Fast Variational Sparse Bayesian Learning With Automatic Relevance Determination for Superimposed Signals

Dmitriy Shutin, Thomas Buchgraber, Sanjeev R. Kulkarni, and H. Vincent Poor

***Abstract*—In this work, a new fast variational sparse Bayesian learning (SBL) approach with automatic relevance determination (ARD) is proposed. The sparse Bayesian modeling, exemplified by the relevance vector machine (RVM), allows a sparse regression or classification function to be constructed as a linear combination of a few basis functions. It is demonstrated that, by computing the stationary points of the variational update expressions with noninformative (ARD) hyperpriors, a fast version of variational SBL can be constructed. Analysis of the computed stationary points indicates that SBL with Gaussian sparsity priors and noninformative hyperpriors corresponds to removing components with signal-to-noise ratio below a 0 dB threshold; this threshold can also be adjusted to significantly improve the convergence rate and sparsity of SBL. It is demonstrated that the pruning conditions derived for fast variational SBL coincide with those obtained for fast marginal likelihood maximization; moreover, the parameters that maximize the variational lower bound also maximize the marginal likelihood function. The effectiveness of fast variational SBL is demonstrated with synthetic as well as with real data.**

***Index Terms*—Automatic relavance determination, sparse Bayesian learning, variational Bayesian inference.**

I. Introduction

During the last decade research of sparse signal representations has received considerable attention [1]–[5]. With a few minor variations, the general goal of sparse reconstruction is to optimally estimate the parameters of the following canonical model:

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

where $\boldsymbol{t} \in \mathbb{R}^N$ is a vector of targets, $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_L]$ is a design matrix with L columns corresponding to basis functions $\boldsymbol{\phi}_l \in \mathbb{R}^N$, $l = 1, \ldots, L$, and $\boldsymbol{w} = [w_1, \ldots, w_L]^T$ is a vector of weights that are to be estimated. The additive perturbation $\boldsymbol{\xi}$ is typically assumed to be a white Gaussian random vector with zero mean and covariance matrix $\tau^{-1}\boldsymbol{I}$, where τ is a noise precision parameter. Imposing constraints on the model parameters $\boldsymbol{w}$ is a key to sparse signal modeling [3].

In sparse Bayesian learning (SBL) [2], [4], [6] the weights $\boldsymbol{w}$ are constrained using a parametric prior $p(\boldsymbol{w}|\boldsymbol{\alpha})$; this prior is a symmetric probability density function (pdf) with zero mean and prior parameters $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_L]^T$—also called sparsity parameters—that are inversely proportional to the width of the pdf. Hence, a large value of α_l drives the posterior value of the corresponding element w_l in the vector $\boldsymbol{w}$ towards zero, thus encouraging a solution with only a few nonzero coefficients.

In the relevance vector machine (RVM) approach to the SBL problem [2], the sparsity parameters $\boldsymbol{\alpha}$ are estimated by maximizing the marginal likelihood $p(\boldsymbol{t}|\boldsymbol{\alpha}, \tau) = \int p(\boldsymbol{t}|\boldsymbol{w}, \tau)p(\boldsymbol{w}|\boldsymbol{\alpha})\mathrm{d}\boldsymbol{w}$, which is also termed model evidence [2], [6], [7]; the corresponding estimation approach is then referred to as the Evidence Procedure [2]. Unfortunately, the RVM solution is known to converge rather slowly and the computational complexity of the algorithm scales as $O(L^3)$[2], [8]; this makes the application of RVMs to large data sets impractical. In [8] an alternative learning scheme was proposed to alleviate this drawback. Specifically, for a Gaussian prior $p(\boldsymbol{w}|\boldsymbol{\alpha}) \propto \exp(-\sum_l \frac{\alpha_l |w_l|^2}{2})$ the maximum of the marginal likelihood function with respect to a single sparsity parameter α_l, assuming the sparsity parameters of the other basis functions are fixed, can be evaluated in closed form.[1]

An alternative approach to SBL is based on approximating the posterior $p(\boldsymbol{w}, \tau, \boldsymbol{\alpha}|\boldsymbol{t})$ with a variational proxy pdf $q(\boldsymbol{w}, \tau, \boldsymbol{\alpha}) = q(\boldsymbol{w})q(\tau)q(\boldsymbol{\alpha})$ [10] so as to maximize the variational lower bound on $\log p(\boldsymbol{t})$ [11]. There are several advantages of the variational approach to SBL as compared to that proposed in [2] and [8]. First, the distributions rather than point estimates of the unobserved variables can be obtained. Second, the variational approach to SBL allows one to obtain analytical approximations to the posterior distributions of interest even when exact inference of these distributions is intractable. Finally, the variational methodology provides a unifying framework for inference on graphical models that represent extensions of (1), such as different sparsity priors, parametric design matrices, etc. (see, e.g., [12] and [13]). Unfortunately, the variational approach to SBL discussed in [10] is similar to the RVM in terms of estimation complexity and is prone to a slow convergence rate. Also, the pdfs $q(\boldsymbol{\alpha})$ and $q(\tau)$ are estimated so as to approximate the true posterior pdfs, thus obscuring the structure of the marginal likelihood that was exploited in [8] to accelerate the convergence rate of the learning scheme.

One possible strategy to improve the convergence rate of variational inference is to reduce coupling between the estimated random variables (see e.g., [14] and [15]). In this paper we propose an alternative approach that in some sense imitates fast marginal likelihood maximization (FMLM). Specifically, we consider the maximization of the variational lower bound with respect to a single factor $q(\alpha_l)$. As we will demonstrate, it then becomes possible to analytically compute the stationary points of the repeated updates of $q(\alpha_l)$ and $q(\boldsymbol{w})$ that maximize the bound and thus accelerate convergence. In [12] this was done both for a Gaussian prior $p(w_l|\alpha_l)$ and a Laplace prior $p(w_l|\alpha_l) \propto \exp(-\alpha_l|w_l|)$ when $q(\boldsymbol{w}) = \prod_l q(w_l)$, i.e., when the correlations between the elements of $\boldsymbol{w}$ are ignored. Here we present an extension of these results for the Gaussian prior case when $q(\boldsymbol{w})$ does not fully factor. We derive the closed form expressions for the stationary points of the variational updates of $q(\alpha_l)$ and determine conditions that ensure convergence to these stationary points. We demonstrate that the convergence condition for each $q(\alpha_l)$ has a simple and intuitive interpretation in terms of the component signal to noise ratio (SNR), which provides further insight into the performance of SBL and eventually allows one to improve it. Moreover, we show that this convergence condition coin-

