$

Although not generally available in closed form, this prior is necessarily concave in $\boldsymbol{x}^2$ in the same sense as the priors (2) and (3) [16]. Unlike the previous prior expressions however, (18) is *non-separable*, meaning $p(\boldsymbol{x}) \neq \prod_i p(x_i)$. This we believe is the key distinction between Type I and Type II; both are finding MAP estimates of $\boldsymbol{x}$, 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 $\boldsymbol{x}$-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 2, the SBL cost can be expressed as

$$\begin{aligned}\mathcal{L}_{\rm SBL}(\boldsymbol{\gamma};\lambda) &= \boldsymbol{y}^T\Sigma_y^{-1}\boldsymbol{y} + \log|\Gamma| + \log\left|\Gamma^{-1} + \lambda^{-1}\Phi^T\Phi\right| \\ &= \mathcal{L}_{\rm FOCUSS}(\boldsymbol{\gamma};\lambda) + \log\left|\Gamma^{-1} + \lambda^{-1}\Phi^T\Phi\right|, \end{aligned} \tag{19}$$

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 $p(\boldsymbol{x})$? 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.

6. REFERENCES

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

[2] C. Févotte and S. Godsill, "Blind separation of sparse sources using Jeffreys inverse prior and the EM algorithm," *Proc. 6th Int. Conf. Independent Component Analysis and Blind Source Separation*, March 2006.

[3] M. Figueiredo, "Adaptive sparseness using Jeffreys prior," *Advances in Neural Information Processing Systems 14*, pp. 697–704, 2002.

[4] M. Girolami, "A variational method for learning sparse and overcomplete representations," *Neural Computation*, 13(11):2517–2532, 2001.

[5] I. Gorodnitsky and B. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Sig. Proc.*, 45(3):600–616, 1997.

[6] K. Kreutz-Delgado, J. Murray, B. Rao, K. Engan, T.-W. Lee, and T. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Computation*, 15(2):349–396, 2003.

[7] Y. Lin and D. Lee, "Bayesian ℓ_1-norm sparse learning," *IEEE Int. Conf. Acoustics, Speech, and Sig. Proc.*, May 2006.

[8] D. MacKay, "Comparison of approximate methods for handling hyperparameters," *Neural Computation*, 11(5):1035–1068, 1999.

[9] R. Neal, *Bayesian Learning for Neural Networks*. New York: Springer-Verlag, 1996.

[10] 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*, 1059–1066, 2006.

[11] R. Ramírez, "Neuromagnetic source imaging of spontaneous and evoked human brain dynamics," PhD Thesis, New York University, May 2005.

[12] B. Rao, K. Engan, S. Cotter, J. Palmer, and K. Kreutz-Delgado, "Subset selection in noise based on diversity measure minimization," *IEEE Trans. Signal Processing*, 51(3):760–770, 2003.

[13] M. Sahani and J. Linden, "Evidence optimization techniques for estimating stimulus-response functions," *Advances in Neural Information Processing Systems 15*, 301–308, 2003.

[14] K. Sekihara, M. Sahani, and S. Nagarajan, "Localization bias and spatial resolution of adaptive and non-adaptive spatial filters for MEG source reconstruction," *NeuroImage*, 25:1056–1067, 2005.

[15] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, 1:211–244, 2001.

[16] D. Wipf, J. Palmer, B. Rao, and K. Kreutz-Delgado, "A general framework for handling latent variable models with sparse priors," in preparation, 2007.