Variational Bayes Group Sparse Time-Adaptive Parameter Estimation With Either Known or Unknown Sparsity Pattern

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

***Abstract*—In this paper, we study the problem of time-adaptive group sparse signal estimation from a Bayesian viewpoint. We propose two online variational Bayes schemes that are specifically designed to estimate and track group sparse signals in time. The proposed schemes address both the cases where the grouping information of the signal is either known or not. For the case of known group sparsity pattern, the proposed scheme builds on a novel hierarchical model for the *Bayesian adaptive group lasso*. Utilizing the variational Bayes framework, update equations for all model parameters are given, for both the batch and time adaptive estimation scenarios. To address the case where the group sparsity pattern is unknown, the hierarchical Bayesian model of the former scheme is extended by organizing the penalty parameters of the Bayesian lasso in a *conditional autoregressive model*. Intrinsic conditional autoregression is exploited to penalize the signal coefficients in a structured manner and thus obtain group sparse solutions automatically. Again, a robust and computationally efficient online variational Bayes estimator is developed, capitalizing on the conjugacy of the proposed hierarchical Bayesian formulation. Experimental results are reported that corroborate the superior estimation performance of the proposed online schemes, when compared with state-of-the-art methods.**

***Index Terms*—Group sparsity, online variational Bayes, conditional autoregressive model, generalized inverse Gaussian distribution.**

I. Introduction

ADAPTIVE parameter (or signal) estimation is a research topic of paramount importance in the area of statistical signal processing. It entails time recursive parameter estimation techniques that take advantage of the statistical properties of *sequentially* observed, streaming data, [1]. Adaptive signal estimation has a plethora of applications, including array beamforming, interference and echo cancellation, system identification, channel estimation and equalization in wireless communications, to name but a few, [2].

Manuscript received September 02, 2015; revised January 07, 2016 and March 01, 2016; accepted March 04, 2016. Date of publication March 17, 2016; date of current version April 25, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Wee Peng Tay. This work was partially funded by the PHySIS project, contract no. 640174, within the H2020 Framework Program of the European Commission.

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.2016.2543204

During recent years, advances in the area of compressive sensing have affected almost every aspect of modern signal processing theory, including adaptive signal estimation. The challenge here is to exploit the attribute of *sparsity*, a common characteristic found in many signals in nature, in order to develop more accurate, robust and computationally efficient adaptive estimators. A parameter vector is considered sparse if the majority of its components is zero or almost zero. Hence, to estimate a sparse signal in a time adaptive environment, we need not only to track the values of its nonzero coefficients in time, but also to track its support set that is also subject to changes as time evolves. To address these challenges, the sparsity imposing ℓ_1 norm has been rigorously embedded in widely used adaptive estimators, such as the recursive least squares (RLS) algorithm, e.g., [3], [4]. A Bayesian method to solve this problem has also been described in [5].

More recently, and following developments pertaining to sparse signal representation, the problem of sparse signal estimation has been enhanced by accounting for *group sparsity*. Group sparsity differs from the traditional notion of sparsity, in the sense that the few nonzero coefficients of a group sparse signal form distinct clusters instead of being randomly positioned in the signal support. Group sparse signals are also commonly found in nature, for example, they can be drawn from applications such as image classification, [6], wireless channel equalization, [7], speech recognition, [8], and image denoising and fusion, [9], to name a few. Thus, it is not surprising that, over the last years, a number of signal processing methods have been specifically tailored to handle group sparse signals.

To account first for the *batch* group sparse signal estimation problem, the group lasso has been proposed in [10], serving as an extension of the original lasso [11]. As its name suggests, the group lasso performs variable selection by imposing sparsity on groups of signal coefficients rather than on individual components. In the same manner, the Bayesian counterpart of the group lasso, [12], expands on the hierarchical Bayesian model of the Bayesian lasso, [13], by placing multivariate Laplace priors on separate groups. This formulation is also proposed and described in [14], where the problem of grouped variable selection with a prior hierarchical structure is discussed too. In [15], more generalized sparsity inducing priors are used for representing group sparse signals and the variational Bayes algorithm is used to perform Bayesian inference. However, it is worth pointing out that the aforementioned batch methods assume that

the signal's sparsity pattern is known *a priori*, which, unfortunately, is a setting inherently difficult to encounter in most applications.

To address situations where the grouping information is entirely unknown, more sophisticated methods for batch group sparse signal estimation have been developed lately. In [16], a block sparse Bayesian learning algorithm is presented, where the strategy of overlapping coefficient groups is introduced. In [17], an additional regression layer is added to the Bayesian lasso formulation, with the aim to exploit the correlation between the sparsity inducing parameters. A Bayesian compressive sensing view of the problem is also proposed in [18], where a local Beta process is assumed on latent variables that encodes the signal's sparsity structure. Besides, in the framework of sparse Bayesian learning, covariance associations among adjacent signal coefficients are taken into account in [19].

Although these methods are reliable in recovering block sparse signals, they do not have the ability to process streaming data sequentially. To the best of our knowledge, attempts on *time-adaptive* estimation of group sparse signals are scarce and treat exclusively the case of known sparsity pattern. For example, an RLS type estimator is proposed in [20], where group sparsity is imposed via the $\ell_{1,\infty}$ norm. In the same fashion, [21] utilizes an approximation to the $\ell_{p,0}$ norm, which is again incorporated in the recursive estimation process of the RLS algorithm. However, both these methods stem from a deterministic framework, and, hence, are highly dependent on the selection of appropriate parameter values.

In this paper, we develop two novel online variational Bayes schemes that estimate group sparse time-varying signals and address both the cases where the sparsity pattern is either known or not. In the first case, we assume that the sparsity pattern is known beforehand, and propose a hierarchical formulation of the Bayesian adaptive group lasso, where independent multivariate Laplace distributions are placed over distinct coefficient groups. An advantage of this formulation is that it uses conjugate prior distributions that facilitate the development of an efficient variational Bayes algorithm to perform inference. Hence, an iterative variational scheme is first presented for the batch estimation problem, which is then properly modified for the task of online inference. In the sequel, we attack the case where the group sparsity pattern is entirely unknown by proposing a modification in the last hierarchical level of the Bayesian model of the former scheme. This level concerns the penalty parameters of the proposed Bayesian adaptive group lasso, whose role is to shrink the signal coefficients towards zero. Specifically, we propose to organize these penalty parameters in a *conditional autoregressive model*. In this way, we impose correlation between adjacent signal coefficients, which shrinks the signal towards zero in a structured manner. Again, the Bayesian formulation has a simple interpretation and an online variational Bayes scheme is developed in a manner similar to the first case. The main advantage of the proposed scheme is that group sparse solutions are automatically obtained with essentially no additional computational cost. It should be noted that, to the best of our knowledge, the time-adaptive estimation of group sparse signals with unknown sparsity pattern has not been dealt with before in the open signal processing literature. We validate the performance of the new schemes in a channel estimation setup, using both synthetic and a real wireless channel, under various conditions. Extensive experimental results illustrate that, in terms of the mean squared estimation error, the proposed online variational Bayes schemes exhibit superior performance compared to state-of-the-art structure-ignorant sparse time-adaptive algorithms.

The rest of the paper is organized as follows. Section II provides the mathematical formulation of the time-adaptive estimation problem. In Section III an online variational Bayes scheme is developed, based on prior knowledge of the signal's group sparsity pattern. To account also for the case where such knowledge is not available, Section IV presents an online variational Bayes scheme that automatically detects the positioning of the grouped nonzero signal coefficients. Section V provides detailed experimental results and, finally, conclusions are reported in Section VI.

Notation: Matrices are denoted by bold capital letters, e.g., $\mathbf{X}$, column vectors are written with bold lowercase letters, e.g., $\mathbf{x}$, while $x_i, \mathbf{x}_i$ denote the ith entry and ith column of $\mathbf{x}$ and $\mathbf{X}$, respectively. $\mathbf{I}_M$ is the $M \times M$ identity matrix, $\mathbf{1}_d$ is the all-ones vector of length d, $\mathbf{0}$ is the zero vector, $(\cdot)^T$ denotes transposition, $\|\cdot\|$ stands for the classical ℓ_2 norm, $\mathrm{tr}(\mathbf{X})$ is the trace of the square matrix $\mathbf{X}$, $\mathrm{diag}(\mathbf{x})$ is a diagonal matrix whose diagonal elements are the entries of $\mathbf{x}$, and $\mathrm{diag}(\mathbf{X})$ denotes a diagonal matrix formed by the diagonal elements of the square matrix $\mathbf{X}$. $\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ denotes the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. $\mathcal{GIG}(x;p,a,b)$ is the one-dimensional generalized inverse Gaussian distribution defined as

$$\mathcal{GIG}(x;p,a,b) = \frac{(a/b)^{p/2}\exp\left[(p-1)\log x - \left(ax+\frac{b}{x}\right)/2\right]}{2K_p(\sqrt{ab})},$$

where $x > 0, a > 0, b > 0, p$ is real, and $K_p(\cdot)$ denotes the modified Bessel function of second kind with p degrees of freedom. The pdf of the Gamma distribution is

$$\mathcal{G}(x;\zeta,\tau) = \exp\left[(\zeta-1)\log x - \frac{x}{\tau} - \log\Gamma(\zeta) - \zeta\log\tau\right],$$

where $\Gamma(\cdot)$ is the gamma function, while

$$\mathcal{IG}(x;\zeta,\tau) = \exp\left[-(\zeta+1)\log x - \frac{\tau}{x} - \log\Gamma(\zeta) + \zeta\log\tau\right],$$

is the inverse Gamma distribution.

II. Problem Formulation

Consider a *group sparse* time-varying weight vector $\mathbf{w}(n) = [w_1(n), w_2(n), \ldots, w_N(n)]^T \in \mathbb{R}^N$, where n is the time index. We assume that $\mathbf{w}(n)$ has $\xi \ll N$ non-zero elements that occur in blocks rather than being independently distributed in random positions. Our objective is to estimate the varying vector $\mathbf{w}(n)$ based on a) some known input data stacked on an $n \times N$ matrix $\mathbf{X}(n) = [\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(n)]^T$, and b) some noisy data observations, $\mathbf{y}(n) = [y(1), y(2), \ldots, y(n)]^T$, up to time n. The measured data are assumed to be generated by the linear regression model

$$\mathbf{y}(n) = \mathbf{X}(n)\mathbf{w}(n) + \boldsymbol{\epsilon}(n), \tag{1}$$

where $\boldsymbol{\epsilon}(n)$ stands for additive noise.

After collecting sufficient measurements so that $n \geq N$, $\mathbf{w}(n)$ can be estimated by minimizing the following least squares (LS) cost function,

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

where λ is the well-known forgetting factor, with $0 \ll \lambda < 1$ and $\mathbf{\Lambda}(n) = \mathrm{diag}\left([\lambda^{n-1}, \lambda^{n-2}, \ldots, 1]^T\right)$. The forgetting factor assigns less significance to past data, and $\mathbf{w}(n)$ is estimated based primarily on more recent observations. An efficient way to minimize (2) recursively in time is by using the celebrated RLS algorithm, which, unfortunately, cannot take advantage of the group sparsity of $\mathbf{w}(n)$ to enhance its estimation performance.

