A Variational Bayes Framework for Sparse Adaptive Estimation

Konstantinos E. Themelis, Athanasios A. Rontogiannis, *Member, IEEE*, and Konstantinos D. Koutroumbas

***Abstract*—Recently, a number of mostly ℓ_1-norm regularized least-squares-type deterministic algorithms have been proposed to address the problem of *sparse adaptive* signal estimation and system identification. From a Bayesian perspective, this task is equivalent to maximum *a posteriori* probability estimation under a sparsity promoting heavy-tailed prior for the parameters of interest. Following a different approach, this paper develops a unifying framework of sparse *variational Bayes* algorithms that employ heavy-tailed priors in conjugate hierarchical form to facilitate posterior inference. The resulting fully automated variational schemes are first presented in a batch iterative form. Then, it is shown that by properly exploiting the structure of the batch estimation task, new sparse adaptive variational Bayes algorithms can be derived, which have the ability to impose and track sparsity during real-time processing in a time-varying environment. The most important feature of the proposed algorithms is that they completely eliminate the need for computationally costly parameter fine-tuning, a necessary ingredient of sparse adaptive deterministic algorithms. Extensive simulation results are provided to demonstrate the effectiveness of the new sparse adaptive variational Bayes algorithms against state-of-the-art deterministic techniques for adaptive channel estimation. The results show that the proposed algorithms are numerically robust and exhibit in general superior estimation performance compared to their deterministic counterparts.**

***Index Terms*—Sparse adaptive estimation, online variational Bayes, Bayesian models, sparse Bayesian learning, Student-t distribution, Laplace distribution, generalized inverse Gaussian distribution, Bayesian inference.**

I. Introduction

ADAPTIVE estimation of time-varying signals and systems is a research field that has attracted tremendous attention in the statistical signal processing literature, has triggered extensive research, and has had a great impact in a plethora of applications [1], [2]. A large number of adaptive estimation techniques have been developed and analyzed during the past decades, which have the ability to process streaming data and provide real-time estimates of the parameters of interest in an online fashion. It has long ago been recognized that apart from being time-varying, most signals and systems, both natural and man-made, also admit a parsimonious or so-called *sparse* representation in a certain domain. This fact has nowadays sparked new interest in the area of adaptive estimation, as the recent advances and tools developed in the compressive sensing (CS) field [3], [4], provide the means to effectively exploit *sparsity* in a time-varying environment. It has been anticipated that by suitably exploiting signal sparsity, significant improvements in convergence rate and estimation performance of adaptive techniques could be achieved. It should be noted that conventional CS deals with the problem of estimating a time-invariant sparse signal using less measurements than the size of the signal. On the other hand, in sparse adaptive estimation, a sparse time-varying signal is estimated time-recursively, by exploiting its sparsity as new measurement data become available.

It is not surprising that the majority of sparsity aware adaptive estimation methods developed so far, stem from a deterministic framework. Capitalizing on the celebrated least absolute shrinkage and selection operator (lasso) [5], an ℓ_1 regularization term is introduced in the cost function of these methods. In this context, by incorporating an ℓ_1 (or a log-sum) penalty term in the cost function of the standard least mean square (LMS) algorithm, adaptive LMS algorithms that are able to recursively identify sparse systems are derived in [6]. Inclusion of an ℓ_1 regularization factor or a more general regularizing term in the least squares (LS) cost function has also been proposed in [7] and [8], respectively. In [7] adaptive coordinate-descent type algorithms are developed with sparsity being imposed via soft-thresholding, while in [8] recursive LS (RLS) type schemes are designed. An ℓ_1 regularized RLS type algorithm that utilizes the expectation maximization (EM) algorithm as a low-complexity solver is described in [9]. In addition, adaptive identification of sparse nonlinear Volterra-type systems is presented in [10], by suitably combining EM with Kalman filtering. From such a general setting, several sparse variants, including RLS, LMS and fast RLS schemes are then derived. In a different spirit, a subgradient projection-based adaptive algorithm that induces sparsity using projections on weighted ℓ_1 balls is developed and analyzed in [11]. Adaptive greedy variable selection schemes have been also recently reported, e.g., [12]. However, these algorithms require, at least, a rough knowledge of the signal sparsity level and work effectively for sufficiently high signal sparsity.

In this paper, we depart from the deterministic setting adopted so far in previous works and deal with the sparse adaptive estimation problem within a Bayesian framework. In such a frame-

Manuscript received January 10, 2014; revised May 02, 2014; accepted July 06, 2014. Date of publication July 11, 2014; date of current version August 14, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tareq Al-Naffouri. This research has been co-financed by the European Union (European Social Fund-ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF)-Research Funding Program: Thales. Investing in knowledge society through the European Social Fund.

The authors are with the Institute for Astronomy, Astrophysics, Space Applications and Remote Sensing (IAASARS), National Observatory of Athens, GR-15236, Penteli, Greece (e-mail: themelis@noa.gr; tronto@noa.gr; koutroum@noa.gr).

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.2014.2338839

work, a Bayesian model is first defined comprising, a) a likelihood function specified by the assumed measurement data generation process and b) prior distributions for all model parameters, (which are thus considered as random variables), properly chosen to adhere to the constraints of the problem. In particular, to induce sparsity, suitable heavy-tailed sparsity promoting priors are assigned to the weight parameters of interest. Then a variational Bayesian inference method is utilized to approximate the joint posterior distribution of all model parameters, from which estimates of the sought parameters can be obtained via suitably defined iterative algorithms. It should be emphasized though that the various Bayesian inference methods are designed to solve the *batch* estimation problem, i.e., they provide the parameter estimates based on a given fixed size block of data and observations.

In the context described above, the contribution of this work is twofold. First, we provide a unified derivation of a family of Bayesian batch estimation techniques. Such a derivation passes through a) the selection of a generalized prior distribution for the *sparsity inducing parameters* of the model and b) the adoption of the mean-field variational approach [13]–[15] to perform Bayesian inference. The adopted *fully* factorized variational approximation method relies on an independence assumption on the joint posterior of *all* involved model parameters and leads to simple sparsity aware iterative batch estimation schemes with proven convergence. The derivation of the above batch estimation algorithms constitutes the prerequisite step that paves the way for the deduction of the novel *adaptive* variational Bayes algorithms, which marks the second contribution and main objective of this work. The proposed adaptive algorithms consist of two parts, namely, a common part encompassing time update formulas of the basic model parameters and a sparsity enforcing mechanism, which depends on the various Bayesian model priors assumed. The algorithms are numerically robust and are based on second order statistics having a computational complexity similar to that of other related sparsity aware deterministic schemes. Moreover, extensive simulations under various time-varying conditions show that they converge faster to sparse solutions and offer, in principle, lower steady-state estimation error compared to existing algorithms. The major advantage, though, of the proposed algorithms is that thanks to their Bayesian origin, they are fully automated (after certain hyperparameters at the highest level of the model are fixed to values close to zero, as is typically done in sparse Bayesian learning). Hence, while related sparse deterministic algorithms (in order to achieve optimum performance) involve application- and conditions-dependent regularization parameters that need to be predetermined via exhaustive fine-tuning, the Bayesian algorithms presented in this paper directly infer all model parameters from the data, and hence, the need for parameter fine-tuning is entirely eliminated. This, combined with their robust sparsity inducing properties, makes them particularly attractive for use in practice[1] (Preliminary versions of parts of this work have been presented in [17], [18]).

The rest of the paper is organized as follows. Section II defines the mathematical formulation of the adaptive estimation problem from a LS point of view. In Section III the adopted hierarchical Bayesian model is described. A family of batch variational Bayes iterative schemes is presented in Section IV. The new sparse adaptive variational Bayes algorithms are developed in Section V. In Section VI an analysis of the proposed algorithms is presented and their relation to other known algorithms is established. Extensive experimental results are provided in Section VII and concluding remarks are given in Section VIII.

Notation: Column vectors are represented as boldface lowercase letters, e.g., $\mathbf{x}$, and matrices as boldface uppercase letters, e.g., $\mathbf{X}$, while the i-th component of vector $\mathbf{x}$ is denoted by x_i and the ij-th element of matrix $\mathbf{X}$ by x_{ij}. Moreover, $(\cdot)^T$ denotes transposition, $\|\cdot\|_1$ stands for the ℓ_1-norm, $\|\cdot\|$ stands for the standard ℓ_2-norm, $|\cdot|$ denotes the determinant of a matrix or absolute value in case of a scalar, $\mathcal{N}(\cdot)$ is the Gaussian distribution, $\mathcal{G}(\cdot)$ is the Gamma distribution, $\mathcal{IG}(\cdot)$ is the inverse Gamma distribution, $\mathcal{GIG}(\cdot)$ is the generalized inverse Gaussian distribution, $\Gamma(\cdot)$ is the Gamma function, $\langle\cdot\rangle$ is the expectation operator, $\mathrm{diag}(\mathbf{x})$ denotes a diagonal matrix whose diagonal entries are the elements of $\mathbf{x}$, and $\mathrm{diag}(\mathbf{X})$ is a column vector containing the main diagonal elements of a square matrix $\mathbf{X}$. Finally, we use the semicolon $(;)$ and the vertical bar $(|)$ characters to express the dependence of a random variable on parameters and other random variables, respectively.

II. Problem Statement

Let $\mathbf{w}(n) = [w_1(n), w_2(n), \ldots, w_N(n)]^T \in \mathbb{R}^N$ denote a *sparse* time-varying weight vector having $\xi \ll N$ non-zero elements, where n is the time index. We wish to estimate and track $\mathbf{w}(n)$ in time by observing a stream of sequential data which are assumed to obey to the following linear regression model,

$$y(n) = \mathbf{x}^T(n)\mathbf{w}(n) + \epsilon(n), \tag{1}$$

where $\mathbf{x}(n) = [x_1(n), x_2(n), \ldots, x_N(n)]^T$ is a *known* $N \times 1$ regression vector, and $\epsilon(n)$ denotes the uncorrelated with $\mathbf{x}(n)$ added Gaussian noise of zero mean and variance β^{-1} (or precision β), i.e., $\epsilon(n) \sim \mathcal{N}(\epsilon(n)|0, \beta^{-1})$. The linear data generation model given in (1) fits very well or, at least, approximates adequately the hidden mechanisms in many signal processing tasks. Let

$$\mathbf{y}(n) = [y(1), y(2), \ldots, y(n)]^T \tag{2}$$

and

$$\mathbf{X}(n) = [\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(n)]^T \tag{3}$$

be the $n \times 1$ vector of observations and the $n \times N$ input data matrix respectively, up to time n. Then, the unknown weight vector $\mathbf{w}(n)$ can be estimated by minimizing with respect to (w.r.t.) $\hat{\mathbf{w}}(n)$ the following exponentially weighted LS cost function[2],

$$\begin{aligned}\mathcal{J}_{\mathrm{LS}}(n) &= \sum_{j=1}^{n} \lambda^{n-j} |y(j) - \mathbf{x}^T(j)\hat{\mathbf{w}}(n)|^2 \\ &= \|\mathbf{\Lambda}^{1/2}(n)\mathbf{y}(n) - \mathbf{\Lambda}^{1/2}(n)\mathbf{X}(n)\hat{\mathbf{w}}(n)\|^2. \end{aligned} \tag{4}$$

The parameter $\lambda, 0 \ll \lambda \leq 1$, is commonly referred to as the *forgetting factor* (because it weights more heavily recent data and 'forgets' gradually old data), and

[1]Note that a Bayesian approach to adaptive *filtering* has been previously proposed in [16]. However, in [16] a type-II maximum likelihood inference method is adopted that leads to a regularized RLS-type scheme. This is completely different from the approach and algorithms described in this work.

[2]Note that a fixed size sliding in time data window could be also used.

$\mathbf{\Lambda}(n) = \mathrm{diag}([\lambda^{n-1}, \lambda^{n-2}, \ldots, 1]^T)$. It is well-known that the vector $\hat{\mathbf{w}}(n)$ that minimizes $\mathcal{J}_{\mathrm{LS}}(n)$ is given by the solution of the celebrated *normal equations*, [1]. In an adaptive estimation setting, the cost function in (4) can be optimized recursively in time by utilizing the RLS algorithm. The RLS algorithm, a) reduces the computational complexity from $\mathcal{O}(N^3)$, which is required for solving the normal equations per time iteration, to $\mathcal{O}(N^2)$, b) has constant memory requirements despite the fact that the size of the data grows with n, and, c) has the ability of tracking possible variations of $\mathbf{w}(n)$ as n increases.