Manuscript received October 26, 2010; revised March 14, 2011; accepted August 22, 2011. Date of publication September 15, 2011; date of current version November 16, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jerome Idier. This research was supported in part by an Erwin Schrödinger Postdoctoral Fellowship, Austrian Science Fund Project J2909-N23, in part by the Austrian Science Fund under Grant NFN-SISE-S10610-N13, and in part by the NSF Science and Technology Center under Grant CCF-0939370, the U.S. Army Research Office under Grant W911NF-07-1-0185, and the U.S. Office of Naval Research under Grant N00014-09-1-0342.

D. Shutin, S. R. Kulkarni, and H. V. Poor are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: dshutin@princeton.edu; kulkarni@princeton.edu; poor@princeton.edu).

T. Buchgraber is with the Signal Processing and Speech Communication Laboratory, Graz University of Technology, Graz, A-8010, Austria (e-mail: thomas.buchgraber@tugraz.at).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2011.2168217

[1] Fast suboptimal solutions to SBL with Laplace prior $p(\boldsymbol{w}|\boldsymbol{\alpha}) \propto \exp(-\sum_l \alpha_l |w_l|)$ have also been proposed [9].

cides with the condition that determines the maximum of the marginal likelihood function with respect to a single sparsity parameter.

Throughout this paper we shall make use of the following notation. The expression $\mathrm{diag}(\boldsymbol{x})$ stands for a diagonal matrix with the elements of $\boldsymbol{x}$ on the main diagonal; $\mathrm{tr}(\boldsymbol{X})$ denotes the trace of the matrix $\boldsymbol{X}$; $[\boldsymbol{B}]_{\bar{l}\bar{k}}$ denotes a matrix obtained by deleting the lth row and kth column from the matrix $\boldsymbol{B}$; similarly, $[\boldsymbol{b}]_{\bar{l}}$ denotes a vector obtained by deleting the lth element from the vector $\boldsymbol{b}$. Finally, for a random vector $\boldsymbol{x}$, $\mathrm{N}(\boldsymbol{x}|\boldsymbol{a}, \boldsymbol{B})$ denotes a multivariate Gaussian pdf with mean $\boldsymbol{a}$ and covariance matrix $\boldsymbol{B}$; similarly, for a random variable x, $\mathrm{Ga}(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$ denotes a gamma pdf with parameters a and b.

II. Variational Sparse Bayesian Learning

In SBL it is assumed that the joint pdf factorizes as $p(\boldsymbol{w}, \tau, \boldsymbol{\alpha}, \boldsymbol{t}) = p(\boldsymbol{t}|\boldsymbol{w}, \tau)p(\boldsymbol{w}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})p(\tau)$ [2], [4], [6]. Under the Gaussian noise assumption the likelihood $p(\boldsymbol{t}|\boldsymbol{w}, \tau)$ is given as $p(\boldsymbol{t}|\boldsymbol{w}, \tau) = \mathrm{N}(\boldsymbol{t}|\boldsymbol{\Phi}\boldsymbol{w}, \tau^{-1}\boldsymbol{I})$. The second term $p(\boldsymbol{w}|\boldsymbol{\alpha})$ is the sparsity prior that is assumed to factorize as $p(\boldsymbol{w}|\boldsymbol{\alpha}) = \prod_{l=1}^{L} p(w_l|\alpha_l)$. Henceforth, we will restrict our analysis to a Gaussian sparsity prior case where $p(w_l|\alpha_l) = \mathrm{N}(w_l|0, \alpha_l^{-1})$. The choice of the prior $p(\tau)$ is arbitrary in the context of this work; a convenient choice would be a gamma distribution , i.e., $p(\tau) = \mathrm{Ga}(\tau|c, d)$, since it is a conjugate prior for the precision of the Gaussian likelihood $p(\boldsymbol{t}|\boldsymbol{w}, \tau)$. The prior $p(\alpha_l)$, also called the hyperprior of the lth component, is selected as a gamma pdf $\mathrm{Ga}(\alpha_l|a_l, b_l)$.

The variational solution to SBL is obtained by maximizing a variational lower bound on a log-evidence $\log p(\boldsymbol{t})$ [10], [11], which can be shown to be equivalent to minimizing the Kullback-Leibler divergence between the approximating pdf $q(\boldsymbol{w}, \boldsymbol{\alpha}, \tau) = q(\boldsymbol{w})q(\tau)\prod_{k=1}^{L} q(\alpha_k)$ and the posterior pdf $p(\boldsymbol{w}, \tau, \boldsymbol{\alpha}|\boldsymbol{t})$. The factors of $q(\boldsymbol{w}, \boldsymbol{\alpha}, \tau)$, selected as $q(\boldsymbol{w}) = \mathrm{N}(\boldsymbol{w}|\hat{\boldsymbol{w}}, \hat{\boldsymbol{S}})$, $q(\alpha_l) = \mathrm{Ga}(\alpha_l|\hat{a}_l, \hat{b}_l)$, and $q(\tau) = \mathrm{Ga}(\tau|\hat{c}, \hat{d})$, are the variational approximating factors. It has been shown [10] that the parameters of the approximating factors—the variational parameters—can be computed as follows:

$$\hat{\boldsymbol{S}} = \left(\hat{\tau}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \mathrm{diag}(\hat{\boldsymbol{\alpha}})\right)^{-1}, \quad \hat{\boldsymbol{w}} = \hat{\tau}\hat{\boldsymbol{S}}\boldsymbol{\Phi}^T\boldsymbol{t} \tag{2}$$

$$\hat{a}_l = a_l + \frac{1}{2}, \; \hat{b}_l = b_l + \frac{(|\hat{w}_l|^2 + \hat{S}_{ll})}{2} \tag{3}$$