In this paper we study the previously described time-adaptive parameter estimation problem from a Bayesian viewpoint and propose two novel online variational Bayes schemes that promote *group sparse* solutions. The first scheme requires that the group sparsity pattern of $\mathbf{w}(n)$ is known *a priori*. Motivated by the group lasso, our Bayesian model considers a sparsity-imposing multivariate Laplace prior for the coefficient blocks that are of known number and size. When the sparsity pattern is unknown (which is a more realistic scenario), we develop a slightly more complex scheme that *automatically* identifies the non-zero groups of $\mathbf{w}(n)$. This is achieved by imposing a conditional autoregressive model on the penalty parameters of the Laplace distribution, which induces correlation among the adjacent coefficients of $\mathbf{w}(n)$. Both schemes are designed to perform variational Bayes inference for the model parameters by processing streaming data in an online setting.

III. Variational Scheme With Known Group Sparsity Pattern

In this section we consider the case where the signal's group sparsity pattern is known beforehand, i.e., we know the exact number of groups formed by the signal coefficients, as well as their sizes. Based on this information, we develop a hierarchical Bayesian model that imposes group sparsity. A variational Bayes algorithm is then developed to perform batch and online inference[1].

A. Hierarchical Bayesian Modeling

Let us start with the description of our group sparsity imposing hierarchical Bayesian model, which can be considered as a block extension of the model presented in [5]. We temporarily drop the time index n from all model parameters in order to simplify notation. Hence, at first, our analysis applies to the batch estimation problem, where a single, fixed-size batch of data is observed. Time indexing is reestablished at Section III.C, where the time-adaptive variational Bayes algorithm is presented.

[1]A version similar to the proposed variational Bayes scheme has been recently presented in [22]. However, in contrast to the current work, in [22] groups of fixed size are considered and a Student-t group sparsity-promoting prior is utilized.

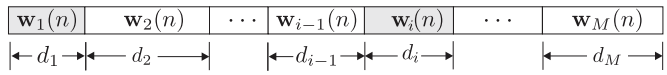


Fig. 1. A group sparse signal $\mathbf{w}(n)$ with floating group size. Non-zero (zero) blocks are shaded (non-shaded).

Considering the linear model in (1), it is typical that our data likelihood is determined by the additive noise distribution. To establish a connection, under maximum likelihood arguments, between the likelihood function and the cost function in (2), we assume that the additive noise is distributed as $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \beta^{-1}\mathbf{\Lambda}^{-1})$. This gives rise to the likelihood function,

$$p(\mathbf{y}|\mathbf{w}, \beta) = \mathcal{N}(\mathbf{X}\mathbf{w}, \beta^{-1}\mathbf{\Lambda}^{-1}). \quad (3)$$

Notice that minimizing (2) is equivalent to maximizing (3) with respect to $\mathbf{w}$. Having defined our likelihood function, we proceed to define appropriate prior distributions for the model parameters $\mathbf{w}$ and β. For the precision parameter β we assume a nonnegative Gamma distribution,

$$p(\beta; \rho, \delta) = \mathcal{G}(\beta; \rho, 1/\delta), \quad (4)$$

which is *conjugate* with respect to the likelihood in (3). The hyperparameters ρ and δ in (4) are set to values close to zero, so as to define a non-informative prior, [23]. Furthermore, we assume that the weight coefficients w_i's are organized in M groups, i.e., $\mathbf{w} = \left[\mathbf{w}_1^T, \mathbf{w}_2^T, \ldots, \mathbf{w}_M^T\right]^T$, where $\mathbf{w}_i$ is the $d_i \times 1$ weight component corresponding to the ith block of $\mathbf{w}$, and $N = \sum_{i=1}^{M} d_i$. Many of these groups are zero, while the remaining ones are nonzero, as shown in Fig. 1. Our objective is to place an independent Laplace distribution over each coefficient block $\mathbf{w}_i$, so as to form the Bayesian analogue of the *adaptive group lasso*, [24]. However, since the Laplace distribution is not conjugate with respect to the Gaussian likelihood in (3), we utilize, instead, its equivalent hierarchical representation, [25]. At the first hierarchical level, each coefficient group $\mathbf{w}_i, i = 1, \ldots, M$, is assigned a zero-mean multivariate Gaussian distribution with a diagonal covariance matrix, i.e.,

$$p(\mathbf{w}|\boldsymbol{\alpha}, \beta) = \prod_{i=1}^{M} \mathcal{N}\left(\mathbf{w}_i|\mathbf{0}, \beta^{-1}\alpha_i^{-1}\mathbf{I}_{d_i}\right). \quad (5)$$

Notice that the noise variance β^{-1} scales the covariance matrix of $\mathbf{w}$ in order for the conditional posterior $p(\mathbf{w}, \beta|\mathbf{y})$ to be unimodal, as explained in [13]. Also, in (5), only a single precision parameter α_i parameterizes the covariance matrix of the ith group $\mathbf{w}_i$. As it will be shown later, during the inference procedure, some of the α_i's obtain high values, which drive the corresponding $\mathbf{w}_i$'s to values very close to zero. These precision parameters $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_M]^T$ are assigned a multivariate inverse Gamma distribution in the second level of hierarchy,

$$p(\boldsymbol{\alpha}|\mathbf{b}) = \prod_{i=1}^{M} \mathcal{IG}\left(\alpha_i \,\middle|\, \frac{d_i + 1}{2}, \frac{b_i}{2}\right), \quad (6)$$

with scale parameters $\mathbf{b} = [b_1, b_2, \ldots, b_M]^T$. In Appendix A it is proven that by utilizing (5) and (6) and by integrating out

$\boldsymbol{\alpha}$, a multivariate Laplace-type prior is established over $\mathbf{w}$, as given in (66). In Appendix A it is also shown that under a maximum a posteriori (MAP) context, the group sparsity promoting multivariate Laplace-type prior in (66) can be considered as the Bayesian group analogue of the adaptive lasso, [26]. Moreover, it can be shown that the scale parameters b_i's in (67) are analogous to the ℓ_1-norm regularizing parameters of the adaptive group lasso, [26], [27]. To directly infer these parameters from the data, we assign a conjugate Gamma prior over $\mathbf{b}$, [13],

$$p(\mathbf{b};\iota,\eta)=\prod_{i=1}^{M}\mathcal{G}(b_i;\iota,1/\eta), \tag{7}$$

where ι and η are shape and rate hyperparameters that are set to values close to zero.

B. Variational Bayesian Inference

Bayesian methods rely on the joint posterior distribution to perform inference on the model parameters. Unfortunately, the complexity of the model proposed in Section III.A does not allow for the exact computation of the posterior $p(\mathbf{w},\beta,\boldsymbol{\alpha},\mathbf{b}|\mathbf{y})$ using Bayes law. In an attempt to overcome this difficulty, Markov chain Monte Carlo (MCMC) sampling can be utilized to approximate our model's posterior distribution. However, this would lead to an estimator unable to operate within the strict time constraints of the time-adaptive estimation scenario under consideration, although sequential MCMC methods have also been recently proposed, [28]. Hence, in this paper we resort to the deterministic framework associated with the variational Bayes algorithm by approximating $p(\mathbf{w},\beta,\boldsymbol{\alpha},\mathbf{b}|\mathbf{y})$ with a simpler distribution, $q(\mathbf{w},\beta,\boldsymbol{\alpha},\mathbf{b})$.

To obtain an analytically tractable approximation, the variational Bayes methodology dictates that the approximating distribution takes on a simple, factorized form. For our purposes, we factorize $q(\mathbf{w},\beta,\boldsymbol{\alpha},\mathbf{b})$ according to the partitioning of the weight vector, i.e.,

$$q(\mathbf{w},\beta,\boldsymbol{\alpha},\mathbf{b})=q(\beta)\prod_{i=1}^{M}q(\mathbf{w}_i)\prod_{i=1}^{M}q(\alpha_i)\prod_{i=1}^{M}q(b_i). \tag{8}$$

The variational Bayes algorithm iteratively minimizes the Kullback-Leibler distance between the true posterior $p(\mathbf{w},\beta,\boldsymbol{\alpha},\mathbf{b}|\mathbf{y})$ and $q(\mathbf{w},\beta,\boldsymbol{\alpha},\mathbf{b})$ in (8), [29], [30]. This operation is equivalent to maximizing a lower bound of the marginal data likelihood with respect to the approximating distribution $q(\mathbf{w},\beta,\boldsymbol{\alpha},\mathbf{b})$, [31]. Let $\Theta=\left\{\mathbf{w}_1^T,\ldots,\mathbf{w}_M^T,\beta,\alpha_1,\ldots,\alpha_M,b_1,\ldots,b_M\right\}$ be the set of all model variables and θ_i its ith element. Then the closed form solution for each posterior factor $q(\theta_i)$ is expressed as, [29], [30],

$$q(\theta_i)=\frac{\exp\left(\mathbb{E}_{j\neq i}\left[\log p(\mathbf{y},\Theta)\right]\right)}{\int\exp\left(\mathbb{E}_{j\neq i}\left[\log p(\mathbf{y},\Theta)\right]\right)d\theta_i}, \tag{9}$$

where $\mathbb{E}_{j\neq i}[\cdot]$ denotes expectation w.r.t. all $q(\theta_j)$'s except for $q(\theta_i)$. Applying (9) for the noise precision variable β, we get a Gamma posterior approximating distribution, [22],

$$q(\beta)=\mathcal{G}(\beta;\tilde{\rho},1/\tilde{\delta}), \tag{10}$$

with

$$\tilde{\rho}=\frac{n+N}{2}+\rho \tag{11}$$

$$\tilde{\delta}=\delta+\frac{1}{2}\left\langle\left\|\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{y}-\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{X}\mathbf{w}\right\|^2\right\rangle+\frac{1}{2}\sum_{i=1}^{M}\langle\alpha_i\rangle\left\langle\|\mathbf{w}_i\|^2\right\rangle, \tag{12}$$

where $\langle\cdot\rangle$ denotes expectation with respect to the corresponding posterior factor $q(\,\cdot\,)$. Next, for each weight block $\mathbf{w}_i, i=1,\ldots,M$, we get

$$q(\mathbf{w}_i)=\mathcal{N}(\mathbf{w}_i;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i), \tag{13}$$

where

$$\boldsymbol{\Sigma}_i=\langle\beta\rangle^{-1}\left(\mathbf{X}_i^T\boldsymbol{\Lambda}\mathbf{X}_i+\langle\alpha_i\rangle\mathbf{I}_{d_i}\right)^{-1}, \tag{14}$$

$$\boldsymbol{\mu}_i=\langle\beta\rangle\boldsymbol{\Sigma}_i\mathbf{X}_i^T\boldsymbol{\Lambda}\left(\mathbf{y}-\mathbf{X}_{\neg i}\boldsymbol{\mu}_{\neg i}\right). \tag{15}$$