However, the RLS algorithm does not specifically exploit the inherent sparsity of the parameter vector $\mathbf{w}(n)$, so as to improve its initial convergence rate and estimation performance. To deal with this issue, a number of adaptive deterministic LS-type algorithms have been recently proposed, e.g., [7]–[10]. In all these schemes, the LS cost function is supplemented with a regularization term that penalizes the ℓ_1-norm of the unknown weight vector, i.e.,

$$\mathcal{J}_{\mathrm{LS}-\ell_1}(n) = \|\mathbf{\Lambda}^{1/2}(n)\mathbf{y}(n) - \mathbf{\Lambda}^{1/2}(n)\mathbf{X}(n)\hat{\mathbf{w}}(n)\|^2 + \tau\|\hat{\mathbf{w}}(n)\|_1, \quad (5)$$

where $\tau > 0$ is a regularization parameter controlling the sparsity of $\hat{\mathbf{w}}(n)$, that should be properly selected. Regularization with the ℓ_1-norm has its origin in the widely known lasso operator, [5], and is known to promote sparse solutions.

In this paper, in contrast to previous studies, we provide an analysis of the sparse adaptive estimation problem from a Bayesian perspective. To this end, we derive a class of variational Bayes estimators that are built upon hierarchical Bayesian models featuring heavy-tailed priors. A basic characteristic of heavy-tailed priors is their sparsity inducing nature. These prior distributions are known to improve robustness of regression and classification tasks to outliers and have been widely used in variable selection problems, [19], [20]. The variational Bayesian inference approach adopted in this paper, a) exhibits low computational complexity compared to (the possible alternative) Markov Chain Monte Carlo (MCMC) sampling methods, [15], and b) performs inference for all model parameters, including the sparsity promoting parameter τ, as opposed to deterministic methods. In the following, we analyze a general hierarchical Bayesian model for the batch estimation problem first (i.e., when n is considered fixed), and then we show how the proposed variational Bayes inference method can be extended in an adaptive estimation setting[3].

III. Bayesian Modeling

To simplify the description of the hierarchical Bayesian model we temporarily drop the dependence of all model quantities from the time indicator n. Time dependency will be re-introduced in Section V, where the proposed adaptive variational schemes are presented. To consider the estimation problem at hand from a Bayesian point of view, we first define a likelihood function based on the given data generation model and then we introduce sparsity to our estimate by assigning a suitable heavy-tailed prior distribution over the parameter vector $\mathbf{w}$. In order to account for the exponentially weighted data windowing used in (4), the following observation model is considered

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon} \quad (6)$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{\varepsilon}|\mathbf{0}, \beta^{-1}\mathbf{\Lambda}^{-1})$. From this observation model and the statistics of the noise vector $\boldsymbol{\varepsilon}$, it turns out that the corresponding likelihood function is

$$p(\mathbf{y}|\mathbf{w}, \beta) = \frac{\beta^{\frac{n}{2}}|\mathbf{\Lambda}|^{\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}}\exp\left[-\frac{\beta}{2}\left\|\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{y} - \mathbf{\Lambda}^{\frac{1}{2}}\mathbf{X}\mathbf{w}\right\|^2\right]. \quad (7)$$

Notice that the maximum likelihood estimator of (7) coincides with the LS estimator that minimizes (4). However, as mentioned previously, our estimator should be further constrained to be sparse. To this end, the likelihood is complemented by suitable *conjugate* priors w.r.t. (7) over the parameters $\mathbf{w}$ and β, [22], [23]. The prior for the noise precision β is selected to be a Gamma distribution with parameters ρ and δ, i.e.,

$$p(\beta; \rho, \delta) = \mathcal{G}(\beta; \rho, \delta) = \frac{\delta^\rho}{\Gamma(\rho)}\beta^{\rho-1}\exp[-\delta\beta]. \quad (8)$$

Next, a hierarchical heavy-tailed prior is selected for the parameter vector $\mathbf{w}$, that reflects our knowledge that many of its components are zero or nearly zero. In the first level of hierarchy, a Gaussian prior is attached on $\mathbf{w}$, i.e.,

$$p(\mathbf{w}|\boldsymbol{\alpha}, \beta) = \mathcal{N}\left(\mathbf{w}|0, \beta^{-1}\mathbf{A}^{-1}\right) = \prod_{i=1}^{N} p\left(w_i|\alpha_i, \beta\right) = \prod_{i=1}^{N}(2\pi)^{-\frac{1}{2}}\beta^{\frac{1}{2}}\alpha_i^{\frac{1}{2}}\exp\left[-\frac{\beta}{2}w_i^2\alpha_i\right]. \quad (9)$$

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_N]^T$ is the vector of the precision parameters of the w_i's, $\mathbf{A} = \mathrm{diag}(\boldsymbol{\alpha})$ and the w_i's have been assumed a priori independent. Now, depending on the choice of the prior distribution for the precision parameters in $\boldsymbol{\alpha}$ at the second level of hierarchy, various heavy-tailed distributions may arise for $\mathbf{w}$, such as the Student-t or the Laplace distribution. To provide a unification of all these distributions in a single model, we assume that the sparsity enforcing parameters α_i follow a generalized inverse Gaussian (GIG) distribution, expressed as[4]

$$p(\alpha_i; a, b_i, c) = \mathcal{GIG}(\alpha_i; a, b_i, c) = \frac{(a/b_i)^{(c/2)}}{2\mathcal{K}_c(\sqrt{ab_i})}\alpha_i^{c-1}\exp[-\frac{1}{2}(a\alpha_i + \frac{b_i}{\alpha_i})], \quad (10)$$

where $a, b_i > 0, c \in \mathbb{R}$ and $\mathcal{K}_c(\,\cdot\,)$ is the modified Bessel function of the second kind. In this paper, hypersparametes a, c and b_i's in (10) are selected so as to formulate the widely used sparsity promoting heavy-tailed Student-t and Laplace priors, e.g., [22], [25], [26]. In particular, in order to infer the sparsity regularizing parameters b_i's from the data, these are assumed to follow a Gamma distribution with parameters κ and ν, i.e.,

$$p(b_i; \kappa, \nu) = \mathcal{G}(b_i; \kappa, \nu) = \frac{\nu^\kappa}{\Gamma(\kappa)}b_i^{\kappa-1}\exp[-\nu b_i]. \quad (11)$$

A directed acyclic graph (DAG) of the proposed hierarchical Bayesian model is shown in Fig. 1, where $\mathbf{b} = [b_1, b_2, \ldots, b_N]^T$. Note that the hyperparameters κ, ν, ρ, and δ at the highest level are set close to zero so as to create

[3]Departing from sparse adaptive estimation, an *online* variational Bayes algorithm for model selection has been presented in [21]. This is the first work to deploy variational Bayes in a "non-batch" setting.

[4]More general models are reported in [24].

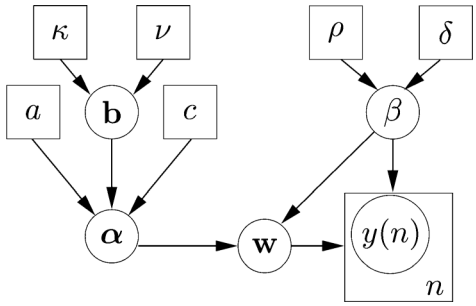

Fig. 1. DAG of the proposed Bayesian model.

(almost) non-informative Jeffreys priors ($p(x) \propto 1/x$) for β and b_i's, [22], [27]. This distribution expresses prior ignorance and allows for the *parameter-free* estimation of β and b_i's directly from the data. Notice also in Fig. 1 the dependence of $\mathbf{w}$ on β, which is due to the normalization by β of the variances of w_i's in (9). It can be shown that this normalization ensures the unimodality of the posterior joint distribution, [23], and leads to simpler and more compact parameter update expressions, as will be seen later.

IV. Mean-Field Variational Bayesian Inference

So far we have presented a generative model for the observations data (6) and a hierarchical Bayesian model (8), (9), (10), (11) treating the model parameters as random variables. To proceed with Bayesian inference, the computation of the joint posterior distribution over the model parameters is required[5]. Using Bayes' law, this distribution is expressed as

$$p(\mathbf{w},\beta,\boldsymbol{\alpha},\mathbf{b}|\mathbf{y}) = \frac{p(\mathbf{y},\mathbf{w},\beta,\boldsymbol{\alpha},\mathbf{b})}{\int p(\mathbf{y},\mathbf{w},\beta,\boldsymbol{\alpha},\mathbf{b}) d\mathbf{w} d\beta d\boldsymbol{\alpha} d\mathbf{b}}. \quad (12)$$

However, due to the complexity of the model, we cannot directly compute the posterior of interest, since the integral in (12) can not be expressed in closed form. Thus, we resort to approximations. In this paper, we adopt the variational framework, [13], [14], [29]–[31], to approximate the posterior in (12) with a simpler, *variational* distribution $q(\mathbf{w},\beta,\boldsymbol{\alpha},\mathbf{b})$. From an optimization point of view, the parameters of $q(\mathbf{w},\beta,\boldsymbol{\alpha},\mathbf{b})$ are selected so as to minimize the Kullback-Leibler divergence metric between the true posterior $p(\mathbf{w},\beta,\boldsymbol{\alpha},\mathbf{b}|\mathbf{y})$ and the variational distribution $q(\mathbf{w},\beta,\boldsymbol{\alpha},\mathbf{b})$, [30]. This minimization is equivalent to maximizing the *evidence lower bound* (ELBO) (which is a lower bound on the logarithm of the data marginal likelihood $\log p(\mathbf{y})$) w.r.t. the variational distribution $q(\mathbf{w},\beta,\boldsymbol{\alpha},\mathbf{b})$, [31]. Based on the mean-field theory from statistical physics, [32], we constrain $q(\mathbf{w},\beta,\boldsymbol{\alpha},\mathbf{b})$ to the family of distributions, which are *fully* factorized w.r.t. their parameters yielding

$$\begin{aligned} q(\mathbf{w},\beta,\boldsymbol{\alpha},\mathbf{b}) &= q(\mathbf{w})q(\beta)q(\boldsymbol{\alpha})q(\mathbf{b}) \\ &= \prod_{i=1}^{N} q(w_i)q(\beta)\prod_{i=1}^{N} q(\alpha_i)\prod_{i=1}^{N} q(b_i), \end{aligned} \quad (13)$$

[5]An alternative approach is the solution of the MAP problem defined by the presented Bayesian model. For such a problem, exact solvers exist, e.g., the iterative re-weighted least squares method, as explained in [28].

i.e., all model parameters are assumed to be *a posteriori* independent. This fully factorized form of the approximating distribution $q(\mathbf{w},\beta,\boldsymbol{\alpha},\mathbf{b})$ turns out to be very convenient, mainly because it results to an optimization problem that is computationally tractable. In fact, if we let θ_i denote the i-th component of the vector $\boldsymbol{\theta} = [w_1,\ldots,w_N,\beta,\alpha_1,\ldots,\alpha_N,b_1,b_2,\ldots,b_N]^T$ containing the parameters of the Bayesian hierarchical model, maximization of the ELBO results in the following expression for $q(\theta_i)$, [15],

$$q(\theta_i) = \frac{\exp\left[\langle \log p(\mathbf{y},\boldsymbol{\theta})\rangle_{j\neq i}\right]}{\int \exp\left[\langle \log p(\mathbf{y},\boldsymbol{\theta})\rangle_{j\neq i}\right] d\theta_i}, \quad (14)$$

where $\langle\cdot\rangle_{j\neq i}$ denotes the expectation w.r.t. $\prod_{j\neq i} q(\theta_j)$. Note that this is not a closed form solution, since every factor $q(\theta_i)$ depends on the remaining factors $q(\theta_j)$, for $j \neq i$. However, the interdependence between the factors $q(\theta_i)$ gives rise to a cyclic optimization scheme, where the factors are initialized appropriately, and each one is then iteratively updated via (14), by holding the remaining factors fixed. Each update cycle is known to increase the ELBO until convergence, [31].