$$\hat{c} = c + \frac{N}{2}, \text{ and } \hat{d} = d + \frac{||\boldsymbol{t} - \boldsymbol{\Phi}\hat{\boldsymbol{w}}||^2 + \mathrm{tr}(\hat{\boldsymbol{S}}\boldsymbol{\Phi}^T\boldsymbol{\Phi})}{2} \tag{4}$$

where $\hat{\tau} = \mathbb{E}_{q(\tau)}\{\tau\} = \frac{\hat{c}}{\hat{d}}$, $\hat{\alpha}_l = \mathbb{E}_{q(\alpha_l)}\{\alpha_l\} = \frac{\hat{a}_l}{\hat{b}_l}$, $\hat{w}_l$ is the lth element of the vector $\hat{\boldsymbol{w}}$, and $\hat{S}_{ll}$ is the lth element on the main diagonal of the matrix $\hat{\boldsymbol{S}}$.

A. Fast Variational SBL

Essentially, the variational update expressions (2)–(4) provide the estimates of the parameters of the corresponding approximating pdfs. These expressions reduce to those obtained in [2] when the approximating factors $q(\tau)$ and $q(\alpha_l)$ are chosen as Dirac measures on the corresponding domains and the expectation-maximization (EM) algorithm is used to maximize the marginal likelihood function. In [8] the authors circumvent the EM-based maximization of the marginal likelihood by computing the maximizer of the marginal log-likelihood function with respect to a single sparsity parameter α_l in closed form. In the variational approach the marginal likelihood function is not available. Nonetheless, as we intend to demonstrate, it is possible to analytically compute the stationary points of the repeated updates of $q(\boldsymbol{w})$ and $q(\alpha_l)$ for a single basis function, which also leads to a similar efficient realization of SBL.

Consider now the variational update expressions (2)–(4). Due to the convexity of the variational lower bound in the approximating factors $q(\boldsymbol{w}), q(\tau), q(\alpha_l),\ l = 1, \ldots, L$, we can update these factors in any order [11]; furthermore, a group of factors can be updated successively while keeping the other factors fixed.[2] Let us consider a noninformative hyperprior $p(\alpha_l)$, obtained by selecting $a_l = b_l = 0$ for all components [2]. We now study the expression for the mean $\hat{\alpha}_l$ of $q(\alpha_l)$ for some fixed basis function.[3] From (3) and the properties of a Gamma distribution it follows that $\hat{\alpha}_l^{-1} = \frac{\hat{b}_l}{\hat{a}_l} = \boldsymbol{e}_l^T(\hat{\boldsymbol{w}}\hat{\boldsymbol{w}}^T + \hat{\boldsymbol{S}})\boldsymbol{e}_l$, where $\boldsymbol{e}_l = [0, \ldots, 0, 1, 0, \ldots, 0]^T$ is a vector of all zeros with 1 at the lth position. Let us now assume that $q(\boldsymbol{w})$ and $q(\alpha_l)$ are successively updated while keeping $q(\tau)$ and $q(\alpha_k)$, $k \neq l$, fixed. This will generate a sequence of estimates $\left\{\hat{\alpha}_l^{[m]} = \frac{\hat{a}_l^{[m]}}{\hat{b}_l^{[m]}}\right\}_{m=1}^{M}$, with each element in the sequence computed according to (3). Our goal is to compute the stationary point $\hat{\alpha}_l^{[\infty]}$ of this sequence as $M \to \infty$.

First we note that $\hat{\boldsymbol{w}}\hat{\boldsymbol{w}}^T = \hat{\tau}^2\hat{\boldsymbol{S}}\boldsymbol{\Phi}^T\boldsymbol{t}\boldsymbol{t}^T\boldsymbol{\Phi}\hat{\boldsymbol{S}}^T$ and thus (3) can be rewritten as

$$\hat{\alpha}_l^{-1} = \boldsymbol{e}_l^T(\hat{\tau}^2\hat{\boldsymbol{S}}\boldsymbol{\Phi}^T\boldsymbol{t}\boldsymbol{t}^T\boldsymbol{\Phi}\hat{\boldsymbol{S}}^T + \hat{\boldsymbol{S}})\boldsymbol{e}_l. \tag{5}$$

Now consider the influence of a single sparsity parameter $\hat{\alpha}_l$ on the matrix $\hat{\boldsymbol{S}}$ in (2). By noting that $\mathrm{diag}(\hat{\boldsymbol{\alpha}}) = \sum_l \hat{\alpha}_l\boldsymbol{e}_l\boldsymbol{e}_l^T$, we rewrite $\hat{\boldsymbol{S}}$ as

$$\hat{\boldsymbol{S}} = \left(\hat{\tau}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \hat{\alpha}_l\boldsymbol{e}_l\boldsymbol{e}_l^T + \sum_{k\neq l}\hat{\alpha}_k\boldsymbol{e}_k\boldsymbol{e}_k^T\right)^{-1} = \bar{\boldsymbol{S}}_l - \frac{\bar{\boldsymbol{S}}_l\boldsymbol{e}_l\boldsymbol{e}_l^T\bar{\boldsymbol{S}}_l}{\hat{\alpha}_l^{-1} + \boldsymbol{e}_l^T\bar{\boldsymbol{S}}_l\boldsymbol{e}_l} \tag{6}$$

where the latter expression was obtained using the matrix inversion lemma [16] and defining

$$\bar{\boldsymbol{S}}_l = \left(\hat{\tau}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \sum_{k\neq l}\hat{\alpha}_k\boldsymbol{e}_k\boldsymbol{e}_k^T\right)^{-1}. \tag{7}$$

Finally, we define

$$\varsigma_l = \boldsymbol{e}_l^T\bar{\boldsymbol{S}}_l\boldsymbol{e}_l \quad \text{and} \quad \omega_l^2 = \hat{\tau}^2\boldsymbol{e}_l^T\bar{\boldsymbol{S}}_l\boldsymbol{\Phi}^T\boldsymbol{t}\boldsymbol{t}^T\boldsymbol{\Phi}\bar{\boldsymbol{S}}_l\boldsymbol{e}_l. \tag{8}$$