Matrix $\mathbf{X}_i\in\mathbb{R}^{n\times d_i}$ is formed by the d_i columns of $\mathbf{X}$ corresponding to the ith group, while matrix $\mathbf{X}_{\neg i}\in\mathbb{R}^{n\times(N-d_i)}$ is formed by the remaining $N-d_i$ columns of $\mathbf{X}$. In addition, $\boldsymbol{\mu}_{\neg i}$ is derived from vector $\boldsymbol{\mu}=\left[\boldsymbol{\mu}_1^T,\boldsymbol{\mu}_2^T,\ldots,\boldsymbol{\mu}_M^T\right]^T$ after excluding $\boldsymbol{\mu}_i$. The approximating distribution for $\alpha_i, i=1,\ldots,M$, is

$$q(\alpha_i)=\mathcal{GIG}\left(\alpha_i;-\frac{1}{2},\langle\beta\rangle\left\langle\|\mathbf{w}_i\|^2\right\rangle,\langle b_i\rangle\right), \tag{16}$$

while for the parameters b_i's we get,

$$q(b_i)=\mathcal{G}\left(b_i;\iota+\frac{d_i+1}{2},\left(\eta+\frac{1}{2}\left\langle\frac{1}{\alpha_i}\right\rangle\right)^{-1}\right). \tag{17}$$

Notice that all approximating distributions in (10), (13), (16) and (17) come in standard exponential forms, thanks to the conjugacy of our Bayesian model. Notice also the interdependency among the parameters of the approximate distributions. The variational Bayes algorithm is essentially a coordinate ascent type algorithm, which updates the parameters of the approximate posteriors in (10), (13), (16) and (17) in a cyclic manner. At each step of the optimization procedure a single parameter gets updated while the remaining are kept fixed to their latest values. The required first and second order moments of the parameters are computed as,

$$\langle\beta\rangle=\tilde{\rho}\tilde{\delta}, \tag{18}$$

$$\left\langle\|\mathbf{w}_i\|^2\right\rangle=\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i+\mathrm{tr}(\boldsymbol{\Sigma}_i), \tag{19}$$

$$\left\langle\left\|\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{y}-\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{X}\mathbf{w}\right\|^2\right\rangle=\left\|\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{y}-\boldsymbol{\Lambda}^{\frac{1}{2}}\sum_{i=1}^{M}\mathbf{X}_i\boldsymbol{\mu}_i\right\|^2+\sum_{i=1}^{M}\mathrm{tr}(\boldsymbol{\Sigma}_i\mathbf{X}_i^T\boldsymbol{\Lambda}\mathbf{X}_i), \tag{20}$$

$$\langle\alpha_i\rangle=\sqrt{\frac{\langle b_i\rangle}{\langle\beta\rangle\left\langle\|\mathbf{w}_i\|^2\right\rangle}}, \tag{21}$$

$$\left\langle\frac{1}{\alpha_i}\right\rangle=\frac{1}{\langle\alpha_i\rangle}+\frac{1}{\langle b_i\rangle}, \tag{22}$$

and

$$\langle b_i\rangle=\frac{\iota+\frac{d_i+1}{2}}{\eta+\frac{1}{2}\left\langle\frac{1}{\alpha_i}\right\rangle}. \tag{23}$$

Using these moments, the proposed *batch* variational Bayes scheme converges to a group sparse estimate, $\boldsymbol{\mu}$, for the unknown signal vector $\mathbf{w}$ in a few iterations. In the next section we properly adjust the proposed group sparse variational Bayes scheme to operate in a time-adaptive setting, where data are sequentially received.

C. Time-Adaptive Group Sparse Variational Bayes

Let us now restore time indexing, remove $\langle\cdot\rangle$, and extend the variational Bayes scheme in a time-adaptive setting, where the weight vector $\mathbf{w}(n)$ is now time-varying. To enable online processing, we define the following quantities, whose size does not change over time as new data become available,

$$\mathbf{R}(n) = \mathbf{X}^T(n)\boldsymbol{\Lambda}(n)\mathbf{X}(n) + \mathbf{A}(n-1), \tag{24}$$

$$\mathbf{z}(n) = \mathbf{X}^T(n)\boldsymbol{\Lambda}(n)\mathbf{y}(n), \tag{25}$$

$$d(n) = \mathbf{y}^T(n)\boldsymbol{\Lambda}(n)\mathbf{y}(n), \tag{26}$$

where $\mathbf{A}(n) = \text{diag}([\boldsymbol{\alpha}_1^T(n), \boldsymbol{\alpha}_2^T(n), \ldots, \boldsymbol{\alpha}_M^T(n)]^T)$ and $\boldsymbol{\alpha}_i(n) = \alpha_i(n)\mathbf{1}_{d_i}, i = 1, 2, \ldots, M$. It is easily observed that (a) $\mathbf{R}(n)$ is the sample auto-correlation matrix of $\mathbf{x}(n)$ regularized by the diagonal matrix $\mathbf{A}(n-1)$, (b) $\mathbf{z}(n)$ is the sample cross-correlation vector between $\mathbf{x}(n)$ and $y(n)$, and (c) $d(n)$ is the energy of the observation vector $\mathbf{y}(n)$. These quantities can be expressed recursively in time, as,

$$\mathbf{R}(n) = \lambda\mathbf{R}(n-1) + \mathbf{x}(n)\mathbf{x}^T(n) - \lambda\mathbf{A}(n-2) + \mathbf{A}(n-1) \tag{27}$$

$$\mathbf{z}(n) = \lambda\mathbf{z}(n-1) + \mathbf{x}(n)y(n), \tag{28}$$

$$d(n) = \lambda d(n-1) + y^2(n). \tag{29}$$

Substituting (14) in (15) and utilizing (24) and (25), we deduce that the adaptive weight estimates $\hat{\mathbf{w}}_i(n)(=\boldsymbol{\mu}_i(n))$ can be efficiently computed as[2]

$$\hat{\mathbf{w}}_i(n) = \mathbf{R}_i^{-1}(n)\left(\mathbf{z}_i(n) - \mathbf{R}_{i\neg i}(n)\hat{\mathbf{w}}_{\neg i}(n)\right), \tag{30}$$

for $i = 1, 2, \ldots, M$. In (30), $\mathbf{z}_i$ is the ith $d_i \times 1$ block of $\mathbf{z}$, $\mathbf{R}_i(n)$ is the ith $d_i \times d_i$ diagonal block of $\mathbf{R}(n)$, $\mathbf{R}_{i\neg i}(n)$ is the $d_i \times (N - d_i)$ matrix resulting from the ith row block of $\mathbf{R}(n)$ after removing its ith group of d_i columns, and $\hat{\mathbf{w}}_{\neg i}(n) = \left[\hat{\mathbf{w}}_1^T(n), \ldots, \hat{\mathbf{w}}_{i-1}^T(n), \hat{\mathbf{w}}_{i+1}^T(n-1), \ldots, \hat{\mathbf{w}}_M^T(n-1)\right]^T$. Moreover, it can be shown that noise precision estimate is efficiently computed as follows, [5], [22],

$$\beta(n) = \left((1-\lambda)^{-1} + N + 2\rho\right) \Big/ \left(2\delta + d(n) - \mathbf{z}^T(n)\hat{\mathbf{w}}(n-1) + \sum_{i=1}^{M} \text{tr}[\boldsymbol{\Sigma}_i(n-1)\mathbf{R}_i(n)]\right), \tag{31}$$

[2]It can be shown that (30) represents a block coordinate descent updating rule, [32].

with $\hat{\mathbf{w}}(n) = \boldsymbol{\mu}(n)$ and $\boldsymbol{\Sigma}_i(n-1) = \beta^{-1}(n-1)\mathbf{R}_i^{-1}(n-1)$ according to (14). Based on (18) and (19), let us also define the time-varying auxiliary quantity $\psi_i(n) = \beta(n)\|\hat{\mathbf{w}}_i(n)\|^2 + \text{tr}(\mathbf{R}_i^{-1}(n))$. The precision parameters $\alpha_i(n), i = 1, \ldots, M$, are then updated as[3]

$$\alpha_i(n) = \sqrt{\frac{b_i(n-1)}{\psi_i(n)}}, \tag{32}$$

while, their first inverse moments, $\left\langle \frac{1}{\alpha_i(n)} \right\rangle \equiv \breve{\alpha}_i(n)$, are computed according to (22) as,

$$\breve{\alpha}_i(n) = \frac{1}{\alpha_i(n)} + \frac{1}{b_i(n-1)}. \tag{33}$$

Finally, the penalty parameters $b_i(n), i = 1, \ldots, M$, are updated as

$$b_i(n) = \frac{2\iota + d_i + 1}{2\eta + \breve{\alpha}_i(n)}. \tag{34}$$

The resulting algorithm, termed *adaptive group sparse variational Bayes based on a multivariate Laplace prior (AGSVB-L)*, is summarized in Table I. The proposed variational scheme converges to a group sparse solution as verified in the experimental results section. To get an insight on the sparsity inducing mechanism of AGSVB-L, as the algorithm executes some $\alpha_i(n)$'s tend to very large values and shrink the corresponding $\hat{\mathbf{w}}_i(n)$'s to zero. More specifically, according to (24) the corresponding diagonal blocks $\mathbf{R}_i(n)$'s of $\mathbf{R}(n)$ will tend to diagonal matrices with very large diagonal entries, which by their inversion in (30) will impose the annihilation of $\hat{\mathbf{w}}_i(n)$'s. The computational complexity of AGSVB-L is dominated by the update operation of $\mathbf{R}(n)$ in (27) and the matrix inversion operation in (30), and, hence, is of the order $\mathcal{O}\left(N^2 + (\max\{d_i\})^3\right)$ per iteration. Nonetheless, in practice it is expected that $\max\{d_i\} \ll N$, that is, the maximum group length is much lower than the signal length. Specifically, it is easily seen that if $\max\{d_i\} < \sqrt{N}$ the computational complexity of AGSVB-L reduces to $\mathcal{O}(N^2)$ per time iteration. Originating from a Bayesian framework, the proposed AGSVB-L algorithm has the advantages of being fully automatic (with no need for parameter fine-tuning) and providing a set of approximating posterior distributions for each model parameter, instead of single point estimates.

IV. Variational Scheme With Unknown Group Sparsity Pattern

We now consider the case where there is no prior information on the group sparsity pattern of the time-varying signal $\mathbf{w}(n)$. This is a more challenging setting, where we need to concurrently recover the signal's clustered support and track its nonzero coefficients in time. In what follows, we utilize the Bayesian model presented in Section III.A and the main idea is to modify its third hierarchical level in order to capture possible

[3]The different time indexing of the parameters involved in each of the (32) and (33) is due to the ordering of the update equations adopted in the proposed algorithm.

TABLE I
THE AGSVB-L ALGORITHM

Initialize $\beta(0), \hat{\mathbf{w}}(0), \mathbf{A}(-1), \mathbf{A}(0), \mathbf{R}(0), \mathbf{z}(0), d(0)$
Set $\rho, \delta, \iota, \eta$ to very small values
for $n = 1, 2, \ldots$
$\mathbf{R}(n) = \lambda \mathbf{R}(n-1) + \mathbf{x}(n)\mathbf{x}^T(n)$
$\quad -\lambda \mathbf{A}(n-2) + \mathbf{A}(n-1)$
$\mathbf{z}(n) = \lambda \mathbf{z}(n-1) + \mathbf{x}(n)y(n)$
$d(n) = \lambda d(n-1) + y^2(n)$
$\beta(n) = ((1-\lambda)^{-1} + N + 2\rho)/(2\delta + d(n)$
$\quad -\mathbf{z}^T(n)\hat{\mathbf{w}}(n-1) + \sum_{i=1}^{M} \mathrm{tr}[\frac{\mathbf{R}_i^{-1}(n-1)}{\beta(n-1)}\mathbf{R}_i(n)])$
for $i = 1, 2, \ldots, M$
$\hat{\mathbf{w}}_i(n) = \mathbf{R}_i^{-1}(n)\,(\mathbf{z}_i(n) - \mathbf{R}_{i\neg i}(n)\hat{\mathbf{w}}_{\neg i}(n))$
$\psi_i(n) = \beta(n)\|\hat{\mathbf{w}}_i(n)\|^2 + \mathrm{tr}(\mathbf{R}_i^{-1}(n))$
$\alpha_i(n) = \sqrt{\frac{b_i(n-1)}{\psi_i(n)}}$
$\breve{\alpha}_i(n) = \frac{1}{\alpha_i(n)} + \frac{1}{b_i(n-1)}$
$b_i(n) = (2\iota + d_i + 1)/(2\eta + \breve{\alpha}_i(n))$
end for
end for

sparsity patterns[4]. As before, we resort to the variational Bayes algorithm to develop an online group sparsity-cognizant inference scheme.