Applying (14) to the proposed model (exact computations are reported in Appendix A), the approximating distribution for each coordinate $w_i, i = 1,2,\ldots,N$, is found to be Gaussian, $\mathcal{N}(w_i;\mu_i,\sigma_i^2)$,

$$q(w_i) = \frac{1}{\sigma_i\sqrt{2\pi}}\exp\left[-\frac{1}{2}\frac{(w_i-\mu_i)^2}{\sigma_i^2}\right], \quad (15)$$

with parameters μ_i and σ_i^2 given by

$$\sigma_i^2 = \langle\beta\rangle^{-1}\left(\mathbf{x}_i^T\boldsymbol{\Lambda}\mathbf{x}_i + \langle\alpha_i\rangle\right)^{-1}, \quad (16)$$

$$\mu_i = \langle\beta\rangle\sigma_i^2\mathbf{x}_i^T\boldsymbol{\Lambda}(\mathbf{y} - \mathbf{X}_{\neg i}\boldsymbol{\mu}_{\neg i}). \quad (17)$$

In (17), $\mathbf{X}_{\neg i}$ results from the data matrix $\mathbf{X}$ after removing its i-th column $\mathbf{x}_i$, μ_i is the posterior mean value of w_i, $\boldsymbol{\mu}_{\neg i}$ results from $\boldsymbol{\mu} = [\mu_1,\mu_2,\ldots,\mu_N]^T$ after the exclusion of its i-th element, and expectation $\langle\cdot\rangle$ is w.r.t. the variational distributions $q(\,\cdot\,)$ of the parameters appearing within each pair of brackets. Notice that, since each element w_i of $\mathbf{w}$ is treated separately, $q(w_i)$ constitutes an individual factor in the right hand side (RHS) of (13), as opposed to having a single compact factor $q(\mathbf{w})$ for the whole vector $\mathbf{w}$, as e.g., in [33]; this is beneficial for the development of the adaptive schemes that will be presented in the next Section. Working in a similar manner for the noise precision β, we get that $q(\beta)$ is a Gamma distribution expressed as

$$q(\beta) = \mathcal{G}(\beta;\tilde{\rho},\tilde{\delta}) = \frac{\tilde{\delta}^{\tilde{\rho}}}{\Gamma(\tilde{\rho})}\beta^{\tilde{\rho}-1}\exp[-\tilde{\delta}\beta], \quad (18)$$

with $\tilde{\rho} = \frac{n+N}{2} + \rho$ and $\tilde{\delta} = \delta + \frac{1}{2}\left\langle\|\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{y} - \boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{X}\mathbf{w}\|^2\right\rangle + \frac{1}{2}\left\langle\mathbf{w}^T\boldsymbol{\Lambda}\mathbf{w}\right\rangle$. Thus, the mean value of β w.r.t. (18) is expressed as

$$\langle\beta\rangle = \frac{\frac{n+N}{2} + \rho}{\delta + \frac{1}{2}\left\langle\|\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{y} - \boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{X}\mathbf{w}\|^2\right\rangle + \frac{1}{2}\sum_{i=1}^{N}\langle w_i^2\rangle\langle\alpha_i\rangle}. \quad (19)$$

In addition, since $\langle w_i^2\rangle = \mu_i^2 + \sigma_i^2$, it can be easily shown that the middle term in the denominator of the RHS of (19) is evaluated as

$$\left\langle\|\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{y} - \boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{X}\mathbf{w}\|^2\right\rangle = \|\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{y} - \boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{X}\boldsymbol{\mu}\|^2 + \sum_{i=1}^{N}\sigma_i^2\mathbf{x}_i^T\boldsymbol{\Lambda}\mathbf{x}_i. \quad (20)$$

The variational distribution of the precision parameters α_i's also turns out to be a generalized inverse Gaussian distribution given by,

$$q(\alpha_i) = \mathcal{GIG}\left(\alpha_i; a + \langle\beta\rangle\left\langle w_i^2\right\rangle, \langle b_i\rangle, c + \frac{1}{2}\right), \tag{21}$$

for $i = 1, 2, \ldots, N$. Finally, the variational distribution of the sparsity regularizing parameters b_i's can be expressed as

$$q(b_i) \propto b_i^{\kappa-1-\frac{c}{2}} \exp\left[-(\nu + \frac{1}{2\alpha_i})b_i\right] \frac{1}{\mathcal{K}_c(\sqrt{ab_i})}, \tag{22}$$

for $i = 1, 2, \ldots, N$. Since our intention here is the development of sparse estimation schemes, in the following subsections, three special cases of the previously described general model are presented, that are based on the sparsity promoting Student-t and Laplace priors.

A. Batch Variational Bayes With a Student-t Prior

As mentioned in Section III, various sparsity inducing prior distributions may arise for $\mathbf{w}$ by exploiting the flexibility of the GIG prior for the precision parameters α_i's in (10). One such prior is obtained by selecting the limit case where the rate hyperparameter $\nu \to \infty$, which implies that $\mathbf{b} = \mathbf{0}$, and $c > 0$. A Gamma distribution then arises with scale parameter c and rate parameter $a/2$, i.e.,

$$p(\alpha_i; c, a) = \frac{(a/2)^c}{\Gamma(c)} \alpha_i^{c-1} \exp\left[-\frac{a}{2}\alpha_i\right], \tag{23}$$

for $i = 1, 2, \ldots, N$. If we integrate out the precision parameter $\boldsymbol{\alpha}$ from (9) using (23), it is easily verified that the two-level hierarchical prior defined by (9) and (23) is equivalent to assigning a Student-t distribution over the parameter vector $\mathbf{w}$, which depends only on the hyperparameters a and c, [25], [26]. Under this hierarchical prior, the variational posterior distributions for $\mathbf{w}$ and β are the same as in (15) and (18), while for the precision parameters α_i's the following Gamma distribution is now computed (the same distribution can also be derived by substituting $b_i = 0$ and $c > 0$ in (21))

$$q(\alpha_i) = \mathcal{G}(\alpha_i; \tilde{c}, \tilde{a}) \tag{24}$$

with $\tilde{c} = c + \frac{1}{2}$ and $\tilde{a} = \frac{\langle\beta\rangle\langle w_i^2\rangle + a}{2}$ for $i = 1, 2, \ldots, N$. Moreover, the mean of (24) is expressed as

$$\langle\alpha_i\rangle = \frac{2c+1}{a + \langle\beta\rangle\langle w_i^2\rangle}. \tag{25}$$

Note that owing to the conjugacy of our hierarchical Bayesian model, the variational distributions in (15), (18), and (21) are expressed in a standard exponential form. Notice also that the parameters of all variational distributions are expressed in terms of expectations of expressions of the other parameters. This gives rise to a variational iterative scheme, which involves updating (16), (17), for $i = 1, 2, \ldots, N$, (19) and (25), for $i = 1, 2, \ldots, N$, in a sequential manner. Due to the convexity of the factors $q(w_i)$, $q(\beta)$, and $q(\alpha_i)$, the variational Bayes algorithm converges to a sparse solution in a few cycles, [15]. The variational algorithm solves the batch estimation problem defined in (4), providing the mean $\boldsymbol{\mu}$ of the approximating posterior $q(\mathbf{w})$ as the final estimate of the sparse vector $\mathbf{w}$.

A summary of the sparse variational Bayes procedure is shown in Table I. The Table includes a description of the Bayesian model, the resulting variational distributions and the corresponding sparse variational Bayes Student-t based (SVB-S) iterative scheme. In SVB-S, besides ρ, δ, hyperparameters a, c at the highest level of the hierarchy are also fixed to values close to zero, thus giving rise to almost non-informative priors for α_i's, but retaining the sparsity promoting Student-t distribution for w_i's.

B. Batch Variational Bayes With Laplace Priors

Next, we adjust the model parameters of the GIG prior in (10) to create a sparsity inducing Laplace prior for the weights $\mathbf{w}$. This prior results by setting the hyperparameters $a = 0$ and $c = -1$ and by selecting a single variable $b \sim \mathcal{G}(b; \kappa, \nu)$ to replace all b_i's in (10). In this case, the following *inverse* Gamma distribution is obtained as a prior for the precision parameters α_i's,

$$p(\alpha_i|b) = \mathcal{IG}\left(\alpha_i|1, \frac{b}{2}\right) = \frac{b}{2}\alpha_i^{-2}\exp\left[-\frac{b}{2}\frac{1}{\alpha_i}\right] \tag{26}$$

for $i = 1, 2, \ldots, N$. As shown in Appendix B, if we integrate out $\boldsymbol{\alpha}$ from the hierarchical prior of $\mathbf{w}$ defined by (9) and (26), a sparsity promoting multivariate Laplace distribution arises for $\mathbf{w}$. In addition, it can be shown that the resulting Bayesian model preserves an equivalence relation with the lasso [5] in that its maximum a posteriori probability (MAP) estimator coincides with the vector that minimizes the lasso criterion [22], [34][6]. A summary of this alternative model accompanied by a description of the resulting sparse variational Bayes iterative scheme based on a Laplace prior (SVB-L), is shown in Table I. Note from (21) and (22) that the variational distributions $q(\alpha_i), i = 1, 2, \ldots, N$, and $q(b)$ now become

$$q(\alpha_i) = \mathcal{GIG}\left(\alpha_i; \langle\beta\rangle\langle w_i^2\rangle, \langle b\rangle, -1/2\right), \tag{27}$$

$$q(b) = \mathcal{G}(b; N + \kappa, \nu + \frac{1}{2}\sum_{i=1}^{N}\left\langle\frac{1}{\alpha_i}\right\rangle). \tag{28}$$

Moreover, the expressions of the means of α_i and b used in the variational updates are computed as

$$\langle\alpha_i\rangle = \sqrt{\frac{\langle b\rangle}{\langle\beta\rangle\langle w_i^2\rangle}}, \tag{29}$$

$$\langle b\rangle = \frac{N + \kappa}{\nu + \frac{1}{2}\sum_{i=1}^{N}\left\langle\frac{1}{\alpha_i}\right\rangle} \tag{30}$$

while the mean $\langle 1/\alpha_i\rangle$ w.r.t. $q(\alpha_i)$ given in (27) is expressed as

$$\left\langle\frac{1}{\alpha_i}\right\rangle \equiv \langle\gamma_i\rangle = \frac{1}{\langle\alpha_i\rangle} + \frac{1}{\langle b\rangle}. \tag{31}$$

As noted in [35], the single shrinkage parameter b of the Laplace prior penalizes both zero and non-zero coefficients equally and it is not flexible enough to express the variability of sparsity among the unknown weight coefficients. In many circumstances, this leads to limited posterior inference and, evidently, to poor estimation performance. Hence, utilizing the full parameter vector $\mathbf{b}$ and setting hyperparameters $a = 0$ and

[6]Note, however, that in [22], [34] a different, (in terms of the parameters that impose sparsity), model is described. Specifically, instead of the precisions α_i's of w_i's, their variances γ_i's are used, with $\gamma_i = \alpha_i^{-1}$, on which Gamma priors of the form $\mathcal{G}(\gamma_i|1, \frac{b}{2})$ are assigned.