Now, by substituting (6) into (5) and using the definitions (8), we obtain

$$\hat{\alpha}_l^{-1} = (\omega_l^2 + \varsigma_l) - \frac{\varsigma_l^2 + 2\varsigma_l\omega_l^2}{\hat{\alpha}_l^{-1} + \varsigma_l} + \frac{\varsigma_l^2\omega_l^2}{(\hat{\alpha}_l^{-1} + \varsigma_l)^2}. \tag{9}$$

Expression (9) is a modified version of (5) that is now an implicit function of $\hat{\alpha}_l$. Solving for $\hat{\alpha}_l$ naturally leads to the desired stationary point $\hat{\alpha}_l^{[\infty]}$. Observe that (9) can be seen as a nonlinear map $\hat{\alpha}_l^{[m+1]} = F(\hat{\alpha}_l^{[m]})$ that at iteration m maps $\hat{\alpha}_l^{[m]}$ to $\hat{\alpha}_l^{[m+1]}$. Naturally, the stationary points of this map are equivalent to the desired (possibly multiple) stationary points $\hat{\alpha}_l^{[\infty]}$. The following theorem provides analytical expressions for the stationary points of the map $\hat{\alpha}_l^{[m+1]} = F(\hat{\alpha}_l^{[m]})$.

[2] Note, however, that the order in which the factors are updated is important since different update orderings might lead to different local optima of the variational lower bound. We will return to this issue later in the text.

[3] Notice that since $a_l = b_l = 0$, the parameters of $q(\alpha_l)$ in (3) can be specified as $\hat{a}_l = \frac{1}{2}$ and $\hat{b}_l = \frac{1}{(2\hat{\alpha}_l)}$, where $\hat{\alpha}_l = \frac{1}{(\hat{w}_l^2 + \hat{S}_{ll})}$. Thus, it makes sense to study the stationary point of the variational update expression in terms of $\hat{\alpha}_l$, rather than in terms of $\hat{a}_l$ and $\hat{b}_l$. However, the following analysis can be similarly performed for arbitrary values $a_l > 0$ and $b_l > 0$, resulting merely in a bit more involved expressions. For the sake of brevity we leave this analysis outside the scope of this paper.

Theorem 1: Assuming an initial condition $\alpha_l^{[0]} \geq 0$, the iterations of the nonlinear map

$$\hat{\alpha}_l^{[m+1]} = F(\hat{\alpha}_l^{[m]}) = \left((\omega_l^2 + \varsigma_l) - \frac{\varsigma_l^2 + 2\varsigma_l\omega_l^2}{\frac{1}{\hat{\alpha}_l^{[m]}} + \varsigma_l} + \frac{\varsigma_l^2\omega_l^2}{\left(\frac{1}{\hat{\alpha}_l^{[m]}} + \varsigma_l\right)^2} \right)^{-1} \tag{10}$$

where ω_l^2 and ς_l are given by (8), converge as $m \to \infty$ to

$$\hat{\alpha}_l^{[\infty]} = \begin{cases} (\omega_l^2 - \varsigma_l)^{-1}, & \omega_l^2 > \varsigma_l \\ \infty, & \omega_l^2 \leq \varsigma_l. \end{cases} \tag{11}$$

Proof: We begin by computing the stationary points of the map $\hat{\alpha}_l^{[m+1]} = F(\hat{\alpha}_l^{[m]})$. By inspecting (9) we observe that $\hat{\alpha}_l^{[\infty]} = \infty$ is a stationary point. The other solution is found by solving $\hat{\alpha}_l^* - F(\hat{\alpha}_l^*) = 0$ with respect to $\hat{\alpha}_l^*$. After rather tedious but straightforward algebraic manipulations we obtain the second stationary point at

$$\hat{\alpha}_l^{[\infty]} = \hat{\alpha}_l^* = (\omega_l^2 - \varsigma_l)^{-1}. \tag{12}$$

We now investigate the stability of (12) by analyzing the map (10) in the vicinity of $\hat{\alpha}_l^* = (\omega_l^2 - \varsigma_l)^{-1}$. It is known that a stationary point of a map is asymptotically stable if the eigenvalues of the Jacobian of the map evaluated at this stationary point are all within a unit circle. Thus, we compute

$$\left.\frac{\mathrm{d}F(\hat{\alpha}_l^*)}{\mathrm{d}\hat{\alpha}_l^*}\right|_{\hat{\alpha}_l^* = (\omega_l^2 - \varsigma_l)^{-1}} = -\frac{\varsigma_l(\varsigma_l - 2\omega_l^2)}{\omega_l^4}. \tag{13}$$

Now, it can be shown[4] that $\left|\frac{\varsigma_l(\varsigma_l - 2\omega_l^2)}{\omega_l^4}\right| < 1$ when

$$\omega_l^2 > \varsigma_l \tag{14}$$

$$\omega_l^2 < \varsigma_l < (1 + \sqrt{2})\omega_l^2. \tag{15}$$

Observe that the condition (15) suggests that the stationary point (12) might become negative. However, the negative value of $\alpha_l^{[m]}$ cannot be reached for $\alpha_l^{[0]} \geq 0$ since the map (10) can be shown to be positive for ς_l and ω_l^2 satisfying (15).[5] Thus, $\hat{\alpha}_l^{[\infty]} = (\omega_l^2 - \varsigma_l)^{-1}$ is a stable positive stationary point only when $\omega_l^2 > \varsigma_l$; if $\varsigma_l \geq \omega_l^2$, then (12) loses its stability and the iterations of the map converge to the other positive stationary point at $\hat{\alpha}_l^{[\infty]} = \infty$, i.e., the iterations simply diverge. ■

Corollary 1: Assume that $a_l = b_l = 0$ and (14) is satisfied for some basis function $\boldsymbol{\phi}_l$. Then the following is true: i) the repeated updates of $q(\boldsymbol{w})$ and $q(\alpha_l)$ from (2) and (3), respectively, maximize the variational lower bound and ii) $q(\alpha_l)$ converges to $q(\alpha_l) = \mathrm{Ga}\left(\alpha_l \middle| \frac{1}{2}, \frac{(\omega_l^2 - \varsigma_l)}{2}\right)$.