A. Hierarchical Bayesian Model

This section provides a detailed description of the proposed clustered sparsity imposing hierarchical Bayesian model. Again, for notational expediency, we drop the dependency of the model parameters on the time index n. We re-introduce time indexing in Section IV.C, where the corresponding time-adaptive variational Bayesian scheme is presented.

As we have already pointed out, we base the development of this section's hierarchical Bayesian model on the one presented in Section III.A. Specifically, we adopt exactly the same priors reported for the first two levels of hierarchy in Section III.A, as expressed by (4), (5), and (6). Notably, owing to the lack of prior knowledge on the grouping information, we now explicitly assume that $d_i = 1, i = 1, \ldots, N$, i.e., all w_i's are deemed to be independently distributed. Under this assumption, the Laplace prior in (66) imposes no structure and matches the one used in [5]. In the sequel, and to take into account the occurrence of the parameter vector's sparsity in groups of coefficients, we revise and refine the prior on the scale parameters b_i's of the Laplace distribution.

As shown in [34], the scale parameters b_i's can be interpreted as the penalty parameters in an adaptive lasso formulation. Their role is to adaptively shrink the signal coefficients w_i's towards zero, in the sense that nonzero coefficients are assigned relatively lower penalty values than zero coefficients. An interesting idea, recently coined in [17], is that the grouping information can be properly embedded in the prior distributions of b_i's, in order to shrink the original signal $\mathbf{w}$ towards zero in a *structured* manner. However, in [17], an additional group-membership matrix is required that provides information on the grouping structure of the signal coefficients. In this paper, group sparsity is induced by imposing correlation among adjacent signal coefficients. This is achieved by organizing their corresponding scale

[4]A preliminary version of the proposed variational Bayes scheme has been recently presented in [33].

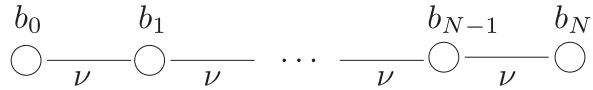


Fig. 2. The proposed Markov chain model.

parameters b_i's in a *conditional autoregressive model*, as described below.

Conditional autoregressive models date back to the pioneering work of [35], and, since then, they have been widely used in data analysis to capture spatial correlations, e.g., [36]. As their name suggests, they are defined via conditional probability distributions over mutually dependent parameters. In our modeling, the "spatial" dependency of the scale parameters is depicted in the undirected graph of Fig. 2. Each node in the graph corresponds to a scale parameter b_i, while the edges between adjacent nodes encode the dependency between them. Next, we assume that the conditional probabilities of the scale parameters b_i's are expressed as,

$$p(b_i|b_{i-1}) = \mathcal{G}(b_i|\kappa, \nu b_{i-1}), \quad i = 1, \ldots, N, \tag{35}$$

where $\kappa > 0$ and $\nu > 0$ are hyperparameters. Using the conditional pdf in (35) and Bayes law, the complete conditional for each scale parameter $b_i, i = 1, \ldots, N-1$, is computed as

$$p(b_i|\mathbf{b}_{\neg i}) = \mathcal{GIG}\left(b_i; 0, \frac{2}{\nu b_{i-1}}, \frac{2b_{i+1}}{\nu}\right), \tag{36}$$

where it is easily observed that each b_i depends only on its direct neighbors b_{i-1} and b_{i+1}. Equation (36) defines a conditional autoregressive model based on the GIG distribution. According to Brook's expansion, [37], the joint probability distribution $p(\mathbf{b})$ can be determined through the set of conditional distributions in (36). In that case, and with respect to the undirected graph in Fig. 2., $\mathbf{b}$ would form a *Markov chain*, [35], [38]. Unfortunately though, the form of the assumed GIG distribution in (36) is too complex to deduce the joint probability from the set of conditional distributions directly. Despite that, this distribution is conjugate with respect to the prior (6) in the second level of our Bayesian model. This allows us to utilize the variational Bayes framework for the development of a computationally efficient inference scheme.

In summary, the proposed hierarchical Bayesian model encompasses the following likelihood and priors

$$p(\mathbf{y}|\mathbf{w}, \beta) = \mathcal{N}(\mathbf{X}\mathbf{w}, \beta^{-1}\mathbf{\Lambda}^{-1}), \tag{37}$$

$$p(\beta; \rho, \delta) = \mathcal{G}(\beta; \rho, \delta), \tag{38}$$

$$p(\mathbf{w}|\boldsymbol{\alpha}, \beta) = \prod_{i=1}^{N} \mathcal{N}\left(w_i|0, \beta^{-1}\alpha_i^{-1}\right), \tag{39}$$

$$p(\boldsymbol{\alpha}|\mathbf{b}) = \prod_{i=1}^{N} \mathcal{IG}\left(\alpha_i|1, b_i/2\right), \tag{40}$$

$$p(b_i|b_{i-1}, b_{i+1}) = \mathcal{GIG}\left(b_i; 0, \frac{2}{\nu b_{i-1}}, \frac{2b_{i+1}}{\nu}\right). \tag{41}$$

B. Variational Bayesian Inference

Working as in Section III.B, we assume that $q(\beta, \mathbf{w}, \boldsymbol{\alpha}, \mathbf{b})$ is now factorized as

$$q(\beta, \mathbf{w}, \boldsymbol{\alpha}, \mathbf{b}) = q(\beta)\prod_{i=1}^{N} q(w_i)\prod_{i=1}^{N} q(\alpha_i)\prod_{i=1}^{N} q(b_i), \quad (42)$$

and use the closed form expression (9) to compute each approximating factor in (42). For the model parameters $\beta, \mathbf{w}$, and $\boldsymbol{\alpha}$ in the two first levels of hierarchy, we get the same distributions as in Section III.B, although in their univariate form, since $d_i = 1, \forall i$. For the sake of completeness, we restate them below, i.e.,

$$q(\beta) = \mathcal{G}(\beta; \tilde{\rho}, 1/\tilde{\delta}), \quad (43)$$

with $\tilde{\rho}$ and $\tilde{\delta}$ given in (11) and (12), after replacing M by N and $\langle \|\mathbf{w}_i\|^2 \rangle$ by $\langle w_i^2 \rangle$,

$$q(w_i) = \mathcal{N}(w_i; \mu_i, \sigma_i^2), \quad i = 1, 2, \ldots, N, \quad (44)$$

with

$$\sigma_i^2 = \langle \beta \rangle^{-1} \left(\mathbf{x}_i^T \boldsymbol{\Lambda} \mathbf{x}_i + \langle \alpha_i \rangle \right)^{-1} \quad (45)$$

$$\mu_i = \langle \beta \rangle \sigma_i^2 \mathbf{x}_i^T \boldsymbol{\Lambda} (\mathbf{y} - \mathbf{X}_{\neg i} \boldsymbol{\mu}_{\neg i}), \quad (46)$$

and

$$q(\alpha_i) = \mathcal{GIG}\left(\alpha_i; -1/2, \langle \beta \rangle \langle w_i^2 \rangle, \langle b_i \rangle\right). \quad (47)$$

Note that, now, in (45), (46), $\mathbf{x}_i$ is the ith column of $\mathbf{X}$ and $\mathbf{X}_{\neg i}$ results from $\mathbf{X}$ after removing $\mathbf{x}_i$. Finally, as shown in the Appendix B, the penalty parameters $b_i, i = 1, 2, \ldots, N-1$, are inferred via the GIG approximating distribution

$$q(b_i) = \mathcal{GIG}\left(b_i; 1, \varrho_i, \omega_i\right), \quad (48)$$

where $\varrho_i = \left\langle \frac{1}{\alpha_i} \right\rangle + \frac{2}{\nu} \left\langle \frac{1}{b_{i-1}} \right\rangle$ and $\omega_i = 2\langle b_{i+1} \rangle / \nu, i = 1, \ldots, N-1$. For the Nth node of the chain the posterior in (48) simplifies to

$$q(b_N) = \mathcal{G}\left(b_N; \kappa + 1, 2/\varrho_N\right). \quad (49)$$

As previously, our variational Bayes scheme iteratively optimizes the parameters of the approximating factors given in (43), (44), (47), (48), and (49). The mean of the noise precision is given in (18), while the additional moments involved in the previous equations are computed as

$$\langle w_i \rangle = (\mathbf{x}_i^T \mathbf{x}_i + \langle \alpha_i \rangle)^{-1} \mathbf{x}_i^T (\mathbf{y} - \mathbf{X}_{\neg i} \langle \mathbf{w}_{\neg i} \rangle), \quad (50)$$

$$\langle \alpha_i \rangle = \sqrt{\frac{\langle b_i \rangle}{\langle \beta \rangle \langle w_i^2 \rangle}}, \quad \left\langle \frac{1}{\alpha_i} \right\rangle \equiv \breve{\alpha}_i = \frac{1}{\langle \alpha_i \rangle} + \frac{1}{\langle b_i \rangle}, \quad (51)$$

$$\langle b_i \rangle = \sqrt{\frac{\omega_i}{\varrho_i}} \frac{K_2(\sqrt{\omega_i \varrho_i})}{K_1(\sqrt{\omega_i \varrho_i})}, \quad \langle b_N \rangle = (\kappa + 1)\frac{2}{\varrho_N}, \quad (52)$$

and,

$$\left\langle \frac{1}{b_i} \right\rangle \equiv \breve{b}_i = \sqrt{\frac{\varrho_i}{\omega_i}} \frac{K_0(\sqrt{\omega_i \varrho_i})}{K_1(\sqrt{\omega_i \varrho_i})}. \quad (53)$$

It should be noted that $\langle \alpha_i \rangle$ and $\left\langle \frac{1}{\alpha_i} \right\rangle$ in (51) are easily obtained from (21) and (22) by setting $d_i = 1$, while the expressions for $\langle b_i \rangle$ and $\left\langle \frac{1}{b_i} \right\rangle$ in (52) and (53) are derived in the Appendix. The resulting *batch* variational Bayes scheme updates, in its core, the expectations $\langle w_i \rangle, \langle \beta \rangle, \langle \alpha_i \rangle$, and $\langle b_i \rangle$, for $i = 1, 2, \ldots, N$, and converges in a few iterations. In the next section, we show how the proposed variational Bayes algorithm can operate online for handling streaming data.