TABLE I
THE SVB-S, SVB-L AND SVB-MPL SCHEMES

Sparse variational Bayes schemes		
Data likelihood	$p(\mathbf{y}\|\mathbf{w},\beta)=(2\pi)^{-\frac{n}{2}}\beta^{\frac{n}{2}}\|\mathbf{\Lambda}\|^{1/2}\exp\left[-\frac{\beta}{2}\|\mathbf{\Lambda}^{1/2}\mathbf{y}-\mathbf{\Lambda}^{1/2}\mathbf{X}\mathbf{w}\|^2\right]$	
Prior distributions	$p(\beta;\rho,\delta)=\mathcal{G}(\beta;\rho,\delta)=\frac{\delta^\rho}{\Gamma(\rho)}\beta^{\rho-1}\exp\left[-\delta\beta\right]$ $p(\mathbf{w}\|\boldsymbol{\alpha},\beta)=\mathcal{N}(\mathbf{w}\|0,\beta^{-1}\mathbf{A}^{-1})=\prod_{i=1}^{N}(2\pi)^{-\frac{1}{2}}\beta^{\frac{1}{2}}\alpha_i^{\frac{1}{2}}\exp\left[-\frac{\beta}{2}w_i^2\alpha_i\right]$	
	SVB-S	$p(\alpha_i;c,a)=\mathcal{G}(\alpha_i;c,a/2)=\frac{(a/2)^c}{\Gamma(c)}\alpha_i^{c-1}\exp\left[-\frac{a}{2}\alpha_i\right], i=1,2,\ldots,N$
	SVB-L	$p(\alpha_i\|b)=\mathcal{IG}(\alpha_i\|1,\frac{b}{2})=\frac{b}{2}\alpha_i^{-2}\exp\left[-\frac{b}{2}\frac{1}{\alpha_i}\right], i=1,2,\ldots,N$ $p(b;\kappa,\nu)=\mathcal{G}(b;\kappa,\nu)=\frac{\nu^\kappa}{\Gamma(\kappa)}b^{\kappa-1}\exp\left[-\nu b\right]$
	SVB-mpL	$p(\alpha_i\|b_i)=\mathcal{IG}(\alpha_i\|1,\frac{b_i}{2})=\frac{b_i}{2}\alpha_i^{-2}\exp\left[-\frac{b_i}{2}\frac{1}{\alpha_i}\right], i=1,2,\ldots,N$ $p(b_i;\kappa,\nu)=\mathcal{G}(b_i;\kappa,\nu)=\frac{\nu^\kappa}{\Gamma(\kappa)}b_i^{\kappa-1}\exp\left[-\nu b_i\right], i=1,2,\ldots,N$
Variational distributions	$q(\beta)=\mathcal{G}(\beta;\tilde{\rho},\tilde{\delta})$, with $\tilde{\rho}=\frac{n+N}{2}+\rho$ and $\tilde{\delta}=\delta+\frac{1}{2}\left\langle\|\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{y}-\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{X}\mathbf{w}\|^2\right\rangle+\frac{1}{2}\left\langle\mathbf{w}^T\mathbf{A}\mathbf{w}\right\rangle$ $q(w_i)=\mathcal{N}(w_i;\mu_i,\sigma_i^2)$, with $\sigma_i^2=\langle\beta\rangle^{-1}(\mathbf{x}_i^T\mathbf{\Lambda}\mathbf{x}_i+\langle\alpha_i\rangle)^{-1}$, and $\mu_i=\langle\beta\rangle\sigma_i^2\mathbf{x}_i^T\mathbf{\Lambda}(\mathbf{y}-\mathbf{X}_{\neg i}\boldsymbol{\mu}_{\neg i}), i=1,2,\ldots,N$	
	SVB-S	$q(\alpha_i)=\mathcal{G}(\alpha_i;\tilde{c},\tilde{a})$, with $\tilde{c}=c+\frac{1}{2}$ and $\tilde{a}=\frac{\langle\beta\rangle\langle w_i^2\rangle+a}{2}, i=1,2,\ldots,N$
	SVB-L	$q(\alpha_i)=\frac{\left(\langle\beta\rangle\langle w_i^2\rangle/\langle b\rangle\right)^{-1/4}}{2K_{1/2}\left(\sqrt{\langle\beta\rangle\langle w_i^2\rangle\langle b\rangle}\right)}\alpha_i^{-3/2}\exp\left[-\frac{1}{2}\langle\beta\rangle\langle w_i^2\rangle\alpha_i-\frac{\langle b\rangle}{2}\frac{1}{\alpha_i}\right], i=1,2,\ldots,N$ $q(b)=\mathcal{G}(b;\tilde{\kappa},\tilde{\nu})$, with $\tilde{\kappa}=N+\kappa$, and $\tilde{\nu}=\nu+\frac{1}{2}\sum_{i=1}^{N}\left\langle\frac{1}{\alpha_i}\right\rangle$
	SVB-mpL	$q(\alpha_i)=\frac{\left(\langle\beta\rangle\langle w_i^2\rangle/\langle b_i\rangle\right)^{-1/4}}{2K_{1/2}\left(\sqrt{\langle\beta\rangle\langle w_i^2\rangle\langle b_i\rangle}\right)}\alpha_i^{-3/2}\exp\left[-\frac{1}{2}\langle\beta\rangle\langle w_i^2\rangle\alpha_i-\frac{\langle b_i\rangle}{2}\frac{1}{\alpha_i}\right], i=1,2,\ldots,N$ $q(b_i)=\mathcal{G}(b_i;\tilde{\kappa},\tilde{\nu})$, with $\tilde{\kappa}=1+\kappa$ and $\tilde{\nu}=\nu+\frac{1}{2}\left\langle\frac{1}{\alpha_i}\right\rangle, i=1,2,\ldots,N$
Variational updates	$\langle\beta\rangle=(n+N+2\rho)/(2\delta+\left\langle\|\mathbf{\Lambda}^{1/2}\mathbf{y}-\mathbf{\Lambda}^{1/2}\mathbf{X}\mathbf{w}\|^2\right\rangle+\sum_{i=1}^{N}\langle w_i^2\rangle\langle\alpha_i\rangle)$ $\sigma_i^2=\langle\beta\rangle^{-1}(\mathbf{x}_i^T\mathbf{\Lambda}\mathbf{x}_i+\langle\alpha_i\rangle)^{-1}, i=1,2,\ldots,N$ $\langle w_i\rangle\equiv\mu_i=\langle\beta\rangle\sigma_i^2\mathbf{x}_i^T\mathbf{\Lambda}(\mathbf{y}-\mathbf{X}_{\neg i}\boldsymbol{\mu}_{\neg i}), i=1,2,\ldots,N$ $\langle w_i^2\rangle=\mu_i^2+\sigma_i^2, i=1,2,\ldots,N$	
	SVB-S	$\langle\alpha_i\rangle=\frac{2c+1}{a+\langle\beta\rangle\langle w_i^2\rangle}, i=1,2,\ldots,N$
	SVB-L	$\langle\alpha_i\rangle=\sqrt{\frac{\langle b\rangle}{\langle\beta\rangle\langle w_i^2\rangle}}, \left\langle\frac{1}{\alpha_i}\right\rangle\equiv\gamma_i=\frac{1}{\langle\alpha_i\rangle}+\frac{1}{\langle b\rangle}, i=1,2,\ldots,N$ $\langle b\rangle=\frac{N+\kappa}{\nu+\frac{1}{2}\sum_{i=1}^{N}\left\langle\frac{1}{\alpha_i}\right\rangle}$
	SVB-mpL	$\langle\alpha_i\rangle=\sqrt{\frac{\langle b_i\rangle}{\langle\beta\rangle\langle w_i^2\rangle}}, \quad \left\langle\frac{1}{\alpha_i}\right\rangle\equiv\gamma_i=\frac{1}{\langle\alpha_i\rangle}+\frac{1}{\langle b_i\rangle}, i=1,2,\ldots,N$ $\langle b_i\rangle=\frac{1+\kappa}{\nu+\frac{1}{2}\left\langle\frac{1}{\alpha_i}\right\rangle}, i=1,2,\ldots,N$

$c=-1$ in (10) as before, the following inverse Gamma prior is obtained for the precision parameters α_i's,

$$p\left(\alpha_i|b_i\right)=\mathcal{IG}(\alpha_i|1,\frac{b_i}{2})=\frac{b_i}{2}\alpha_i^{-2}\exp\left[-\frac{b_i}{2}\frac{1}{\alpha_i}\right] \tag{32}$$

for $i=1,2,\ldots,N$. Working as in Appendix B, it can be easily shown that for such a prior for α_i's, the resulting prior for $\mathbf{w}$ is a multivariate, *multi-parameter* Laplace distribution (each b_i corresponds to a single w_i). Furthermore, the MAP estimator for this model is identical to the vector that minimizes the so-called adaptive (or weighted) lasso cost function [36]–[38]. A summary of the above sparse variational Bayes scheme, which is based on a multiparameter Laplace prior (SVB-mpL) is also shown in Table I.

By inspecting Table I we see that SVB-S, SVB-L and SVB-mpL share common rules concerning the computation of the "low in the hierarchy" model parameters $\mathbf{w},\beta$, while they differ in the way the sparsity imposing precision parameters $\boldsymbol{\alpha}$ are computed. To the best of our knowledge, it is the first time that these three schemes are derived via a mean-field fully factorized variational Bayes inference approach, under a unified framework. Such a presentation not only highlights their common features and differences, but it also facilitates a unified derivation of the corresponding adaptive algorithms that will be described in the next Section.

V. SPARSE VARIATIONAL BAYES ADAPTIVE ESTIMATION

The variational schemes presented in Table I deal with the batch estimation problem associated with (4), that is, given the $n\times N$ data matrix $\mathbf{X}$ and the $n\times 1$ vector of observations $\mathbf{y}$, they provide a sparse estimate ($\hat{\mathbf{w}}\equiv\boldsymbol{\mu}$) of $\mathbf{w}$ after a few iterations. However, in an adaptive estimation setting, solving the size-increasing (by n) batch problem in each time iteration is computationally prohibitive. Therefore, SVB-S, SVB-L and SVB-mpL should be properly modified and adjusted in order to perform adaptive processing in a computationally efficient manner, giving rise to ASVB-S, ASVB-L and ASVB-mpL respectively. In this regard, the time index n is reestablished here and the expectation operator $\langle\cdot\rangle$ is removed from the respective parameters, keeping in mind that henceforth these will refer to *posterior* distribution parameters. By carefully inspecting (16), (17), (19), and (25) (which are common for all three schemes) we reveal the following time-dependent quantities that are commonly met in LS estimation tasks,

$$\mathbf{R}(n)=\mathbf{X}^T(n)\mathbf{\Lambda}(n)\mathbf{X}(n)+\mathbf{A}(n-1), \tag{33}$$

$$\mathbf{z}(n)=\mathbf{X}^T(n)\mathbf{\Lambda}(n)\mathbf{y}(n), \tag{34}$$

$$d(n)=\mathbf{y}^T(n)\mathbf{\Lambda}(n)\mathbf{y}(n). \tag{35}$$

Note that in (33) a time-delayed regularization term $\mathbf{A}(n-1)$ is considered. This is related to the update ordering of the various

algorithmic quantities and does affect the derivation and performance of the new algorithms. From the definitions of $\mathbf{y}(n)$ and $\mathbf{X}(n)$ in (2) and (3) and that of $\mathbf{\Lambda}(n)$, it is easily shown that $\mathbf{R}(n), \mathbf{z}(n)$ and $d(n)$ can be efficiently time-updated as follows:

$$\mathbf{R}(n) = \lambda \mathbf{R}(n-1) + \mathbf{x}(n)\mathbf{x}^T(n) - \lambda \mathbf{A}(n-2) + \mathbf{A}(n-1) \tag{36}$$

$$\mathbf{z}(n) = \lambda \mathbf{z}(n-1) + \mathbf{x}(n)y(n), \tag{37}$$

$$d(n) = \lambda d(n-1) + y^2(n). \tag{38}$$

It is readily recognized that $\mathbf{R}(n)$ is the exponentially weighted sample autocorrelation matrix of $\mathbf{x}(n)$ regularized by the diagonal matrix $\mathbf{A}(n-1), \mathbf{z}(n)$ is the exponentially weighted cross-correlation vector between $\mathbf{x}(n)$ and $y(n)$, and $d(n)$ is the exponentially weighted energy of the observation vector $\mathbf{y}(n)$. By substituting (16) in (17) (with the time index n now included) and using (33) and (34), it is straightforward to show that the adaptive weights $\hat{w}_i(n)(\equiv \mu_i(n))$ can be efficiently computed in time for $i = 1, 2, \ldots, N$, as follows

$$\hat{w}_i(n) = \frac{1}{r_{ii}(n)} \left(z_i(n) - \mathbf{r}_{\neg i}^T(n) \hat{\mathbf{w}}_{\neg i}(n) \right). \tag{39}$$

In the last equation, $z_i(n) = \mathbf{x}_i^T(n)\mathbf{\Lambda}(n)\mathbf{y}(n)$ is the i-th element of $\mathbf{z}(n), r_{ii}(n) = \mathbf{x}_i^T(n)\mathbf{\Lambda}(n)\mathbf{x}_i(n) + \alpha_i(n-1)$ is the i-th diagonal element of $\mathbf{R}(n), \mathbf{r}_{\neg i}^T(n) = \mathbf{x}_i^T(n)\mathbf{\Lambda}(n)\mathbf{X}_{\neg i}(n)$ is the i-th row of $\mathbf{R}(n)$ after removing its i-th element $r_{ii}(n)$, and

$$\hat{\mathbf{w}}_{\neg i}(n) = [\hat{w}_1(n), \ldots, \hat{w}_{i-1}(n), \hat{w}_{i+1}(n-1), \ldots, \hat{w}_N(n-1)]^T. \tag{40}$$

From (39) and (40) it is easily noticed that each weight estimate $\hat{w}_i(n)$ depends on the most recent estimates in time of the other $N-1$ weights. This is in full agreement with the spirit of the variational Bayes approach and the batch SVB schemes presented in the previous Section, where each model parameter is computed based on the most recent values of the remaining parameters. As far as the noise precision parameter $\beta(n)$ is concerned, despite its relatively complex expression given in (19), it is shown in Appendix C that it can be approximated in $\mathcal{O}(N)$ operations per time iteration as follows

$$\beta(n) = \frac{(1-\lambda)^{-1} + N + 2\rho}{2\delta + d(n) - \mathbf{z}^T(n)\hat{\mathbf{w}}(n-1) + \mathbf{r}^T(n)\boldsymbol{\sigma}(n-1)}. \tag{41}$$

In (41), the term $(1-\lambda)^{-1}$ represents the active time window size in an exponentially weighted LS setting, $\mathbf{r}(n) = \mathrm{diag}(\mathbf{R}(n))$ and $\boldsymbol{\sigma}(n-1) = \left[\sigma_1^2(n-1), \sigma_2^2(n-1), \ldots, \sigma_N^2(n-1)\right]^T$ is the vector of posterior weight variances at time $n-1$ with

$$\sigma_i^2(n-1) = \frac{1}{\beta(n-1) r_{ii}(n-1)}, \tag{42}$$

according to (16). Note that (39) and (41) are common in all adaptive schemes described in this paper. What differentiates the algorithms is the way their sparsity enforcing precision parameters $\alpha_i(n)$ are computed in time. More specifically, from (25), (16) and the fact that $\langle w_i^2 \rangle = \hat{w}_i^2 + \sigma_i^2$, we get for ASVB-S,

$$\alpha_i(n) = \frac{2c+1}{a + \beta(n)\hat{w}_i^2(n) + r_{ii}^{-1}(n)}. \tag{43}$$