Expression (12) together with the pruning condition (14) allow one to assess the impact of the lth basis vector $\boldsymbol{\phi}_l$ in the matrix $\boldsymbol{\Phi}$ on the variational lower bound by computing (11): a finite value of $\hat{\alpha}_l^{[\infty]}$ instructs us to keep the lth component since it should increase the bound, while an infinite value of $\hat{\alpha}_l^{[\infty]}$ indicates that the basis vector l is superfluous. In this way all L basis vectors can be processed sequentially in a round-robin fashion.

B. Analysis of the Pruning Condition

Since the pruning condition (14) is a key to the model sparsity, let us study it in greater detail. From (8) it follows that ω_l^2 and ς_l are, respectively, a squared weight of the basis $\boldsymbol{\phi}_l$ and the estimated variance of this weight obtained when $\hat{\alpha}_l = 0$ and $\hat{\alpha}_k$, $k \neq l$, are fixed. Notice that the ratio $\frac{\omega_l^2}{\varsigma_l}$ can be recognized as an estimate of the lth component SNR. Furthermore, according to (14) the basis function $\boldsymbol{\phi}_l$ is retained in the model provided $\mathrm{SNR}_l = \frac{\omega_l^2}{\varsigma_l} > 1$, i.e., when the component's SNR is above 0 dB; otherwise, the component is pruned.

This simple interpretation of the pruning condition can be used to generalize (14) to any desired SNR above 0 dB. More specifically, given a certain desired $\mathrm{SNR}'_l \geq 1$, the pruning condition (14) can be empirically adjusted as

$$\omega_l^2 > \varsigma_l \times \mathrm{SNR}'_l \tag{16}$$

which allows for a removal of the lth component with the SNR satisfying $\mathrm{SNR}_l \leq \mathrm{SNR}'_l$. Note, however, that the adjustment (16) might potentially decrease the variational lower bound since it will remove basis functions with finite sparsity parameters. Despite the empirical nature of (16), it has, however, a firm theoretical justification. It can be shown that $\mathrm{SNR}_l = \frac{\omega_l^2}{\varsigma_l}$ follows a χ^2 distribution when the actual weight w_l is 0 and a noncentral χ^2 distribution when $w_l \neq 0$. This allows formulating a statistical composite hypothesis test to determine whether an estimate $\frac{\omega_l^2}{\varsigma_l}$ follows a χ^2 distribution, and thus $\hat{w}_l$ must be 0, or a noncentral χ^2 distribution, which means $\hat{w}_l \neq 0$. Based on this interpretation, it can then be shown that (14) corresponds to a test with a rather large test size. In contrast, (16) corresponds to a test with a smaller test size determined by $\mathrm{SNR}'_l \geq 1$. Space limitations prohibit a detailed study of this interpretation of (16).

III. Implementation Aspects and Simulation Results

A. Efficient Computation of the Test Parameters ω_l^2 and ς_l

Consider the matrix $\bar{\boldsymbol{S}}_l$ in (7) in an alternative (permuted) form, where $\boldsymbol{\phi}_l$ is shifted to the last column of $\boldsymbol{\Phi}$:

$$\begin{pmatrix} \hat{\tau}\boldsymbol{\Phi}_{\bar{l}}^T\boldsymbol{\Phi}_{\bar{l}} + \mathrm{diag}(\hat{\boldsymbol{\alpha}}_{\bar{l}}) & \hat{\tau}\boldsymbol{\Phi}_{\bar{l}}^T\boldsymbol{\phi}_l \\ \hat{\tau}\boldsymbol{\phi}_l^T\boldsymbol{\Phi}_{\bar{l}} & \hat{\tau}\boldsymbol{\phi}_l^T\boldsymbol{\phi}_l \end{pmatrix}^{-1}. \tag{17}$$

Here, $\boldsymbol{\Phi}_{\bar{l}}$ is a matrix obtained from $\boldsymbol{\Phi}$ by removing the lth basis $\boldsymbol{\phi}_l$ and $\hat{\boldsymbol{\alpha}}_{\bar{l}} = [\hat{\boldsymbol{\alpha}}]_{\bar{l}}$. Using a standard result for block matrix inversion [16], the expression (17) can be rewritten as

$$\begin{pmatrix} \left(\hat{\boldsymbol{S}}_{\bar{l}}^{-1} - \hat{\tau}\frac{\boldsymbol{\Phi}_{\bar{l}}^T\boldsymbol{\phi}_l\boldsymbol{\phi}_l^T\boldsymbol{\Phi}_{\bar{l}}}{\boldsymbol{\phi}_l^T\boldsymbol{\phi}_l}\right)^{-1} & -\hat{\tau}\hat{\boldsymbol{S}}_{\bar{l}}\boldsymbol{\Phi}_{\bar{l}}^T\boldsymbol{\phi}_l\gamma_l^{-1} \\ -\hat{\tau}\gamma_l^{-1}\boldsymbol{\phi}_l^T\boldsymbol{\Phi}_{\bar{l}}\hat{\boldsymbol{S}}_{\bar{l}} & \gamma_l^{-1} \end{pmatrix} \tag{18}$$

where $\gamma_l = \hat{\tau}\boldsymbol{\phi}_l^T\boldsymbol{\phi}_l - \hat{\tau}^2\boldsymbol{\phi}_l^T\boldsymbol{\Phi}_{\bar{l}}\hat{\boldsymbol{S}}_{\bar{l}}\boldsymbol{\Phi}_{\bar{l}}^T\boldsymbol{\phi}_l$ and $\hat{\boldsymbol{S}}_{\bar{l}} = (\hat{\tau}\boldsymbol{\Phi}_{\bar{l}}^T\boldsymbol{\Phi}_{\bar{l}} + \mathrm{diag}(\hat{\boldsymbol{\alpha}}_{\bar{l}}))^{-1} = \left[\frac{\hat{\boldsymbol{S}} - \hat{\boldsymbol{S}}\boldsymbol{e}_l\boldsymbol{e}_l^T\hat{\boldsymbol{S}}}{\boldsymbol{e}_l^T\hat{\boldsymbol{S}}\boldsymbol{e}_l}\right]_{\bar{l}\bar{l}}$.