C. Adaptive Group Sparse Variational Bayes

We now reestablish time indexing for all model parameters in order to develop the time-adaptive version of the batch variational algorithm presented in Section IV.B. To achieve this, we utilize again the recursive (27), (28), and (29), and map batch iterations to time iterations.

To begin with the parameter updates, β and $\mathbf{w}$ are time updated in the same fashion as in Section III.C. Setting $d_i = 1$, (30) and (31) become

$$\hat{w}_i(n) = \frac{1}{r_i(n)} \left(z_i(n) - \mathbf{r}_{i\rightarrow i}^T(n) \hat{\mathbf{w}}_{\neg i}(n) \right), \quad (54)$$

and

$$\beta(n) = \frac{(1-\lambda)^{-1} + N + 2\rho}{2\delta + d(n) - \mathbf{z}^T(n)\hat{\mathbf{w}}(n-1) + \mathbf{r}^T(n)\boldsymbol{\sigma}(n-1)}. \quad (55)$$

According to (51), α_i's and $\breve{\alpha}_i$'s are time updated as

$$\alpha_i(n) = \sqrt{\frac{b(n-1)}{\beta(n)\hat{w}_i^2(n) + r_i^{-1}(n)}}, \quad (56)$$

and

$$\breve{\alpha}_i(n) = \frac{1}{\alpha_i(n)} + \frac{1}{b_i(n-1)}. \quad (57)$$

Also, let $\varrho_i(n) = \breve{\alpha}_i(n) + 2\breve{b}_{i-1}(n-1)/\nu$ and $\omega_i(n) = 2b_{i+1}(n-1)/\nu$. Then, the time updates of the scale parameters $b_i, i = 1, 2, \ldots, N-1$, are expressed as,

$$b_i(n) = \sqrt{\frac{\omega_i(n)}{\varrho_i(n)}} \frac{K_2(\sqrt{\omega_i(n)\varrho_i(n)})}{K_1(\sqrt{\omega_i(n)\varrho_i(n)})}, \quad (58)$$

while for the last chain node we have that $b_N(n) = 2(\kappa + 1)/\varrho_N(n)$. Finally, according to (53), $\breve{b}_i$'s are updated as

$$\breve{b}_i(n) = \sqrt{\frac{\varrho_i(n)}{\omega_i(n)}} \frac{K_0(\sqrt{\omega_i(n)\varrho_i(n)})}{K_1(\sqrt{\omega_i(n)\varrho_i(n)})}, \; i = 1, \ldots, N1. \quad (59)$$

The proposed algorithm, termed *adaptive group sparse variational Bayes using conditional auto-regression (AGSVB-CAR)*, is summarized in Table II. AGSVB-CAR converges in a few iterations and successfully detects any sparsity pattern, as also verified experimentally in the next section. The algorithm automatically infers all model parameters, after setting the hyperparameters ρ, δ, κ, and ν to fixed values. Note that ρ and δ are set to values close to zero in order to get non-informative priors, while an extra maximization step could be employed over the hyperparameters κ and ν. However, due to the complexity of the maximization step and the fact that κ and ν have less effect on inference, since they are deep in the Bayesian model hierarchy, [39], we have adopted fixed values for κ and ν. This choice is also supported by experimental results which show that the AGSVB-CAR algorithm is quite robust to the selection of these parameters (it is sensitive only to their order of magnitude) and,

TABLE II
THE PROPOSED AGSVB-CAR ALGORITHM

Initialize $\lambda, \hat{\mathbf{w}}(0), \mathbf{A}(-1), \mathbf{A}(0), \mathbf{R}(0), \mathbf{z}(0), d(0), \boldsymbol{\sigma}(0), \mathbf{b}(0)$
Set $\rho = \delta = 10^{-6}$, $\kappa = 10^{-3}, \nu = 10^3$
for $n = 1, 2, \ldots$
$\mathbf{R}(n) = \lambda \mathbf{R}(n-1) + \mathbf{x}(n)\mathbf{x}^T(n)$
$\quad -\lambda \mathbf{A}(n-2) + \mathbf{A}(n-1)$
$\mathbf{z}(n) = \lambda \mathbf{z}(n-1) + \mathbf{x}(n)y(n)$
$d(n) = \lambda d(n-1) + y^2(n)$
$\beta(n) = \frac{N + (1-\lambda)^{-1} + 2\rho}{2\delta + d(n) - \mathbf{z}^T(n)\hat{\mathbf{w}}(n-1) + \mathbf{r}^T(n)\boldsymbol{\sigma}(n-1)}$
for $i = 1, 2, \ldots, N$
$\sigma_i^2(n) = 1/(\beta(n) r_{ii}(n))$
$\hat{w}_i(n) = r_i^{-1}(n)\left(z_i(n) - \mathbf{r}_{i\neg i}^T(n)\hat{\mathbf{w}}_{\neg i}(n)\right)$
$\alpha_i(n) = \sqrt{b_i(n-1)/(\beta(n)\hat{w}_i^2(n) + r_i^{-1}(n))}$
$\breve{\alpha}_i(n) = 1/\alpha_i(n) + 1/b_i(n-1)$
end for
for $i = 1, 2, \ldots, N-1$
$\omega_i(n) = 2b_{i+1}(n-1)/\nu$
$\varrho_i(n) = \breve{\alpha}_i(n) + 2\breve{b}_{i-1}(n-1)/\nu$
$b_i(n) = \sqrt{\frac{\omega_i(n)}{\varrho_i(n)}} \frac{K_2(\sqrt{\omega_i(n)\varrho_i(n)})}{K_1(\sqrt{\omega_i(n)\varrho_i(n)})}$
$\breve{b}_i(n) = \sqrt{\frac{\varrho_i(n)}{\omega_i(n)}} \frac{K_0(\sqrt{\omega_i(n)\varrho_i(n)})}{K_1(\sqrt{\omega_i(n)\varrho_i(n)})}$
end for
$b_N(n) = (\kappa + 1)/(\breve{\alpha}_N(n)/2 + \breve{b}_{N-1}(n)/\nu)$
end for

hence, these are set to the values $\kappa = 10^{-3}$ and $\nu = 10^3$, respectively (i.e., κ and $1/\nu$ take also small values). The proposed algorithm is numerically robust, while its computational complexity is $\mathcal{O}(N^2)$ per iteration, similar to that of the classical RLS and other competing time-adaptive sparsity promoting algorithms. Moreover, due to its Bayesian nature, the proposed algorithm gives not only single point estimates, but also approximate posterior distributions. To the best of our knowledge, AGSVB-CAR is the first time-adaptive parameter estimation algorithm that promotes group sparsity without a priori knowledge of the signal's group sparsity pattern.

V. EXPERIMENTAL RESULTS

A. Experiments on Synthetic Data

In this section we use simulated data to experimentally examine the estimation performance of the proposed online variational schemes[5] described in the previous sections. In our experiments, we consider the standard time-adaptive filtering setup, where our goal is to track a Rayleigh fading, group-sparse, wireless channel. The proposed schemes are compared with four sparsity-imposing time-adaptive algorithms, namely, a) the EM-RLS algorithm proposed in [4], b) the time and norm weighted lasso (TNWL) algorithm, [3], c) the time-adaptive group sparse variational method AGSVB-S which is based on a Student-t prior, [22], and d) the time-adaptive sparse variational Bayes method ASVB-mpL recently reported in [5]. Moreover, we compare the proposed AGSVB-L algorithm with the author's implementation of the state-of-the-art deterministic online group RLS algorithm $\ell_{0,2}$-RLS proposed in [21]. In all experiments, we use as a benchmark an RLS algorithm, termed the "genie-aided" RLS (GARLS), which knows the signal's support set beforehand and operates only on the nonzero signal coefficients. The estimation performance is evaluated using the normalized mean squared error, which is defined as

$$\text{NMSE} = \frac{\mathbb{E}\left[\|\mathbf{w} - \hat{\mathbf{w}}\|^2\right]}{\mathbb{E}\left[\|\mathbf{w}\|^2\right]}, \tag{60}$$

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

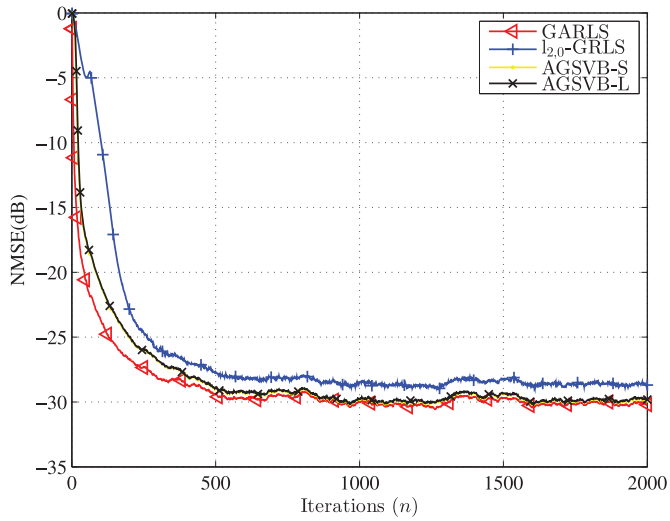


Fig. 3. Comparison of group sparse time-adaptive algorithms with known sparsity pattern.

In the *first experiment*, we consider the case where the sparsity pattern is known beforehand and compare the proposed AGSVB-L algorithm with the deterministic group $\ell_{0,2}$-RLS algorithm of [21][6] and the recently proposed variational Bayes AGSVB-S method based on a Student-t prior, [22]. To this end, we have adopted the experimental settings of [21], for the reason that the $\ell_{0,2}$-RLS algorithm has been fine-tuned for these settings and it has been shown to perform best in comparison to other schemes proposed in [21]. Specifically, the estimation task considers a time-invariant channel of 64 coefficients generated by the standard normal distribution, and having one non-zero group of 4 coefficients. The forgetting factor is set to $\lambda = 0.995$ and the channel input is i.i.d. Gaussian of zero mean and unit variance. Gaussian noise has been also added to the channel output to account for an SNR level of approximately 9 dB. The resulting NMSE curves of our comparison are displayed in Fig. 3. It is observed that both the AGSVB-S and the proposed AGSVB-L algorithm outperform the $\ell_{0,2}$-RLS algorithm, since they reach a lower error floor, closer to that of the benchmark GARLS algorithm. Moreover, there is a notable difference in the convergence speed of the deterministic and the variational Bayes schemes, with the Bayesian methods achieving faster convergence. Notably, as also shown in Fig. 3, there is no evident difference in the performance of the two Bayesian schemes based on the sparsity-imposing Student-t (AGSVB-S)

[5]To make our research reproducible, a Matlab implementation of the proposed online group sparse variational Bayes schemes is available online at http://members.noa.gr/themelis/lib/exe/fetch.php?media=code:agsvb-car_demo_code.zip.

[6]Note that apart from the sparsity pattern, $\ell_{0,2}$-RLS also needs to know the number of non-zero groups.

and multivariate Laplace priors (AGSVB-L), since their NMSE curves almost coincide. Notice, however, that it is the utilization of the Laplace prior that facilitates the development of the conditional autoregressive model described in Section IV.