TABLE II
THE PROPOSED ASVB-S, ASVB-L, AND ASVB-MPL ALGORITHMS

Initialize $\lambda, \hat{\mathbf{w}}(0), \mathbf{A}(-1), \mathbf{A}(0), \mathbf{R}(0), \mathbf{z}(0), d(0), \boldsymbol{\sigma}(0)$
Set $c, a, \rho, \delta, \kappa, \nu$ to very small values (of the order of 10^{-6})
for $n = 1, 2, \ldots$
$\mathbf{R}(n) = \lambda \mathbf{R}(n-1) + \mathbf{x}(n)\mathbf{x}^T(n) - \lambda \mathbf{A}(n-2) + \mathbf{A}(n-1)$
$\mathbf{z}(n) = \lambda \mathbf{z}(n-1) + \mathbf{x}(n)y(n)$
$d(n) = \lambda d(n-1) + y^2(n)$
$\beta(n) = \frac{N + (1-\lambda)^{-1} + 2\rho}{2\delta + d(n) - \mathbf{z}^T(n)\hat{\mathbf{w}}(n-1) + \mathbf{r}^T(n)\boldsymbol{\sigma}(n-1)}$
for $i = 1, 2, \ldots, N$
$\sigma_i^2(n) = 1/(\beta(n) r_{ii}(n))$
$\hat{w}_i(n) = r_{ii}^{-1}(n) \left(z_i(n) - \mathbf{r}_{\neg i}^T(n)\hat{\mathbf{w}}_{\neg i}(n) \right)$
ASVB-S
$\alpha_i(n) = (2c+1) \,/\, \left(a + \beta(n)\hat{w}_i^2(n) + r_{ii}^{-1}(n) \right)$
ASVB-L
$\alpha_i(n) = \sqrt{b(n-1)/(\beta(n)\hat{w}_i^2(n) + r_{ii}^{-1}(n))}$
$\gamma_i(n) = 1/\alpha_i(n) + 1/b(n-1)$
ASVB-mpL
$\alpha_i(n) = \sqrt{b_i(n-1)/(\beta(n)\hat{w}_i^2(n) + r_{ii}^{-1}(n))}$
$\gamma_i(n) = 1/\alpha_i(n) + 1/b_i(n-1)$
$b_i(n) = (1+\kappa)/\left(\nu + \gamma_i(n)/2\right)$
end for
ASVB-L
$b(n) = (N+\kappa)/\left(\nu + \frac{1}{2}\sum_{i=1}^N \gamma_i(n) \right)$
end for

Concerning ASVB-L, from Table I we obtain the following time update recursions,

$$\alpha_i(n) = \sqrt{\frac{b(n-1)}{\beta(n)\hat{w}_i^2(n) + r_{ii}^{-1}(n)}}, \tag{44}$$

$$\gamma_i(n) = \frac{1}{\alpha_i(n)} + \frac{1}{b(n-1)}, \tag{45}$$

$$b(n) = \frac{N+\kappa}{\nu + \frac{1}{2}\sum_{i=1}^N \gamma_i(n)}. \tag{46}$$

Finally, for ASVB-mpL we get expressions similar to (44) and (45) with $b(n-1)$ being replaced by $b_i(n-1)$, while $b_i(n)$ is now expressed as

$$b_i(n) = \frac{1+\kappa}{\nu + \frac{1}{2}\gamma_i(n)} \tag{47}$$

The main steps of the proposed adaptive sparse variational Bayes algorithms are given in Table II. Here again, the hyperparameters $a, c, \rho, \delta, \kappa$ and ν are set equal to very small values (of the order of 10^{-6}) as explained in the previous section. All three algorithms have robust performance, which could be attributed to the absence of matrix inversions or other numerically sensitive computation steps. The algorithms are based on second-order statistics and have an $\mathcal{O}(N^2)$ complexity, similar to that of the classical RLS and other recently proposed sparse adaptive schemes, [7], [9]. This is shown in Table III, where complexity is expressed in terms of the number of multiplications per time iteration. The most computationally costly steps of the proposed algorithms, which require $\mathcal{O}(N^2)$ operations, are those related to the updates of $\mathbf{R}(n)$ and $\hat{\mathbf{w}}(n)$. Note, though, that in an adaptive *filtering* setting, this complexity can be dramatically reduced (and become practically $\mathcal{O}(N)$) by taking ad-

TABLE III
COMPUTATIONAL COMPLEXITY OF SPARSE ADAPTIVE ESTIMATION ALGORITHMS. ξ DENOTES THE SUPPORT OF $\mathbf{w}(n)$

Adaptive algorithm	Complexity
RLS	$2N^2 + \mathcal{O}(N)$
SPARLS [9]	$N^2 + \mathcal{O}(\xi N)$
TWL [7]	$2N^2 + \mathcal{O}(N)$
TNWL [7]	$4N^2 + \mathcal{O}(N)$
ASVB-(S,L,mpL)	$2N^2 + \mathcal{O}(N)$

vantage of the underlying shift invariance property of the data vector $\mathbf{x}(n)$ [7]. As shown in the simulations of Section VII, the algorithms converge very fast to sparse estimates for $\mathbf{w}(n)$ and in the case of ASVB-S and ASVB-mpL, offer lower steady-state estimation error compared to other competing deterministic sparse adaptive schemes. Additionally, while the latter require knowledge of the noise variance beforehand[7], this variance is naturally estimated in time as $1/\beta(n)$ during the execution of the new algorithms.

Most recently reported deterministic sparse adaptive estimation algorithms are sequential variants of the lasso estimator, performing variable selection via soft-thresholding, e.g., the algorithms developed in [7]. To achieve their best possible performances though, such approaches necessitate the use of suitably selected regularization parameters, whose values, in most cases, are determined via time-demanding cross-validation and fine-tuning. Moreover, this procedure should be repeated depending on the application and the application conditions. Unlike the approach followed in deterministic schemes, a completely different sparsity inducing mechanism is used in the proposed algorithms. More specifically, as the algorithms progress in time, many of the exponentially distributed precision parameters $\alpha_i(n-1)$'s are automatically driven to very large values, forcing also the corresponding diagonal elements $r_{ii}(n)$ of $\mathbf{R}(n)$ to become excessively large (33). As a result, according to (39), many weight parameters are forced to become almost zero, thus imposing sparsity. Notably, this sparsity inducing mechanism alleviates the need for fine-tuning or cross-validating of any parameters, which makes the proposed schemes fully automated, and thus, particularly attractive from a practical point of view.

VI. DISCUSSION ON THE PROPOSED ALGORITHMS

Let us now concentrate on the weight updating mechanism given in (39), which is common in all proposed schemes, and attempt to get some further insight on this. To this end, we define the following regularized LS cost function,

$$\mathcal{J}_{\mathrm{LS-R}}(n) = \|\mathbf{\Lambda}^{1/2}(n)\mathbf{y}(n) - \mathbf{\Lambda}^{1/2}(n)\mathbf{X}(n)\hat{\mathbf{w}}(n)\|^2 + \hat{\mathbf{w}}^T(n)\mathbf{A}(n-1)\hat{\mathbf{w}}(n), \quad (48)$$

where the diagonal matrix $\mathbf{A}(n-1)$ has positive diagonal entries and is assumed known, (i.e., for the moment we ignore the procedure that produces $\mathbf{A}(n-1)$). As it is well-known, the vector $\hat{\mathbf{w}}(n)$ that minimizes $\mathcal{J}_{\mathrm{LS-R}}(n)$ is the solution of the following system of equations,

$$\mathbf{R}(n)\hat{\mathbf{w}}(n) = \mathbf{z}(n) \quad (49)$$

[7]With the exception of the algorithms reported in [10], where the noise parameters are adaptively estimated using a smoothing/EM procedure.

where $\mathbf{R}(n)$ and $\mathbf{z}(n)$ are given in (33) and (34), respectively. Let us now decompose $\mathbf{R}(n)$ as,

$$\mathbf{R}(n) = \mathbf{L}(n) + \mathbf{D}(n) + \mathbf{U}(n), \quad (50)$$

where $\mathbf{L}(n)$ is the strictly lower triangular component of $\mathbf{R}(n)$, $\mathbf{D}(n)$ its diagonal component and $\mathbf{U}(n)$ its strictly upper triangular component. This matrix decomposition is the basis of the Gauss-Seidel method [39], and, if substituted in (49), leads to the following iterative scheme for obtaining the optimum $\hat{\mathbf{w}}(n)$,

$$(\mathbf{D}(n) + \mathbf{L}(n))\hat{\mathbf{w}}^{(k)}(n) = \mathbf{z}(n) - \mathbf{U}(n)\hat{\mathbf{w}}^{(k-1)}(n), \quad (51)$$

where k is the iterations index for a given time index n. From the last equation, it is easily verified that by using forward substitution, the elements of $\mathbf{w}^{(k)}(n)$ can be computed sequentially as follows for $i = 1, 2, \ldots, N$,

$$\hat{w}_i^{(k)}(n) = \frac{1}{r_{ii}(n)} \times \left(z_i(n) - \sum_{j<i} r_{ij}(n)\hat{w}_j^{(k)}(n) - \sum_{j>i} r_{ij}(n)\hat{w}_j^{(k-1)}(n) \right). \quad (52)$$

Since the regularized autocorrelation matrix $\mathbf{R}(n)$ is symmetric and positive definite, the Gauss-Seidel scheme in (51) converges (for n fixed) after a few iterations to the solution of (49), irrespective of the initial choice for $\hat{\mathbf{w}}^{(0)}(n)$ [39]. Therefore, in an adaptive estimation setting, optimization is achieved by executing a sufficiently high number of Gauss-Seidel iterations in each time step n. An alternative, more computationally efficient approach though, is to match the iteration and time indices, k and n in (52); i.e., to consider that a single time iteration n of the adaptive algorithm entails just a *single* iteration of the Gauss-Seidel procedure over each coordinate of $\hat{\mathbf{w}}(n)$. By doing so, we end up with the weight updating formula given previously in (39). Such a Gauss-Seidel adaptive algorithm has been previously reported in [40], [41] for the conventional LS cost function $\mathcal{J}_{\mathrm{LS}}(n)$ given in (4), without considering any regularization and/or sparsity issues. It has been termed as the *Euclidean direction set* (EDS) algorithm. Relevant convergence results have been also presented in [42]. However, in that analysis the *time-invariant* limiting values of the autocorrelation and cross-correlation quantities have been employed and thus, the obtained convergence results are not valid for the adaptive Gauss-Seidel algorithm described in [40], [41].

Apart from the Gauss-Seidel viewpoint presented above, a different equivalent approach to arrive at the same weight updating formula as in (39) is the following. We start with the cost function in (48) and minimize it w.r.t. a single weight component in a cyclic fashion. This leads to a cyclic coordinate descent (CCD) algorithm [43] for minimizing $\mathcal{J}_{\mathrm{LS-R}}(n)$ for n fixed. If we now execute only one cycle of the CCD algorithm per time iteration n, we obtain an adaptive algorithm whose weight updating formula is expressed as in (39). CCD algorithms for *sparse adaptive* estimation have been recently proposed in [7]. These algorithms, however, are based on the minimization of $\mathcal{J}_{\mathrm{LS}-\ell_1}(n)$ given in (5), which explicitly incorporates an ℓ_1 penalizing term. In [7] the proposed algorithms have been supported theoretically by relevant convergence results. To the best

of our knowledge, [7] is the only contribution where a proof of convergence of CCD adaptive algorithms has been presented and documented.

From the previous analysis, we conclude that the proposed fully factorized variational methodology described in this paper leads to adaptive estimation schemes where, a) the model weights are adapted in time by using a Gauss-Seidel or CCD type updating rule and b) explicit mechanisms (different for each algorithm) are embedded for computing in time the regularization matrix $\mathbf{A}(n)$ that imposes sparsity to the adaptive weights. The algorithms are fully automated, alleviating the need for predetermining and/or fine-tuning of any penalizing or other regularization parameters.

The convergence properties of the proposed algorithmic family is undoubtedly of major importance. Such results have already been derived for some deterministic sparse adaptive algorithms. More specifically, by assuming that the input sequence is persistently exciting, analytical results for the convergence and the steady-state mean squared error (MSE) of the SPARLS algorithm have been presented in [9]. A result on convergence in the mean is also given in [10]. In a different spirit, in [7] the following ergodicity assumptions are made as a prerequisite for proving convergence,

$$\lim_{n\to\infty} \text{Prob}\left[\frac{1}{n}\mathbf{R}(n) = \mathbf{R}_\infty\right] = 1 \text{ and } \mathbf{R}_\infty \text{ positive definite} \tag{53}$$

$$\lim_{n\to\infty} \text{Prob}\left[\frac{1}{n}\mathbf{z}(n) = \mathbf{z}_\infty\right] = 1, \tag{54}$$

where $\mathbf{R}(n) = \mathbf{X}^T(n)\mathbf{X}(n)$ and $\mathbf{z}(n) = \mathbf{X}^T(n)\mathbf{y}(n)$ in [7]. If these assumptions hold in our case (with $\mathbf{R}(n)$ defined as in (33)) then the convergence analysis presented in [7] would be also valid for the adaptive algorithms described in this paper, with only slight modifications. For this to happen, matrix $\mathbf{A}(n)$ should be either constant, or dependent solely on the data. This is, however, not true owing to the nonlinear dependence of $a_i(n)$'s on the corresponding weight components as shown in (43) and (44). Such a nonlinear interrelation among the parameters of the adaptive algorithms renders the analysis of their convergence an extremely difficult task. In any case, relevant efforts have been undertaken and the problem is under current investigation.