Using (18), ω_l^2 and ς_l in (8) can now be computed as

$$\varsigma_l = (\hat{\tau}\boldsymbol{\phi}_l^T\boldsymbol{\phi}_l - \hat{\tau}^2\boldsymbol{\phi}_l^T\boldsymbol{\Phi}_{\bar{l}}\hat{\boldsymbol{S}}_{\bar{l}}\boldsymbol{\Phi}_{\bar{l}}^T\boldsymbol{\phi}_l)^{-1}$$

and

$$\omega_l^2 = (\hat{\tau}\varsigma_l\boldsymbol{\phi}_l^T\boldsymbol{t} - \hat{\tau}^2\varsigma_l\boldsymbol{\phi}_l^T\boldsymbol{\Phi}_{\bar{l}}\hat{\boldsymbol{S}}_{\bar{l}}\boldsymbol{\Phi}_{\bar{l}}^T\boldsymbol{t})^2. \tag{19}$$

Notice that when the basis $\boldsymbol{\phi}_l$ is retained in the model, the matrix $\hat{\boldsymbol{S}}$ can be efficiently updated:

$$\hat{\boldsymbol{S}} = \hat{\boldsymbol{S}} - \frac{\hat{\boldsymbol{S}}\boldsymbol{e}_l\boldsymbol{e}_l^T\hat{\boldsymbol{S}}}{(\hat{\alpha}_l^{[\infty]} - \hat{\alpha}_l)^{-1} + \boldsymbol{e}_l^T\hat{\boldsymbol{S}}\boldsymbol{e}_l} \tag{20}$$

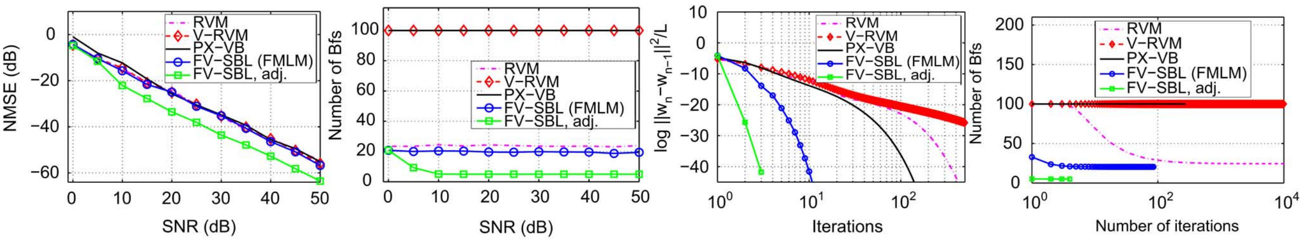
where $\hat{\alpha}_l^{[\infty]} = (\omega_l^2 - \varsigma_l)^{-1}$ and $\hat{\alpha}_l$ is a previous mean of $q(\alpha_l)$.

B. Equivalence Between Fast Variational SBL and Fast Marginal Likelihood Maximization

Theorem 2: Consider a basis function $\boldsymbol{\phi}_l$ and assume that the pdfs $q(\alpha_k)$, $k \neq l$, and $q(\tau)$ are fixed. Let $\hat{\alpha}_l^{[\infty]}$ be the mean of $q(\alpha_l)$ ob-

[4] Assuming $\varsigma_l \geq 0$ according to definition (8).

[5] Naturally, the map (10) is also positive for ς_l and ω_l^2 satisfying (14).



(a) (b) (c) (d)

Fig. 1. Sparse vector estimation results averaged over 50 independent noise realizations. (a) Normalized mean-square error (NMSE) versus the SNR. (b) The estimated number of components versus the SNR. (c) The convergence of the estimated weights versus the number of iterations for SNR = 10 dB. (d) The estimated number of components versus the number of iterations for SNR = 10 dB.

tained by repeatedly updating *ad infinitum* $q(\boldsymbol{w})$ and $q(\alpha_l)$ from (2) and (3), respectively. Then, $\hat{\alpha}_l^{[\infty]}$ given by (11) is a maximizer of the marginal likelihood function $p(\boldsymbol{t}|\alpha_l, \hat{\boldsymbol{\alpha}}_{\bar{l}}, \hat{\tau})$ with respect to the sparsity parameter α_l.

Proof: It has been shown [8] that the maximum of $\log p(\boldsymbol{t}|\alpha_l, \hat{\boldsymbol{\alpha}}_{\bar{l}}, \hat{\tau})$ with respect to α_l is obtained at

$$\hat{\alpha}_l = \begin{cases} s_l^2(q_l^2 - s_l)^{-1}, & q_l^2 > s_l \\ \infty, & q_l^2 \le s_l \end{cases} \tag{21}$$

where $s_l = \boldsymbol{\phi}_l^T \boldsymbol{C}_{\bar{l}}^{-1} \boldsymbol{\phi}_l$, $q_l = \boldsymbol{\phi}_l^T \boldsymbol{C}_{\bar{l}}^{-1} \boldsymbol{t}$, and $\boldsymbol{C}_{\bar{l}} = \hat{\tau}^{-1}\boldsymbol{I} + \sum_{k \neq l} \hat{\alpha}_k^{-1} \boldsymbol{\phi}_k \boldsymbol{\phi}_k^T$. Using the Woodbury matrix identity [16] applied to $\boldsymbol{C}_{\bar{l}}^{-1}$ it is easy to show that $\varsigma_l = s_l^{-1}$ and $\omega_l^2 = \frac{q_l^2}{s_l^2}$. Thus, (i) the pruning condition $q_l^2 > s_l$ in (21) is equivalent to the pruning condition $\omega_l^2 > \varsigma_l$ in (14), and (ii) the value of $\hat{\alpha}_l$ in (21) and that of $\hat{\alpha}_l^{[\infty]}$ in (11) coincide. ■

C. Algorithm Summary

We summarize the main steps of the proposed fast variational SBL scheme in Algorithm 1. It should be mentioned that updating $q(\tau)$ requires recomputing the $L \times L$ covariance matrix $\hat{\boldsymbol{S}}$, an $O(L^3)$ operation. Notice that for both the FMLM algorithm [8] and the fast variational SBL algorithm the functions $\boldsymbol{\phi}_l$ can theoretically be processed in any order. However, the exact order in which basis functions $\boldsymbol{\phi}_l$ are processed matters since different update protocols can lead to different local optima. In our work we first update components with large values of $\hat{\alpha}_l$, i.e., those basis functions that are least well aligned with the measurement $\boldsymbol{t}$. This might potentially reduce the dimensionality of the model already at early iterations.