We next proceed to alter the experimental settings setup, and consider also the case where the sparsity structure is unknown. We now consider a time-varying wireless channel of length $N = 270$. The channel coefficients are generated according to Jake's model, [40], and follow a Rayleigh distribution with normalized Doppler frequency $f_d T_s = 5 \times 10^{-5}$. The channel sparsity pattern varies in each experiment, and it may consist of groups of uniform or arbitrary lengths. The channel input is binary, consisting of ± 1 BPSK symbols and the forgetting factor is set to $\lambda = 0.99$. We assume that the symbols are transmitted in packets of length 2000. Gaussian noise is added to the channel output, while the SNR level is set to 12 dB by adjusting the additive noise variance accordingly.

In the *second experiment*, we examine the ability of the proposed variational schemes to converge to a group sparse solution. For this experiment we simulate a Rayleigh fading channel having a *random* sparsity pattern. Specifically, the nonzero coefficients are randomly organized in groups of length 3 to 5, while the total number of groups in each channel realization varies randomly from 2 to 4. To simulate an abrupt change on the channel coefficients, an extra nonzero group is added to the channel at the $n = 1000$ time mark. Fig. 4 displays the resulting NMSE curves for the benchmark GARLS algorithm, the ASVB-mpL algorithm, the AGSVB-S algorithm with known sparsity structured, the proposed AGSVB-CAR algorithm, and two instances of the proposed AGSVB-L algorithm, one (v1) where the exact knowledge of the signal's group sparsity pattern is provided, and another (v2) where inexact knowledge is used. It is easily observed that the proposed schemes are able to exploit the group sparsity of the parameter vector, since they outperform the ASVB-mpL algorithm, which can be considered as their structure-ignorant analogue. The proposed schemes converge relatively fast to an error floor that is lower than that of the ASVB-mpL algorithm. Especially the AGSVB-CAR algorithm offers a 1 dB improvement over the estimation performance of the ASVB-mpL algorithm, and this improvement comes at negligible additional computational cost. Notice also the impact that the prior knowledge of the sparsity pattern has on the performance of the AGSVB-L algorithm. If the AGSVB-L algorithm knows exactly the signal's group sparsity pattern, it achieves an almost identical steady-state error performance to that of the GARLS algorithm. On the other hand, as expected, if we erroneously inform the AGSVB-L algorithm that the signal is composed of uniform groups of length $d_i = 5, i = 1, \ldots, M$, its performance automatically degrades and becomes worse than that of ASVB-mpL. Again, the estimation performance of the Student-t based scheme (AGSVB-S), proposed in [22], and the proposed multivariate Laplace based scheme AGSVB-L is almost identical, as shown in Fig. 4.

In the *third experiment*, we investigate how the performance of the proposed variational schemes compares to that of existing state-of-the-art methods. To this end, we simulate a wireless channel with groups of fixed length $d_i = 3, i = 1, \ldots, M$, while 2 to 4 nonzero groups are randomly activated in each

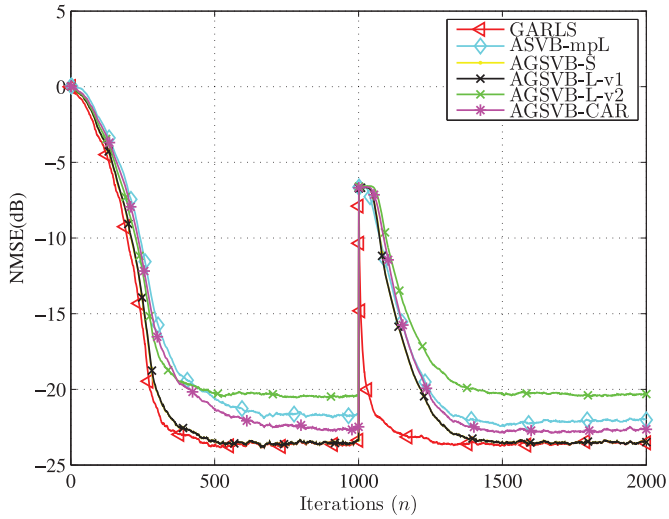


Fig. 4. NMSE curves under slow fading and a sudden channel change.

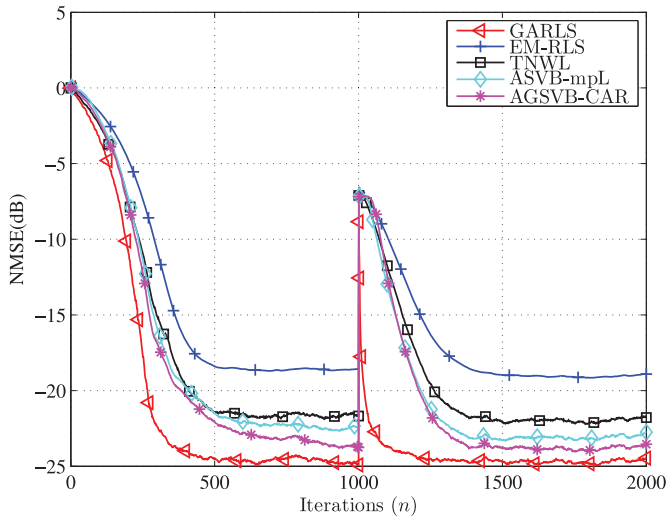


Fig. 5. NMSE curves under slow fading and a sudden channel change.

TABLE III
DETECTION RATES OF THE ALGORITHMS TNWL, ASVB-MPL AND AGSVB-CAR

	false pos.	false neg.	true pos.	true neg.
TNWL	0.011	0.125	0.989	0.874
ASVB-mpL	0.01	0.089	0.989	0.91
AGSVB-CAR	0.007	0.055	0.992	0.944

channel realization. As in the previous experiment, an extra nonzero group is generated at the thousandth time iteration. Fig. 5 displays the NMSE curves for all comparing algorithms versus time iterations. Notice again that the proposed AGSVB-CAR algorithm has the steady state error that reaches closest to the lower bound set by the GARLS. At this point, we should notice that extra experiments have been conducted so as to fine-tune the parameters of the deterministic EM-RLS and TNWL algorithms. In contrast, Bayesian methods nullify such computational costs, since, all their parameters are directly inferred from the data.

Furthermore, to elaborate on the capability of the competing algorithms to correctly identify the zero and nonzero coefficients, Table III provides the relevant detection rates for the

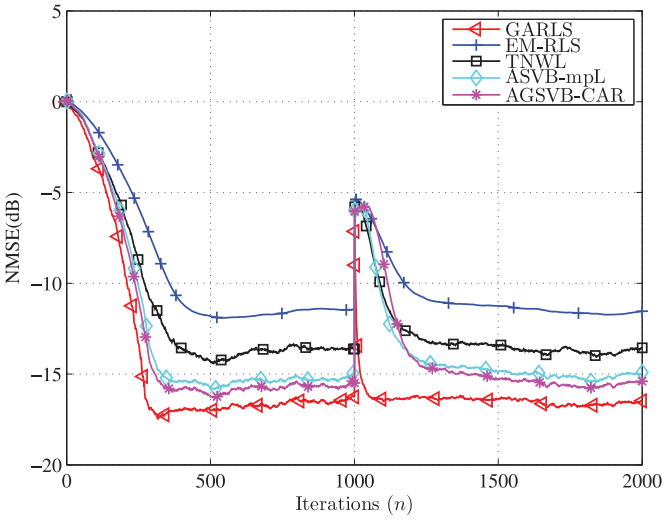


Fig. 6. NMSE curves under fast fading and a sudden channel change.

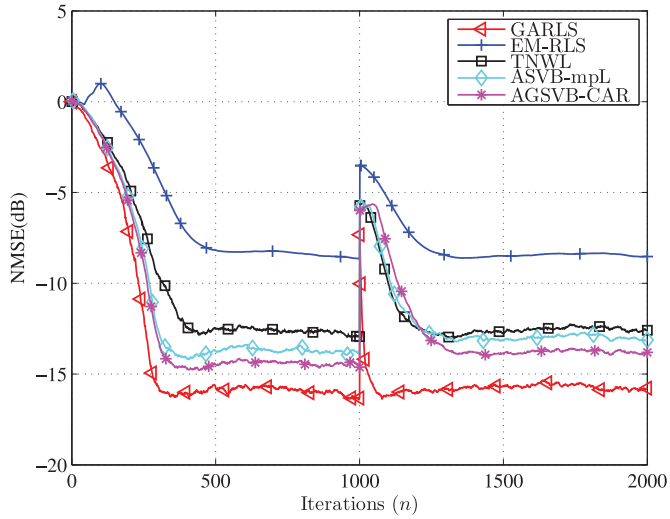


Fig. 7. NMSE curves under slow fading and correlated input.

algorithms TNWL, ASVB-mpL and AGSVB-CAR. Specifically, in each table row, the false positive, the true positive, the false negative and the true negative rates are given for each algorithm, with positive and negative referring to the existence of nonzero and zero coefficients respectively. These rates are computed based on the algorithms' final estimate at the last time iteration of each experimental realization. A brief inspection of these rates reveals that the true positive and true negative rates are much higher that their false counterparts, so that we may say that all algorithms are vastly successful in detecting the signal's sparsity. However, as shown in Table III, the proposed AGSVB-CAR provides the best detection rates, which is an indication that the proposed algorithm is able to identify the signal's clustered sparsity structure more accurately than its competitors.

In the next experiments, we test the performance of the comparing algorithms in the cases of a) a fast fading channel, b) correlated input, c) different SNR levels, and d) different sparsity levels. First, to simulate a fast fading channel, we increase the normalized Doppler frequency to $f_d T_s = 5 \times 10^{-4}$. The experimental settings of the second experiment is used, with the difference that the forgetting factor is now set at $\lambda = 0.98$. Fig. 6 displays the resulting NMSE curves that are in accordance with the previous experiment, apart from the fact that we observe a reasonable increase on all error floors. Moreover, by utilizing the settings of the second experiment, a correlated symbol sequence (generated by a second order autoregressive model with parameters $c_1 = 0.7$ and $c2 = 0.1$ and white Gaussian noise variance equal to 10^{-4}) is used for the channel input, and the resulting NMSE curves are displayed in Fig. 7. Again, we observe that the proposed AGSVB-CAR algorithm achieves the best performance. Next, we investigate the estimation performance of the proposed online schemes for various SNR levels. Fig. 8 plots the computed NMSE for all algorithms. It is easily observed that the closeness of the performance of AGSVB-CAR to that of GARLS is consistent in all SNR levels. Finally, to demonstrate the performance of the proposed method in terms of sparsity, in our last experiment we utilize again the settings of the second experiment, with the difference that the channel length is now set at $N = 570$. By generating random channel and noise realizations, the NMSE curves of Fig. 9 are retrieved. A simple inspection of Fig. 9 reveals that the proposed AGSVB-CAR algorithm performs better in the high sparsity region, where it

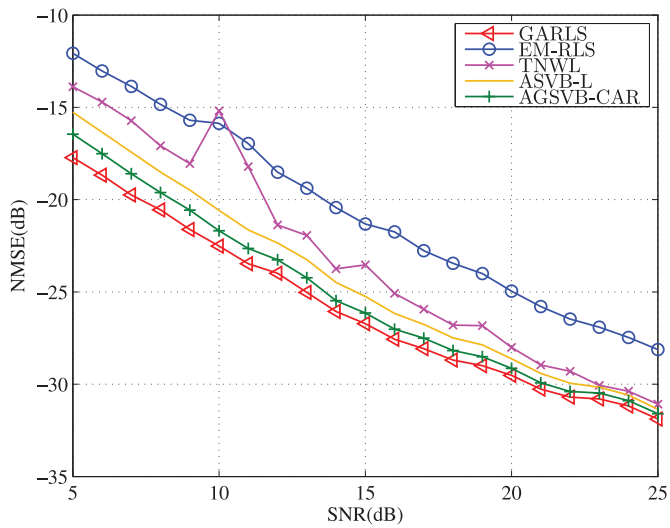


Fig. 8. NMSE curves for different levels of SNR.