VII. Experimental Results

In this section we present experimental results obtained from applying the proposed variational algorithms to the estimation of a time-varying sparse wireless channel. To assess the estimation performance of the proposed adaptive sparse variational Bayesian algorithms[8], a comparison against a number of state-of-the-art deterministic adaptive algorithms is made, such as the sparsity agnostic RLS, [1], the sparse RLS (SPARLS), [9], the time weighted lasso (TWL), [7], and the time and norm weighted lasso (TNWL), [7]. Moreover, an RLS that operates only on the *a priori* known support set of the channel coefficients, termed as the genie aided RLS (GARLS), is also included in the experiments, in order to serve as a benchmark. To set a fair comparison from a performance point of view, the optimal parameters of the deterministic algorithms are obtained via exhaustive cross-validation in order to acquire the best of their performances.

[8]A Matlab implementation of the variational framework presented in this paper is publicly available at http://members.noa.gr/themelis/lib/exe/fetch.php?media=code:asvb_demo_code.zip.

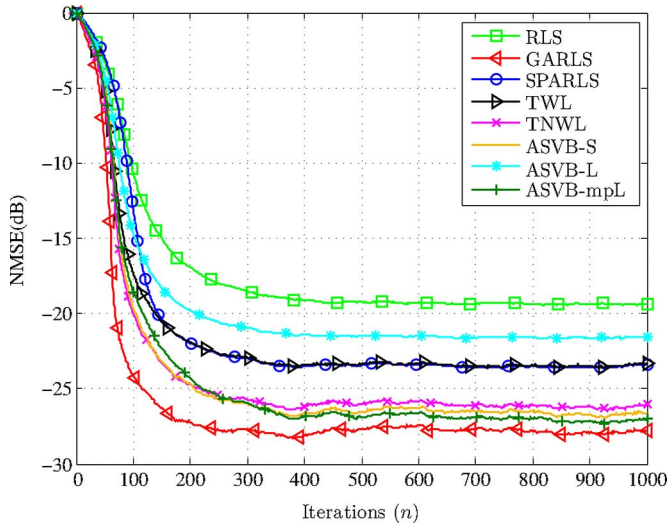

Fig. 2. NMSE curves of adaptive algorithms applied to the estimation of a sparse 64-length time-varying channel with 8 nonzero coefficients. The SNR is set to 15 dB.

We consider a wireless channel with 64 coefficients, which are generated according to Jake's model, [44]. Unless otherwise stated, only 8 of these coefficients are nonzero, having arbitrary positions (support set), and following a Rayleigh distribution with normalized Doppler frequency $f_d\, T_s = 5 \times 10^{-5}$. The forgetting factor is set to $\lambda = 0.99$. The channel's input is a random sequence of binary phase-shift keying (BPSK) ± 1 symbols. The symbols are organized in packets of length 1000 per transmission. Gaussian noise is added to the channel, whose variance is adjusted according to the SNR level of each experiment. The estimation performance of the algorithms is measured in terms of the normalized mean square error (NMSE), which is defined as

$$\text{NMSE} = \frac{\left\langle \|\mathbf{w} - \hat{\mathbf{w}}\|^2 \right\rangle}{\left\langle \|\mathbf{w}\|^2 \right\rangle}, \tag{55}$$

where $\hat{\mathbf{w}}$ is the estimate of the actual channel vector $\mathbf{w}$. All performance curves are ensemble average of 200 transmission packets, channels, and noise realizations.

The first experiment demonstrates the estimation performance of the sparse adaptive estimation algorithms. Fig. 2 shows the NMSE curves of the RLS, GARLS, SPARLS, TWL, TNWL, ASVB-S, ASVB-L, and ASVB-mpL versus time. The SNR is set to 15 dB. Observe that all sparsity aware algorithms perform better than the RLS algorithm, whose channel tap estimates always take non-zero values, even if the actual channel coefficients are zero. Interestingly, there is an improvement margin of about 8 dB in the steady-state NMSE between the RLS and the GARLS, which, as expected, achieves the overall best performance. Moreover, the proposed ASVB-L algorithm has better performance than RLS, but although it promotes sparse estimates, it does not reach the performance level of ASVB-S and ASVB-mpL. From Fig. 2 it is clear that both ASVB-S and ASVB-mpL outperform TNWL, which, in turn, has the best performance among the deterministic algorithms.

TABLE IV
EMPIRICAL RUNTIME FOR THE CONSIDERED ADAPTIVE ALGORITHMS

Adaptive algorithm	Time (s)
RLS	22.99
GARLS	8.38
SPARLS	58.27
TWL	70.85
TNWL	112.61
ASVB-S	66.27
ASVB-L	87.44
ASVB-mpL	87.75

The ASVB-mpL algorithm reaches an error floor that is closer to the one of GARLS, and it provides an NMSE improvement of 1 dB over TNWL and 3 dB over SPARLS and TWL. The empirical runtime of all considered adaptive algorithms for the first experiment is reported in Table IV. Simulations are conducted on an Intel Core i7 machine at 2.20 Ghz and the runtime needed for parameter cross-validation (required by SPARLS, TWL and TNWL) is not included in Table IV.

At this point we grab the chance to shed some light on the relationship between the estimation performance and the complexity of the deterministic algorithms. In a nutshell, the key objective of SPARLS, TWL and TNWL is to optimize the ℓ_1 regularized LS cost function given in (5) w.r.t. $\hat{\mathbf{w}}(n)$ and in a sequential manner. Their estimates, however, are inherently sensitive to the selection of the sparsity imposing parameter τ. The NMSE curves shown in Fig. 2 are obtained after fine-tuning the values of the respective parameters of SPARLS, TWL and TNWL through extensive experimentation. Nonetheless, the thus obtained gain in estimation accuracy adds to the computational complexity of the optimization task. On the other hand, the proposed adaptive variational methods are fully automatic, all parameters are directly inferred from the data, and a single execution suffices to provide the depicted experimental results.

Observe also in Fig. 2 that, as expected, all sparsity aware algorithms converge faster than RLS, requiring an average of approximately 100 fewer iterations in order to reach the NMSE level of -17 dB compared to RLS. Among the deterministic algorithms, TNWL is the one with the fastest convergence rate. In comparison, ASVB-mpL needs almost 10 iterations more than TNWL to converge, but it converges to a lower error floor. Again, the convergence speed of the GARLS is unrivaled.

The next experiment explores the performance of the proposed algorithms for a fast fading channel. The settings of the first experiment are kept the same, with the difference that the normalized Doppler frequency is now increased to $f_d\, T_s = 8.35 \times 10^{-4}$, that suits better to a high mobility application. Specifically, this Doppler results for a system operating at a carrier frequency equal to 1.8 GHz, with a sampling period $T_s = 5 \times 10^{-6}$ and a mobile user velocity 100 Km/h. To account for fast channel variations, the forgetting factor is reduced to $\lambda = 0.96$ (except for ASVB-L, where $\lambda = 0.98$ is used). Fig. 3 shows the resulting NMSE curves for all algorithms versus the number of iterations. In comparison to Fig. 2, we observe that the steady-state NMSE of all algorithms has an expected increase. The algorithms' relative performance is the same, with the exception of ASVB-L, which has higher relative steady-state NMSE and is sensitive to λ. Nevertheless, the

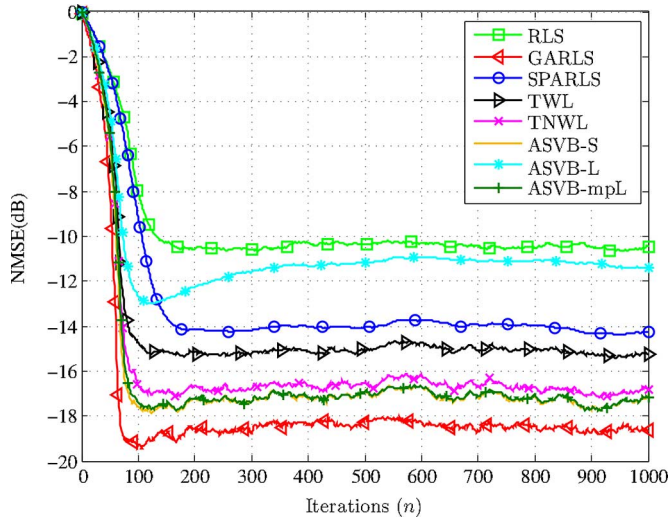

Fig. 3. NMSE curves of adaptive algorithms applied to the estimation of a fast-fading sparse 64-length time-varying channel with 8 nonzero coefficients. The SNR is set to 15 dB.

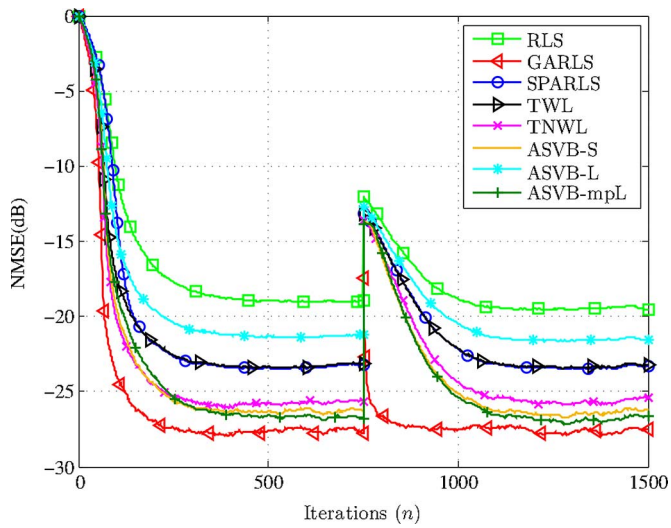

Fig. 4. NMSE curves of adaptive algorithms applied to the estimation of a sparse 64-length time-varying channel with 8 nonzero coefficients, with a non-zero coefficient added at the 750th time mark. The SNR is set to 15 dB.

proposed ASVB-S and ASVB-mpL converge to a better error floor compared to all deterministic algorithms and their NMSE margin to TNWL is more perceptible now.

In the next simulation example, we investigate the tracking performance of the proposed sparse variational algorithms. The experimental settings are identical to those of Fig. 2, with the exception that the packet length is now increased to 1500 symbols, and an extra non-zero Rayleigh fading coefficient is added to the channel at the 750th time instant. Note that until the 750th time mark all algorithms have converged to their steady state. The resulting NMSE curves versus time are depicted in Fig. 4. The abrupt change of the channel causes all algorithms to record a sudden fluctuation in their NMSE curves. Nonetheless, the proposed ASVB-S and ASVB-mpL respond faster than the other algorithms to the sudden change and they successfully track the channel coefficients until they converge to error floors that are again closer to the benchmark GARLS.

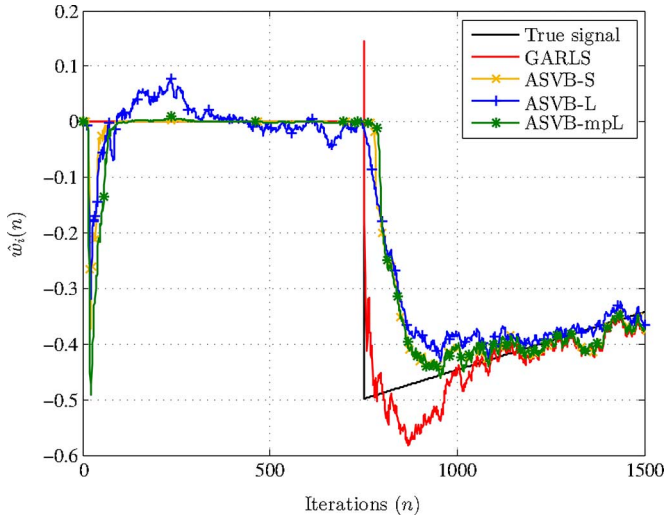

Fig. 5. Tracking of a time-varying channel coefficient.