Algorithm 1

```
Initialize $q(\boldsymbol{w})$, $q(\boldsymbol{\alpha})$, $q(\tau)$
while Continue if not converged do
  for $l \in \{1, \ldots, L\}$ do
    Compute: $\varsigma_l$ and $\omega_l^2$ from (19)
    if $\omega_l^2 > \varsigma_l$ then
      $\hat{\alpha}_l^{[\infty]} = \frac{1}{(\omega_l^2 - \varsigma_l)}$, update $\hat{\boldsymbol{S}}$ from (20)
    else
      $\hat{\boldsymbol{S}} = \hat{\boldsymbol{S}}_{\bar{l}}$, $\hat{\boldsymbol{\alpha}} = [\hat{\boldsymbol{\alpha}}]_{\bar{l}}$, $L = L - 1$;
    end if
  end for
  Compute $\hat{\boldsymbol{w}}$ from (2), $q(\tau)$ from (4) and recompute $\hat{\boldsymbol{S}}$ from (2)
  Check for convergence
end while
```

For algorithm initialization we use the following simple procedure. First, the initial value for the mean $\hat{\tau}$ of the noise pdf $q(\tau)$ is chosen. Then, the parameters of $q(\boldsymbol{w})$ are initialized as $\hat{\boldsymbol{S}} = (\hat{\tau}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \hat{\tau}^{-1}\boldsymbol{I})^{-1}$ and $\hat{\boldsymbol{w}} = \hat{\tau}\hat{\boldsymbol{S}}\boldsymbol{\Phi}^T\boldsymbol{t}$; finally, parameters of $q(\alpha_l), l = 1, \ldots, L$, are found using (3). Such initialization also provides an initial ranking of the components $\boldsymbol{\phi}_l$ based on the values of $\hat{\boldsymbol{\alpha}}$, which is then used to define a sequence of basis function updates used by the FMLM and the fast variational SBL algorithms. Notice that the fast variational SBL can also start with an empty model and add basis functions sequentially, each time testing whether a new basis function should be retained in the model or pruned.

D. Simulation Results

In this section we compare the simulation results of our proposed algorithm (FV-SBL) with the RVM algorithm [2] that does not use EM-based marginal likelihood maximization, the variational RVM algorithm (V-RVM) [10], and parameter-expanded variational Bayesian inference (PX-VB) [15]. Due to space limitations only two experiments are studied. In the first experiment we consider a sparse vector estimation problem with random $\boldsymbol{\Phi}$ and a fixed number of nonzero weights. In the other experiment we test the algorithms using the Concrete Compressive Strength (CCS) data set [17] from the UCI Machine Learning Repository [18]; this is a multivariate regression data set with 8 attributes and 1030 instances.

RVM, V-RVM and PX-VB methods require one to specify a pruning threshold for the sparsity parameters $\hat{\boldsymbol{\alpha}}$; specifically, when some $\hat{\alpha}_l$ exceeds the pruning threshold, the corresponding basis function is removed from the model. In all simulations we set this threshold to 10^{12}, as suggested in [2]. Obviously the estimated sparsity depends on a particular choice of this threshold. In contrast, for FV-SBL and FMLM methods the divergence of sparsity parameters is detected using closed form pruning conditions.

For all methods the same convergence criteria has been used: the algorithm stops i) when the number of basis functions between two consecutive iterations has stabilized and ii) when the ℓ^2-norm of the difference between the values of hyperparameters at two consecutive iterations is less than 10^{-5}. In the first experiment we also interrupt the algorithms when the number of iterations exceeds 10^4. Also, to simplify the analysis of the simulation results we assume $q(\tau)$ to be known and fixed in all simulations.

We begin now with the analysis of the synthetic data experiment. To generate data for sparse vector estimation we construct a random design matrix $\boldsymbol{\Phi} \in \mathbb{R}^{N \times N}$ by drawing N^2 samples from a standard Gaussian distribution; the target vector $\boldsymbol{t}$ is then generated according to (1) with the weight vector $\boldsymbol{w}$ having only 5 nonzero elements equal to 1 at random locations. We also test the performance of the fast variational SBL with adjusted pruning condition (16) (FV-SBL) by simulating an oracle estimator that knows the true signal SNR; the true SNR is then

TABLE I
PERFORMANCE RESULTS FOR CONCRETE COMPRESSIVE STRENGTH DATA

	#iterations	NMSE (dB)	#basis functions
RVM	1774	-15.74	66
V-RVM	(10^5)	-16.06	722
PX-VB	294	$\mathbf{-16.96}$	722
FV-SBL/FMLM	13	-15.56	55
FV-SBL ($\mathrm{SNR}_l' = 10\mathrm{dB}$)	**6**	-14.41	**31**

used as the SNR_l' adjustment in (16). The corresponding simulation results are summarized in Fig. 1. It can be seen from the plots in Fig. 1(a) and (b) that FV-SBL with adjusted pruning condition is able to estimate the true sparsity of the signal, achieving the smallest normalized mean-square error (NMSE) almost over the whole tested SNR range. V-RVM as well as PX-VB, although they perform well in terms of the reached NMSE, do not produce sparse estimates. The reason for that is a very high threshold value: V-RVM requires many more than 10^4 iterations for the hyperparameters to reach the 10^{12} pruning threshold; PX-VB converges faster than V-RVM [see Fig. 1(c)], but lands in a local nonsparse optimum of the variational objective function. FV-SBL without SNR-based adjustment and FMLM perform similar to RVM, V-RVM and PX-VB methods in terms of NMSE and slightly better in terms of the number of estimated components. In Fig. 1(c) and (d), we demonstrate the convergence properties of the algorithms for SNR fixed at 10 dB. Observe that FV-SBL clearly outperform other estimation schemes; FV-SBL with adjusted pruning criterion converges extremely rapidly (in roughly three to four iterations). PX-VB converges faster than RVM and V-RVM, but does not lead to a sparse estimate with the used pruning threshold.