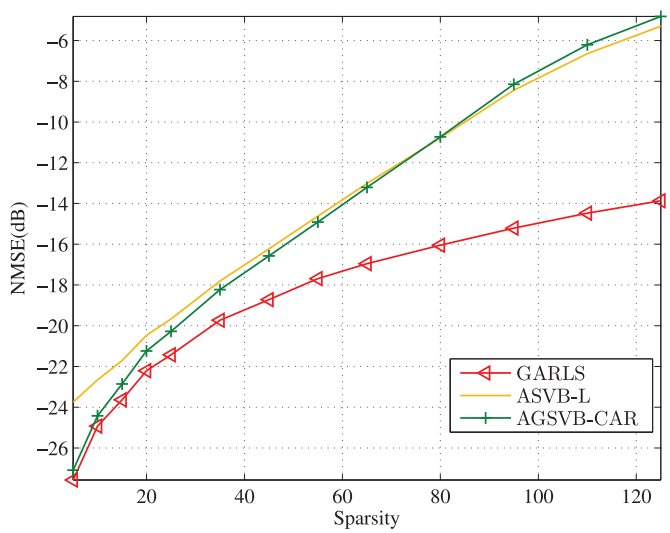


Fig. 9. NMSE curves for different levels of sparsity.

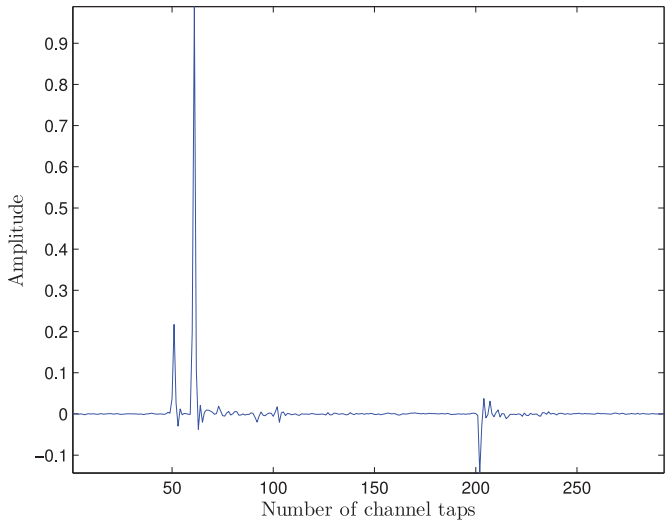


Fig. 10. The real part of a terrestrial HDTV channel with 293 coefficients.

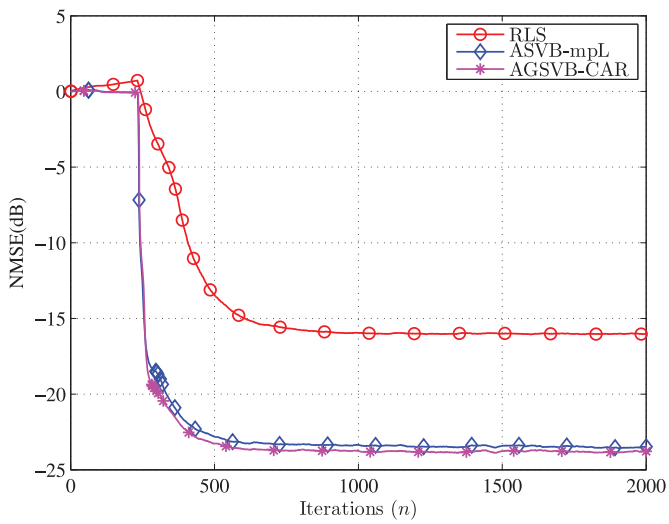


Fig. 11. NMSE curves for the estimation of a real terrestrial HDTV channel.

nearly reaches the performance of GARLS. However, when the number of nonzero coefficients in the signal exceeds a certain ratio (80/570 in this experiment), then it is the structure-ignorant ASVB-L that offers slightly better estimation performance.

B. Experiments on a Real Wireless Channel

In this experiment, we consider the same adaptive filtering setup as in Section V.A, and we now proceed to estimate a measured multipath terrestrial HDTV channel, [7]. The real part of the channel has $N = 293$ coefficients and is displayed in Fig. 10. As shown in the figure, the channel's impulse response can be characterized as group sparse owning to the clustered positioning of its nonzero coefficients. For this experiment we use again a binary input sequence of ± 1 symbols, organized in packets of length 2000. Zero-mean Gaussian noise is added at the channel output with an SNR level of 18 dB. Considering no prior knowledge on the channel group sparsity pattern, we compare in this experiment a) the classical RLS, b) the ASVB-mpL, and c) the AGSVB-CAR algorithms. Fig. 11 displays the resulting NMSE curves. It is observed that a burn-in period of almost 250 time moments

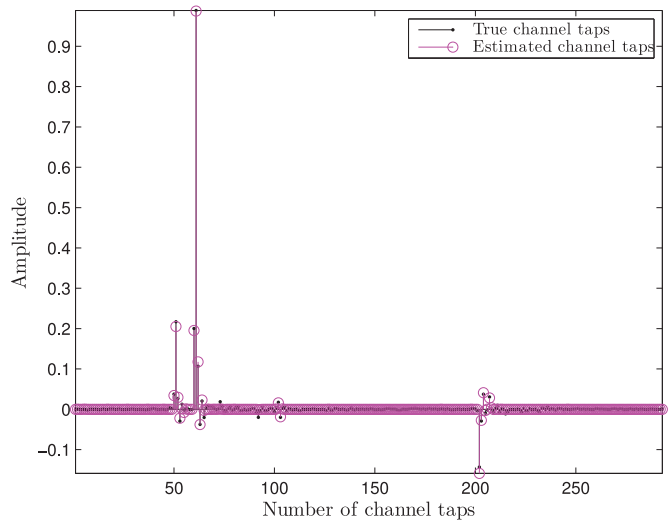


Fig. 12. Estimated channel taps obtained using the proposed AGSVB-CAR algorithm.

is required for the channel input to convolve with the channel nonzero coefficients. After this burn-in period, the algorithms converge to an error floor within about 200 time iterations for ASVB-mpL and AGSVB-CAR and 400 iterations for RLS. We observe in Fig. 11 that AGSVB-CAR exploits the group sparse nature of the channel to converge faster and outperform ASVB-mpL for about 1 dB. As expected, the performance of the sparsity-ignorant RLS algorithm is inferior enough to compare with the remaining sparsity-cognizant schemes. Finally, Fig. 12 verifies that the proposed AGSVB-CAR scheme estimates very accurately the channel coefficients.

VI. Conclusion

Two novel online variational Bayes schemes have been proposed in this paper, that recursively estimate *group sparse* time-varying signals, for both cases of known and unknown group sparsity pattern. The first one places a multivariate Laplace prior over separate coefficient groups defined by the sparsity pattern and the variational Bayes framework is exploited to perform online inference. The problem becomes much more challenging when under time-varying conditions the sparsity pattern is unknown, and it has been, thus far, not addressed by existing signal processing methods. The second scheme tackles this problem by modifying the Bayesian model of the first scheme so as to organize the scale parameters of the Laplace distribution in a conditional autoregressive model. In this way, correlation among individual parameters shrinks the signal towards zero in a structured manner, and, hence, *group sparse* solutions are promoted. Experiments on simulated and real data show that the proposed schemes exploit successfully the signal group sparsity and yield improved estimation performance, when compared to state-of-the-art algorithms.

Appendix A

In this Appendix we analytically derive the multivariate Laplace-type distribution that is placed over $\mathbf{w}$ in the hierarchical Bayesian model of Section III and show that in a MAP estimation setting, this prior is equivalent to utilizing the group

sparsity promoting $\ell_{1,2}$-norm. First, from (5) and (6) we may write,

$$p(\mathbf{w}|\beta,\mathbf{b})=\prod_{i=1}^{M}p(\mathbf{w}_i|\beta,b_i). \tag{61}$$

Then, each factor $p(\mathbf{w}_i|\beta,b_i)$ can be computed as

$$\begin{aligned}p(\mathbf{w}_i|\beta,b_i)&=\int_0^\infty p(\mathbf{w}_i|\beta,\alpha_i)p(\alpha_i|b_i)d\alpha_i\\&=\frac{(2\pi)^{\frac{d_i}{2}}\beta^{\frac{d_i}{2}}}{\Gamma(\frac{d_i+1}{2})}\left(\frac{b_i}{2}\right)^{\frac{d_i+1}{2}}\\&\int_0^\infty\alpha_i^{-\frac{3}{2}}\exp\left[-\frac{1}{2}\left(\beta\|\mathbf{w}_i\|^2\alpha_i+\frac{b_i}{\alpha_i}\right)\right]d\alpha_i.\end{aligned} \tag{62}$$

The integral at the right hand side of (62) can be thought of as the inverse of the normalizing constant of a GIG distribution over α_i, with parameters $-1/2,\beta\|\mathbf{w}_i\|^2$ and b_i. Hence, we easily get

$$p(\mathbf{w}_i|\beta,b_i)=\frac{(2\pi)^{\frac{d_i}{2}}\beta^{\frac{d_i}{2}}}{\Gamma(\frac{d_i+1}{2})}\left(\frac{b_i}{2}\right)^{\frac{d_i+1}{2}}\frac{2K_{-1/2}(\sqrt{\beta b_i\|\mathbf{w}_i\|^2})}{(\beta\|\mathbf{w}_i\|^2)^{-\frac{1}{4}}b_i^{\frac{1}{4}}}. \tag{63}$$

Then, exploiting the identity

$$K_{-1/2}(z)=\sqrt{\frac{\pi}{2z}}\exp[-z] \tag{64}$$

yields

$$p(\mathbf{w}_i|\beta,b_i)=\frac{2^{-d_i}\pi^{-\frac{d_i-1}{2}}b_i^{d_i}\beta^{\frac{d_i}{2}}}{\Gamma(\frac{d_i+1}{2})}\exp\left[-\sqrt{b_i\beta}\|\mathbf{w}_i\|\right], \tag{65}$$

and (61) becomes

$$p(\mathbf{w}|\beta,\mathbf{b})=\prod_{i=1}^{M}\frac{b_i^{d_i}\beta^{\frac{d_i}{2}}}{2^{d_i}\pi^{\frac{d_i-1}{2}}\Gamma(\frac{d_i+1}{2})}\exp\left[-\sqrt{b_i\beta}\|\mathbf{w}_i\|\right]. \tag{66}$$

Equation (66) is a multivariate Laplace-type distribution. In a MAP estimation scenario, maximizing the posterior of $\mathbf{w}$ with respect to the Laplace-type distribution in (66), leads to the following optimization problem,

$$\max_{\mathbf{w}}\left\{p(\mathbf{y}|\mathbf{w},\beta)p(\mathbf{w}|\beta,\mathbf{b})\right\}, \tag{67}$$

which utilizing (3) and (66) becomes

$$\min_{\mathbf{w}}\left\{\frac{\beta}{2}\left\|\mathbf{\Lambda}^{\frac{1}{2}}(\mathbf{y}-\mathbf{X}\mathbf{w})\right\|^2+\sum_{i=1}^{M}\sqrt{b_i\beta}\|\mathbf{w}_i\|\right\}. \tag{68}$$

The regularizing term in (68) is the *weighted* $\ell_{1,2}$ norm of $\mathbf{w}$, which is known to promote group sparsity, [10].