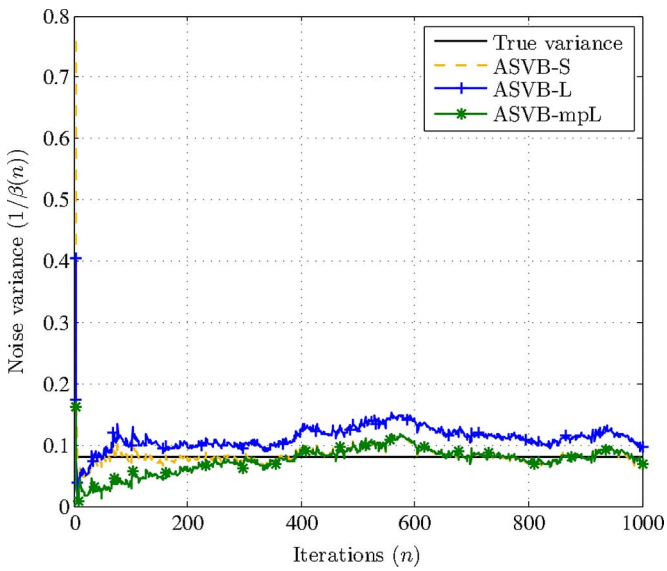

Fig. 6. Estimation of the noise variance in time by the proposed algorithms.

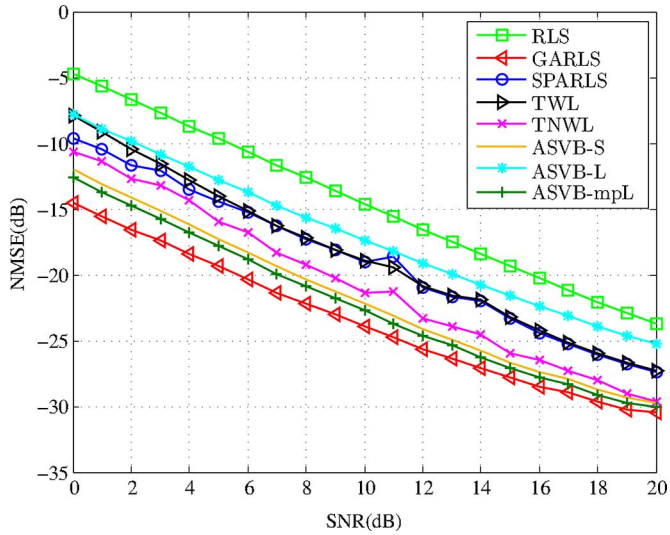

Fig. 7. NMSE versus SNR for all adaptive algorithms applied to the estimation of a sparse 64-length time-varying channel with 8 nonzero coefficients.

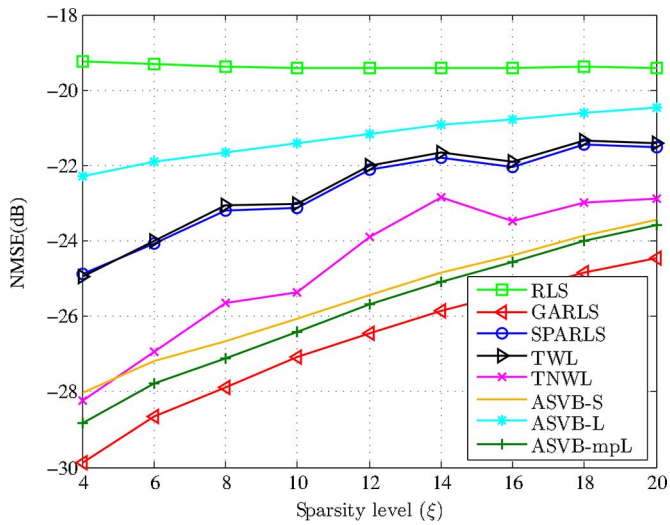

Fig. 8. NMSE versus the level of sparsity of the channel. The SNR is set to 15 dB.

To get a closer look, Fig. 5 depicts the variations in time of the added channel coefficient and the respective estimates of the proposed algorithms. Notice by Fig. 5 that after the first 100 iterations the ASVB-S and ASVB-mpL have converged to a zero estimate for the specific channel coefficient, as opposed to the ASVB-L algorithm, whose estimate is around zero but with higher variations in time. When the value of the true signal suddenly changes, all algorithms track the change after a few iterations. The ASVB-S and ASVB-mpL algorithms converge faster than AVSBL-L to the new signal values. In the sequel, all three algorithms track the slowly fading coefficient, with the estimates of ASVB-S and ASVB-mpL being closer to that of GARLS.

As mentioned previously in Section V, in contrast to all deterministic algorithms the proposed variational algorithmic framework offers the advantage of estimating not only the channel coefficients, but also the noise variance. This is a useful byproduct that can be exploited in many applications, e.g., in the area of wireless communications, where the noise variance estimate can be used when performing minimum mean square error (MMSE) channel estimation and equalization. Fig. 6 depicts the estimation of the noise variance offered by the Bayesian algorithms ASVB-S, ASVB-L and ASVB-mpL across time. Observe that ASVB-S and ASVB-mpL estimate accurately the true noise variance, as opposed to ASVB-L which constantly overestimates it. This is probably the reason why ASVB-L has in general inferior performance compared to ASVB-S and ASVB-mpL. It is worth mentioning that another useful byproduct of the variational framework is the variance of the estimates $\hat{\mathbf{w}}(n)$, given in (16). These variances can be used to build confidence intervals for the weight estimate $\hat{\mathbf{w}}(n)$.

The next experiments evaluate the performance of the proposed algorithms as a function of the SNR and the level of sparsity using the general settings of the first experiment. The corresponding simulation results are summarized in Figs. 7 and 8. It can be seen in Fig. 7 that both ASVB-S and ASVB-mpL outperform all deterministic algorithms for all SNR levels. Specifically, ASVB-mpL achieves an NMSE improvement in all SNR levels of approximately 1 dB over TNWL and 3 dB over SPARLS and TWL, as noted earlier. Moreover, in Fig. 8 the curves affirm the natural increase in the NMSE of the sparsity inducing algorithms as the level of sparsity decreases. The

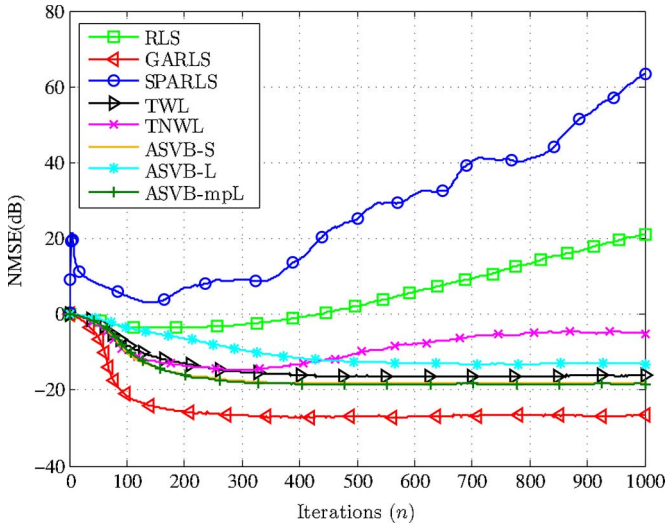

Fig. 9. NMSE curves of adaptive algorithms applied to the estimation of a sparse 64-length time-varying channel with 8 nonzero coefficients. The channel input sequence is colored using a low pass filter. The SNR is set to 15 dB.

simulation results suggest that the performance of the proposed ASVB-S and ASVB-mpL is closest to the optimal performance of GARLS, for all sparsity levels. We should also comment that only the sparsity agnostic RLS algorithm is not affected by the increase of the number of the channel's nonzero components.

As a final experiment, we test the performance of the sparse adaptive algorithms for a colored input signal. To produce a colored input sequence, a Gaussian sequence of zero mean and unit variance is lowpass filtered. For our purposes, a 5th order Butterworth filter is used with a cut-off frequency 1/4 the sampling rate. The remaining settings of our experiment are the same as in the first experiment. Fig. 9 depicts the corresponding NMSE curves for all adaptive algorithms considered in this Section. It is clear from the figure that all algorithms' NMSE performance degrades, owning to the worse conditioning of the autocorrelation matrix $\mathbf{R}(n)$. The convergence speed of all algorithms is also slower than in Fig. 2. Interestingly, RLS and SPARLS diverge. In addition, the poor performance of RLS has a direct impact on TNWL, since, by construction, the inverses of the RLS coefficient estimates are used to weight the ℓ_1-norm in TNWL's cost function. In contrast, both ASVB-S and ASVB-mpL are robust, exhibiting immunity to the coloring of the input sequence.

VIII. Concluding Remarks

In this paper a unifying variational Bayes framework featuring heavy-tailed priors is presented for the estimation of sparse signals and systems. Both batch and adaptive coordinate-descent type estimation algorithms with versatile sparsity promoting capabilities are described with the emphasis placed on the latter, which, to the best of our knowledge, are reported for the first time within a variational Bayesian setting. As opposed to state-of-the-art deterministic techniques, the proposed adaptive schemes are fully automated and, in addition, they naturally provide useful by-products, such as the estimate of the noise variance in time and the variance in the estimate of the parameters, that may provide confidence intervals. Experimental results have shown that the new Bayesian algorithms are robust under various scenarios and in general perform better than their deterministic counterparts in terms of NMSE. Extension of the proposed schemes for complex signals can be made in a straightforward manner. Further developments concerning analytical convergence results and faster versions of the algorithms that update only the non-zero weights (support set) in each time iteration are currently under investigation.

Appendix A
Derivation of the Variational Distribution $q(w_i)$

Starting from (14), the variational distribution $q(w_i)$ is computed as in

$$\begin{aligned}
q(w_i) &\propto \exp\left[\langle \log p(\mathbf{y}|\mathbf{w},\beta) + \log p(\mathbf{w}|\boldsymbol{\alpha},\beta)\rangle\right] \\
&\propto \exp\left[\left\langle -\frac{\beta}{2}\|\boldsymbol{\Lambda}^{1/2}\mathbf{y} - \boldsymbol{\Lambda}^{1/2}\mathbf{X}_{\neg i}\mathbf{w}_{\neg i} \right.\right. \\
&\quad \left.\left. - \boldsymbol{\Lambda}^{1/2}\mathbf{x}_i w_i\|^2 - \frac{\beta}{2}\alpha_i w_i^2 \right\rangle\right] \\
&\propto \exp\left[\left\langle -\frac{\beta}{2}\left(\mathbf{x}_i^T\boldsymbol{\Lambda}\mathbf{x}_i w_i^2 \right.\right.\right. \\
&\quad \left.\left.\left. - 2\mathbf{x}_i^T\boldsymbol{\Lambda}\left(\mathbf{y} - \mathbf{X}_{\neg i}\mathbf{w}_{\neg i}\right) w_i + \alpha_i w_i^2\right) \right\rangle\right] \\
&\propto \exp\left[\left\langle -\frac{1}{2}\left(\beta\left(\mathbf{x}_i^T\boldsymbol{\Lambda}\mathbf{x}_i + \alpha_i\right) w_i^2 \right.\right.\right. \\
&\quad \left.\left.\left. -2\beta\mathbf{x}_i^T\boldsymbol{\Lambda}\left(\mathbf{y} - \mathbf{X}_{\neg i}\mathbf{w}_{\neg i}\right) w_i\right) \right\rangle\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\langle\beta\rangle\left(\mathbf{x}_i^T\boldsymbol{\Lambda}\mathbf{x}_i + \langle\alpha_i\rangle\right) w_i^2 \right.\right. \\
&\quad \left.\left. - 2\langle\beta\rangle\mathbf{x}_i^T\boldsymbol{\Lambda}\left(\mathbf{y} - \mathbf{X}_{\neg i}\langle\mathbf{w}_{\neg i}\rangle\right) w_i\right)\right] \\
&\Rightarrow q(w_i) = (2\pi)^{-1/2}\sigma_i^{-1}\exp\left[-\frac{1}{2}\frac{(w_i - \mu_i)^2}{\sigma_i^2}\right] \qquad (56)
\end{aligned}$$

where μ_i and σ_i^2 are given in (17) and (16) respectively,

Appendix B
Hierarchical Laplace Prior

From (9) and (26) we can write,

$$\begin{aligned}
p(\mathbf{w}|b,\beta) &= \int p(\mathbf{w}|\boldsymbol{\alpha},\beta)p(\boldsymbol{\alpha}|b)d\boldsymbol{\alpha} \\
&= \prod_{i=1}^{N}\int_0^\infty p(w_i|\alpha_i,\beta)p(\alpha_i|b)d\alpha_i \\
&= (2\pi)^{-\frac{N}{2}}\beta^{\frac{N}{2}}\left(\frac{b}{2}\right)^N \\
&\quad \times \prod_{i=1}^{N}\int_0^\infty \alpha_i^{-\frac{3}{2}}\exp\left[-\frac{1}{2}\left(\beta w_i^2\alpha_i + \frac{b}{\alpha_i}\right)\right]d\alpha_i
\end{aligned} \qquad (57)$$

From the definition of the GIG distribution $\mathcal{GIG}(\alpha_i;\beta w_i^2, b, -\frac{1}{2})$ (cf. (10)) the integral in the last equation can be computed, and (57) is then rewritten as

$$p(\mathbf{w}|b,\beta) = (2\pi)^{-\frac{N}{2}}\beta^{\frac{N}{2}}\left(\frac{b}{2}\right)^N 2^N\prod_{i=1}^{N}\frac{\mathcal{K}_{-1/2}(\sqrt{\beta w_i^2 b})}{\left(\frac{\beta w_i^2}{b}\right)^{-\frac{1}{4}}} \qquad (58)$$

In addition,

$$\mathcal{K}_{-1/2}(x) = \mathcal{K}_{1/2}(x) = \sqrt{\frac{\pi}{2}} x^{-\frac{1}{2}} \exp[-x]. \tag{59}$$

Utilizing (59) in (58) and after some straightforward simplifications, we get the multivariate Laplace distribution with parameter $\sqrt{\beta b}$,

$$p(\mathbf{w}|b,\beta) = \left(\frac{\sqrt{\beta b}}{2}\right)^N \exp\left[-\sqrt{\beta b}\|\mathbf{w}\|_1\right], \tag{60}$$

which proves our statement.