Now, let us discuss the CCS data set. The data is normalized to zero mean and unit variance; 70% of the data were picked at random for training and the remaining set was used for testing. The design matrix $\mathbf{\Phi}$ consisted of a constant bias term $\boldsymbol{\phi}_0 = [1, \ldots, 1]^T$ and N Gaussian kernels $\boldsymbol{\phi}_l$ centered at the measurement samples. The variance of the additive noise $\hat{\tau}^{-1}$ as well as the variance of the Gaussian kernels η^2 were estimated using cross-validation with the FV-SBL algorithm; these parameters were then fixed at the estimated values $\hat{\tau}^{-1} = 0.1$ and $\eta^2 = 4.3$ for all compared methods. For the FV-SBL method with the adjusted pruning criterion we used $\mathrm{SNR}_l' = 10\,\mathrm{dB}$ for all components. The corresponding performance results are summarized in Table I. Here again the FV-SBL approach outperforms the other methods: it achieves very sparse results with only 55 basis functions in 13 iterations, only marginally losing in NMSE as compared to RVM, V-RVM and PX-VB. The latter method achieves the best NMSE and is second to only FV-SBL schemes in terms of convergence rate; however, it does not lead to a sparse solution. Observe that the V-RVM algorithm is interrupted after 10^5 iterations. Although the sparsity parameters α_l continue to diverge, the rate of divergence is very low, which prevents them from reaching the used pruning threshold. Notice also that FV-SBL with adjusted pruning condition leads to the sparsest estimator and converges very rapidly, but nonetheless loses in the achieved NMSE. This happens due to the fact that SNR_l' in (16) has been chosen to be equal for all basis functions, which might not be valid for the CCS data set. As a result, this choice leads to an overly sparse estimator that removes relevant basis functions.

IV. CONCLUSION

In this work, a fast variational SBL framework has been considered. For the SBL problem with a Gaussian likelihood model and Gaussian sparsity priors the stationary points of sparsity parameter update expressions as well as conditions that guarantee convergence to these stationary points—pruning conditions—have been obtained in a closed form. This eliminates the need to iterate the variational update expressions and boosts the convergence rate of the algorithm. It has been shown that for the case of noninformative hyperpriors, the mean of the sparsity parameter pdf that maximizes the variational lower bound also maximizes the marginal likelihood function with respect to this sparsity parameter; furthermore, the corresponding pruning conditions for fast variational SBL and the fast marginal likelihood maximization method are equivalent. However, in contrast to marginal likelihood maximization method, the pruning conditions obtained with the fast variational method reveal the relationship between the sparsity properties of SBL and a measure of SNR. This relationship enables an adjustment of the pruning condition such that only the components with a predefined quality (in terms of SNR) are retained in the model. Simulation studies demonstrate that this adjustment allows for an estimation of the true signal sparsity in simulated scenarios and further accelerates the convergence rate of the algorithm.

REFERENCES

[1] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM J. Sci. Comput.*, vol. 20, pp. 33–61, 1998.

[2] M. Tipping, “Sparse Bayesian learning and the relevance vector machine,” *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, Jun. 2001.

[3] M. Wakin, “An introduction to compressive sampling,” *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[4] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, “The variational approximation for Bayesian inference,” *IEEE Signal Process. Mag.*, vol. 25, no. 6, pp. 131–146, Nov. 2008.

[5] W. Bajwa, J. Haupt, A. Sayeed, and R. Nowak, “Compressed channel sensing: A new approach to estimating sparse multipath channels,” *Proc. IEEE*, vol. 98, no. 6, pp. 1058–1076, Jun. 2010.

[6] D. Wipf and B. Rao, “Sparse Bayesian learning for basis selection,” *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2164, Aug. 2004.

[7] R. Neal, *Bayesian Learning for Neural Networks*, ser. Lecture Notes in Stat.. New York: Springer-Verlag, 1996, vol. 118.

[8] M. E. Tipping and A. C. Faul, “Fast marginal likelihood maximisation for sparse Bayesian models,” presented at the 9th Int. Workshop Artif. Intell. Stat., Key West, FL, Jan. 2003.

[9] S. Babacan, R. Molina, and A. K. Katsaggelos, “Bayesian compressive sensing using Laplace priors,” *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 53–63, Jan. 2010.

[10] C. M. Bishop and M. E. Tipping, “Variational relevance vector machines,” in *Proc. 16th Conf. Uncertainty in Artif. Intell. (UAI)* , San Francisco, CA, 2000, pp. 46–53.

[11] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[12] D. Shutin and B. H. Fleury, “Sparse variational Bayesian SAGE algorithm with application to the estimation of multipath wireless channels,” *IEEE Trans. Signal Process.*, vol. 59, no. 8, pp. 3609–3623, Aug. 2011.

[13] D. Shutin, H. Zheng, B. H. Fleury, S. R. Kulkarni, and H. V. Poor, “Space-alternating attribute-distributed sparse learning,” in *Proc. 2nd Int Cognitive Inf. Process. (CIP) Workshop*, 2010, pp. 209–214.

[14] J. Luttinen, A. Ilin, and T. Raiko, “Transformations for variational factor analysis to speed up learning,” presented at the Eur. Symp. Artif. Neural Netw.—Adv. Comput. Intell. Learn., Bruges, Belgium, Apr. 22–24, 2009.

[15] Y. Qi and T. Jaakkola, “Parameter expanded variational Bayesian methods,” in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2007, vol. 19.

[16] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: The Johns Hopkins Univ. Press, 1996.

[17] I. C. Yeh, “Modeling of strength of high-performance concrete using artificial neural networks,” *Cement Concrete Res.*, vol. 28, no. 12, pp. 1797–1808, 1998.

[18] A. Frank and A. Asuncion, UCI Machine Learning Repository, 2010 [Online]. Available: http://archive.ics.uci.edu/ml