APPENDIX B

In this Appendix we derive the posterior approximating distribution for the scale parameter vector $\mathbf{b}$ in (48). Utilizing (9), (40) and (41), we compute

$$\begin{aligned}\log q(b_i)&\propto\langle\log p(\alpha_i|b_i)+\log p(b_i)\rangle\\&\propto\left\langle-2\log\alpha_i-\frac{b_i}{2}\frac{1}{\alpha_i}+\log\frac{b_i}{2}-\log b_i-\frac{b_{i+1}}{\nu}\frac{1}{b_i}-\frac{b_i}{\nu b_{i-1}}\right\rangle\\&\propto\left\langle-b_i\left(\frac{1}{2\alpha_i}+\frac{1}{\nu b_{i-1}}\right)-\frac{b_{i+1}}{\nu}\frac{1}{b_i}\right\rangle\Rightarrow\\q(b_i)&\propto\exp\left[-b_i\left(\frac{1}{2}\left\langle\frac{1}{\alpha_i}\right\rangle+\frac{1}{\nu}\left\langle\frac{1}{b_{i-1}}\right\rangle\right)-\frac{\langle b_{i+1}\rangle}{\nu}\frac{1}{b_i}\right],\end{aligned} \tag{69}$$

and since (69) has to integrate to one, $q(b_i)$ is a GIG distribution with parameters $\varrho_i=\left\langle\frac{1}{\alpha_i}\right\rangle+\frac{2}{\nu}\left\langle\frac{1}{b_{i-1}}\right\rangle$ and $\omega_i=2\langle b_{i+1}\rangle/\nu$. The mean of the GIG pdf is computed as

$$\begin{aligned}\langle b_i\rangle&=\frac{(\varrho_i/\omega_i)^{1/2}}{2K_1\left(\sqrt{\varrho_i\omega_i}\right)}\int_0^\infty b_i\exp\left[-\frac{1}{2}\left(\varrho_i b_i+\omega_i\frac{1}{b_i}\right)\right]db_i\\&=\left(\frac{\varrho_i}{\omega_i}\right)^{1/2}\frac{2}{2K_1\left(\sqrt{\varrho_i\omega_i}\right)}\left(\frac{\omega_i}{\varrho_i}\right)^{2/2}K_1\left(\sqrt{\varrho_i\omega_i}\right)\\&=\left(\frac{\omega_i}{\varrho_i}\right)^{1/2}\frac{K_2\left(\sqrt{\varrho_i\omega_i}\right)}{K_1\left(\sqrt{\varrho_i\omega_i}\right)},\end{aligned} \tag{70}$$

from which (52) follows. In a similar way, the inverse moment $\left\langle\frac{1}{b_i}\right\rangle$ is computed as

$$\begin{aligned}\left\langle\frac{1}{b_i}\right\rangle&=\frac{(\varrho_i/\omega_i)^{1/2}}{2K_1\left(\sqrt{\varrho_i\omega_i}\right)}\int_0^\infty\frac{1}{b_i}\exp\left[-\frac{1}{2}\left(\varrho_i b_i+\omega_i\frac{1}{b_i}\right)\right]db_i\\&=\left(\frac{\varrho_i}{\omega_i}\right)^{1/2}\frac{2}{2K_1\left(\sqrt{\varrho_i\omega_i}\right)}\left(\frac{\omega_i}{\varrho_i}\right)^{0}K_0\left(\sqrt{\varrho_i\omega_i}\right)\\&=\left(\frac{\varrho_i}{\omega_i}\right)^{1/2}\frac{K_0\left(\sqrt{\varrho_i\omega_i}\right)}{K_1\left(\sqrt{\varrho_i\omega_i}\right)},\end{aligned} \tag{71}$$

and (53) follows.

REFERENCES

[1] S. S. Haykin, *Adaptive Filter Theory*. : Pearson Education India, 2007.

[2] A. H. Sayed, *Adaptive Filters*. New York, NY, USA: Wiley-IEEE Press, 2008.

[3] 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.

[4] 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.

[5] K. E. Themelis, A. A. Rontogiannis, and K. D. Koutroumbas, "A variational Bayes framework for sparse adaptive estimation," *IEEE Trans. Signal Process.*, vol. 62, no. 18, pp. 4723–4736, Sep. 2014.

[6] S. Bengio, F. Pereira, Y. Singer, and D. Strelow, "Group sparse coding," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, Eds. Red Hook, NY, USA: Curran Associates, 2009, pp. 82–89.

[7] A. A. Rontogiannis and K. Berberidis, "Efficient decision feedback equalization for sparse wireless channels," *IEEE Trans. Wireless Commun.*, vol. 2, no. 3, pp. 570–581, May 2003.

[8] Q. F. Tan and S. Narayanan, "Novel variations of group sparse regularization techniques with applications to noise robust automatic speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1337–1346, May 2012.

[9] S. Li, H. Yin, and L. Fang, "Group-sparse representation with dictionary learning for medical image denoising and fusion," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 12, pp. 3450–3459, Dec. 2012.

[10] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc. B (Statist. Methodol.)*, vol. 68, no. 1, pp. 49–67, 2006.

[11] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc.*, vol. 58, no. 1, pp. 267–288, 1996.

[12] S. Raman, T. J. Fuchs, P. J. Wild, E. Dahl, and V. Roth, "The Bayesian group-lasso for analyzing contingency tables," in *Proc. 26th Annu. ACM Int. Conf. Mach. Learn.*, 2009, pp. 881–888.

[13] T. Park and G. Casella, "The Bayesian lasso," *J. Amer. Statist. Assoc.*, vol. 103, no. 482, pp. 681–686, Jun. 2008.

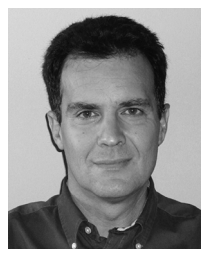
[14] C. Leng, M.-N. Tran, and D. Nott, "Bayesian adaptive lasso," *Ann. Inst. Statist. Math.*, vol. 66, no. 2, pp. 221–244, 2014.

[15] S. Babacan, S. Nakajima, and M. Do, "Bayesian group-sparse modeling and variational inference," *IEEE Trans. Signal Process.*, vol. 62, no. 11, pp. 2906–2921, Jun. 2014.

[16] Z. Zhang and B. Rao, "Extension of SBL algorithms for the recovery of block sparse signals with intra-block correlation," *IEEE Trans. Signal Process.*, vol. 61, no. 8, pp. 2009–2015, Apr. 2013.
[17] V. Rockova and E. Lesaffre, "Incorporating grouping information in Bayesian variable selection with applications in genomics," *Bayesian Anal.*, vol. 9, no. 1, pp. 221–258, 2014.
[18] L. Yu, H. Sun, G. Zheng, and J. P. Barbot, "Model based Bayesian compressive sensing via local Beta process," *Signal Process.*, vol. 108, no. 0, pp. 259–271, 2015.
[19] J. Fang, Y. Shen, H. Li, and P. Wang, "Pattern-coupled sparse Bayesian learning for recovery of block-sparse signals," *IEEE Trans. Signal Process.*, vol. 63, no. 2, pp. 360–372, Jan. 2015.
[20] Y. Chen and A. Hero, "Recursive $\ell_{1,\infty}$ group lasso," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 3978–3987, Aug. 2012.
[21] E. M. Eksioglu, "Group sparse RLS algorithms," *Int. J. Adapt. Control Signal Process.*, vol. 28, no. 12, pp. 1398–1412, 2014.
[22] K. E. Themelis, A. A. Rontogiannis, and K. D. Koutroumbas, "Group-sparse adaptive variational Bayes estimation," in *Proc. Signal Process. Conf. (EUSIPCO)*, Sep. 2014, pp. 1342–1346.
[23] A. Gelman, "Prior distributions for variance parameters in hierarchical models," *Bayesian Anal.*, vol. 1, no. 3, pp. 515–534, 2006.
[24] H. Wang and C. Leng, "A note on adaptive group lasso," *Comput. Statist. Data Anal.*, vol. 52, no. 12, pp. 5277–5286, 2008.
[25] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *J. Roy. Statist. Soc. B (Methodol.)*, vol. 36, no. 1, pp. 99–102, 1974.
[26] H. Zou, "The adaptive lasso and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 101, no. 476, pp. 1418–1429, Dec. 2006.
[27] L. Meier, S. Van De Geer, and P. Bhlmann, "The group lasso for logistic regression," *J. Roy. Statist. Soc. B (Statist. Methodol.)*, vol. 70, no. 1, pp. 53–71, 2008.
[28] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statist. Comput.* vol. 10, no. 3, pp. 197–208, 2000 [Online]. Available: http://dx.doi.org/10.1023/A%3A1008935410038
[29] T. J. Jaakkola and M. I. Jordan, "Bayesian parameter estimation via variational methods," *Statist. Comput.*, vol. 10, no. 1, pp. 25–37, 2000.
[30] H. Attias, "A variational Bayesian framework for graphical models," in *Advances in Neural Information Processing Systems*. Cambridge: MIT Press, 2000, vol. 12, pp. 209–215.
[31] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag, 2006.
[32] P. Tseng, "Convergence of a block coordinate descent method for non-differentiable minimization," *J. Optim. Theory Appl.*, vol. 109, no. 3, pp. 475–494, 2001.
[33] K. E. Themelis, A. A. Rontogiannis, and K. D. Koutroumbas, "Online Bayesian group sparse parameter estimation using a generalized inverse Gaussian Markov chain," in *Proc. Signal Process, Conf, (EUSIPCO*, Sep. 2015.
[34] 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, 2012.
[35] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *J. Roy. Statist. Soc. B (Methodol.)*, pp. 192–236, 1974.
[36] A. E. Gelfand and P. Vounatsou, "Proper multivariate conditional autoregressive models for spatial data analysis," *Biostatistics*, vol. 4, no. 1, pp. 11–15, 2003.
[37] D. Brook, "On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems," *Biometrika*, pp. 481–483, 1964.
[38] J. Besag and C. Kooperberg, "On conditional and intrinsic autoregressions," *Biometrika*, vol. 82, no. 4, pp. 733–746, 1995.
[39] E. L. Lehmann and G. Casella, *Theory of Point Estimation*. New York, NY, USA: Springer Science & Business Media, 1998, vol. 31.
[40] , 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 (5-yr) Diploma degree 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. in communications and signal processing from the University of Athens, Greece, in 1997. From 1998 to 2003, he was with the University of Ioannina, Ionnina, Greece. 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 has been a Senior Researcher. His research interests are in the general areas of statistical signal processing and wireless communications with emphasis on adaptive estimation, hyperspectral image processing, Bayesian compressive sensing, channel estimation/equalization and cooperative communications. He has served 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). 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.