APPENDIX C
UPDATE EQUATION FOR $\beta(n)$

By substituting (20) in (19), removing $\langle\cdot\rangle$ and replacing $\boldsymbol{\mu}$ with $\hat{\mathbf{w}}$ yields,

$$\beta = (n+N+2\rho)/\left(2\delta + \|\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{y} - \boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{X}\hat{\mathbf{w}}\|^2 + \sum_{i=1}^{N}\sigma_i^2\mathbf{x}_i^T\boldsymbol{\Lambda}\mathbf{x}_i + \sum_{i=1}^{N}(\hat{w}_i^2+\sigma_i^2)\alpha_i\right). \tag{61}$$

Since exponential data weighting is used, the actual time window size n should be replaced by the effective time window size $(1-\lambda)^{-1} = \sum_{j=0}^{\infty}\lambda^j$ and (61) is rewritten as

$$\beta = \left((1-\lambda)^{-1} + N + 2\rho\right) / \left(2\delta + \mathbf{y}^T\boldsymbol{\Lambda}\mathbf{y} - 2\mathbf{z}^T\hat{\mathbf{w}} + \hat{\mathbf{w}}^T\mathbf{X}^T\boldsymbol{\Lambda}\mathbf{X}\hat{\mathbf{w}} + \hat{\mathbf{w}}^T\mathbf{A}\hat{\mathbf{w}} + \sum_{i=1}^{N}\sigma_i^2\underbrace{(\mathbf{x}_i^T\boldsymbol{\Lambda}\mathbf{x}_i + \alpha_i)}_{r_{ii}}\right) \tag{62}$$

or

$$\beta = \frac{(1-\lambda)^{-1} + N + 2\rho}{2\delta + d - 2\mathbf{z}^T\hat{\mathbf{w}} + \hat{\mathbf{w}}^T\mathbf{R}\hat{\mathbf{w}} + \boldsymbol{\sigma}^T\mathbf{r}} \tag{63}$$

This is an exact expression for estimating the *posterior* noise precision β, which can be used in the proposed algorithms. However, in order to avoid the computation of $\hat{\mathbf{w}}^T\mathbf{R}\hat{\mathbf{w}}$, which entails N^2 operations, we set $\hat{\mathbf{w}} = \mathbf{R}^{-1}\mathbf{z}$ in (63), that is we assume that in each time iteration, $\hat{\mathbf{w}}$ attains its optimum value according to (49)[9]. Then (63) is expressed as,

$$\beta = \frac{(1-\lambda)^{-1} + N + 2\rho}{2\delta + d - \mathbf{z}^T\hat{\mathbf{w}} + \boldsymbol{\sigma}^T\mathbf{r}} \tag{64}$$

Based on the update ordering of the various parameters of the algorithms in time, the respective quantities in (64) are expressed in terms of either $n-1$ or n, leading to (41).

[9]Note that this would be accurate if we let the Gauss-Seidel scheme iterate a few times for each n. On the contrary, as mentioned in Section VI, in the proposed adaptive algorithms a *signle* Gauss-Seidel iteration takes place per time iteration n.

REFERENCES

[1] S. O. Haykin, *Adaptive Filter Theory*, 4th ed. New York, NY, USA: Springer, 2002.
[2] A. H. Sayed, *Adaptive Filters*. New York, NY, USA: Wiley-IEEE Press, 2008.
[3] E. Candes and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, 2008.
[4] D. Donoho, "Compressed sensing," *IIEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.
[5] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Statist. Soc.*, vol. 58, no. 1, pp. 267–288, 1996.
[6] Y. Chen, Y. Gu, and A. Hero, "Sparse LMS for system identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2009, pp. 3125–3128.
[7] D. Angelosante, J. Bazerque, and G. Giannakis, "Online adaptive estimation of sparse signals: Where RLS meets the ℓ_1-norm," *IEEE Trans. Signal Process.*, vol. 58, pp. 3436–3447, July 2010.
[8] E. Eksioglu and A. Tanc, "RLS algorithm with convex regularization," *IEEE Signal Process. Lett.*, vol. 18, no. 8, pp. 470–473, 2011.
[9] B. Babadi, N. Kalouptsidis, and V. Tarokh, "SPARLS: The sparse RLS algorithm," *IEEE Trans. Signal Process.*, vol. 58, pp. 4013–4025, Aug. 2010.
[10] N. Kalouptsidis, G. Mileounis, B. Babadi, and V. Tarokh, "Adaptive algorithms for sparse system identification," *Signal Process.*, vol. 91, no. 8, pp. 1910–1919, 2011.
[11] Y. Kopsinis, K. Slavakis, and S. Theodoridis, "Online sparse system identification and signal reconstruction using projections onto weighted ℓ_1 balls," *IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 936–952, Mar. 2011.
[12] G. Mileounis, B. Babadi, N. Kalouptsidis, and V. Tarokh, "An adaptive greedy algorithm with application to nonlinear communications," *IEEE Trans. Signal Process.*, vol. 58, pp. 2998–3007, June 2010.
[13] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, pp. 183–233, Jan. 1999.
[14] T. S. Jaakkola and M. I. Jordan, "Bayesian parameter estimation via variational methods," *Statistics and Computing*, vol. 10, pp. 25–37, Jan. 2000.
[15] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. New York, NY, USA: Springer-Verlag, 2006.
[16] H. Koeppl, G. Kubin, and G. Paoli, "Bayesian methods for sparse RLS adaptive filters," in *Proc. 37th IEEE Asilomar Conf. Signals, Syst., Comput.*, 2003, vol. 2, pp. 1273–1278, IEEE.
[17] K. E. Themelis, A. A. Rontogiannis, and K. Koutroumbas, "Variational Bayesian sparse adaptive filtering using a Gauss-Seidel recursive approach," presented at the 21st Eur. Signal Process. Conf. (EUSIPCO), Marrakesch, Morrocco, Sep. 2013.
[18] K. E. Themelis, A. A. Rontogiannis, and K. Koutroumbas, "Adaptive variational sparse Bayesian estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2014.
[19] J. Bioucas-Dias, "Bayesian wavelet-based image deconvolution: A GEM algorithm exploiting a class of heavy-tailed priors," *IEEE Trans. Image Process.*, vol. 15, no. 4, pp. 937–951, 2006.
[20] M. Girolami, "A variational method for learning sparse and overcomplete representations," *Neural Comput.*, vol. 13, no. 11, pp. 2517–2532, Nov. 2001.
[21] M. Sato, "Online model selection based on the variational Bayes," *Neural Comput.*, vol. 13, no. 7, pp. 1649–1681, Jul. 2001.
[22] M. A. T. Figueiredo, "Adaptive sparseness for supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1150–1159, Sept. 2003.
[23] T. Park and C. George, "The Bayesian Lasso," *J. Amer. Statist. Assoc.*, vol. 103, no. 482, pp. 681–686, June 2008.
[24] Z. Zhang, S. Wang, D. Liu, and M. I. Jordan, "EP-GIG priors and applications in Bayesian sparse learning," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 2031–2061, June 2012.
[25] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.
[26] S. Ji, X. Y. , and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2346–2356, June 2008.
[27] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. : John Wiley & Sons, 2009.
[28] D. Ba, B. Babadi, P. Purdon, and E. Brown, "Convergence and stability of iteratively re-weighted least squares algorithms," *IEEE Trans. Signal Process.*, vol. 62, no. 1, pp. 183–195, Jan. 2014.
[29] H. Attias, "A variational Bayesian framework for graphical models," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2000, vol. 12, pp. 209–215.

[30] 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, 2008.
[31] M. J. Beal, "Variational Algorithms for Approximate Bayesian Inference," Ph.D. Dissertation, Gatsby Computational Neuroscience Unit, Univ. College London, London, U.K., 2003.
[32] C. Peterson and J. Anderson, "A mean field theory learning algorithm for neural networks," *Complex Syst.*, vol. 1, pp. 995–1019, 1987.
[33] D. Shutin, T. Buchgraber, S. R. Kulkarni, and H. V. Poor, "Fast variational sparse Bayesian learning with automatic relevance determination for superimposed signals," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 6257–6261, 2011.
[34] S. Babacan, R. Molina, and A. Katsaggelos, "Bayesian compressive sensing using Laplace priors," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 53–63, 2010.
[35] J. E. Griffin and P. J. Brown, "Inference with normal-gamma prior distributions in regression problems," *Bayesian Anal.*, vol. 5, no. 1, pp. 171–188, 2010.
[36] H. Zou, "The adaptive lasso and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 101, no. 476, pp. 1418–1429, Dec. 2006.
[37] K. E. Themelis, A. A. Rontogiannis, and K. D. Koutroumbas, "A novel hierarchical Bayesian approach for sparse semisupervised hyperspectral unmixing," *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 585–599, Feb. 2012.
[38] A. A. Rontogiannis, K. E. Themelis, and K. Koutroumbas, "A fast algorithm for the Bayesian adaptive lasso," in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2012, pp. 974–978.
[39] G. Golub and C. Van Loan, *Matrix Computations*. Baltimore, MD, USA: The Johns Hopkins Univ. Press, 1996.
[40] X. G.-F. , T. Bose, W. Kober, and J. Thomas, "A fast adaptive algorithm for image restoration," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 46, no. 1, pp. 216–220, 1999.
[41] T. Bose, *Digital Signal and Image Processing*. New York, NY, USA: Wiley, 2004.
[42] X. G.-F. and T. Bose, "Analysis of the Euclidean direction set adaptive algorithm," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1998, vol. 3, pp. 1689–1692.
[43] D. Bertsekas, *Nonlinear Programming*. Belmont, MA, USA: Athena Scientific, Sep. 1999.
[44] , W. C. Jakes and D. C. Cox, Eds., *Microwave Mobile Communications*. New York, NY, USA: Wiley-IEEE Press, 1994.

Konstantinos E. Themelis was born in Piraeus, Greece, in 1981. He received the diploma degree in computer engineering and informatics from the University of Patras in 2005, and the Ph.D. degree in signal processing from the University of Athens, Greece, in 2012.

Since 2012 he is a postdoctoral research associate at IAASARS, National Observatory of Athens. His research interests are in the area of statistical signal processing and probabilistic machine learning with application to image processing. He is a member of the Technical Chamber of Greece.

Athanasios A. Rontogiannis (M'97) was born in Lefkada Island, Greece, in 1968. He received the Diploma degree (5 years) in electrical engineering from the National Technical University of Athens (NTUA), Greece, in 1991, the M.A.Sc. in electrical and computer engineering from the University of Victoria, Canada, in 1993, and the Ph.D. degree in communications and signal processing from the University of Athens, Greece, in 1997.

From 1998 to 2003, he was with the University of Ioannina. In 2003 he joined the Institute for Astronomy, Astrophysics, Space Applications and Remote Sensing (IAASARS) of the National Observatory of Athens (NOA), where since 2011 he is a Senior Researcher.

Dr. Rontogiannis serves at the Editorial Boards of the EURASIP *Journal on Advances in Signal Processing*, Springer (since 2008) and the EURASIP *Signal Processing Journal*, Elsevier (since 2011). His research interests are in the general areas of signal processing and wireless communications. He is a member of the IEEE Signal Processing and Communication Societies and the Technical Chamber of Greece.

Konstantinos D. Koutroumbas received the Diploma degree from the University of Patras (1989), an M.Sc. Degree in advanced methods in computer science from the Queen Mary College of the University of London (1990) and a Ph.D. degree from the University of Athens (1995).

Since 2001 he is with the Institute for Astronomy, Astrophysics, Space Applications and Remote Sensing of the National Observatory of Athens, Greece, where currently he is a Senior Researcher. His research interests include mainly Pattern Recognition, Time Series Estimation and their application (a) to remote sensing and (b) to the estimation of characteristic quantities of the upper atmosphere. He is a co-author of the books *Pattern Recognition* (1st, 2nd, 3rd, 4th editions) and *Introduction to Pattern Recognition: A MATLAB Approach*. He has over 2500 citations in his work.