ONLINE BAYESIAN GROUP SPARSE PARAMETER ESTIMATION USING A GENERALIZED INVERSE GAUSSIAN MARKOV CHAIN

Konstantinos E. Themelis, Athanasios A. Rontogiannis, Konstantinos D. Koutroumbas

IAASARS, National Observatory of Athens, GR-15236, Penteli, Greece
{themelis,tronto,koutroum}@noa.gr

ABSTRACT

In this paper we develop a variational Bayes algorithm for the adaptive estimation of time-varying, *group sparse* signals. First, we propose a hierarchical Bayesian model that captures the sparsity structure of the signal. Sparsity is imposed by a multivariate Laplace distribution, which is known to be the Bayesian analogue of the adaptive lasso. Sparsity structure is then expressed via a novel *generalized inverse Gaussian Markov chain*, defined on the parameters of the Laplace distribution. The conjugacy of the model's prior distributions permits the development of an efficient online variational Bayes algorithm that performs inference on the model parameters. Experimental results verify that capturing sparsity structure leads to improvements on estimation performance.

Index Terms— online inference, variational Bayes, Markov random field, generalized inverse Gaussian distribution, group sparsity

1. INTRODUCTION

In recent years, advances in the area of compressive sensing have sparkled new interest in almost every aspect of modern signal processing theory, including, adaptive signal estimation. The challenge here is to adaptively estimate time varying signals that exhibit *sparsity*. Scarce research attempts have also been made to the direction of estimating varying signals that are *group sparse*, which is the case studied in this paper too.

Actually, the development of adaptive estimation techniques for time-varying group sparse signals is driven by group-lasso based techniques. In this realm, mixed norms are utilized, such as the $\ell_{1,\infty}$ or the $\ell_{1,2}$ norm, that are known to penalize signal coefficients in a "group-wise" manner. In [1], an $\ell_{1,\infty}$ norm regularizer is incorporated in the adaptive setting of the recursive least squares (RLS) algorithm. A deterministic group sparse RLS algorithm is also put forth in [2], where an approximation of the $\ell_{p,0}$ pseudonorm is used for the regularizing term. Unfortunately, these deterministic methods suffer from two main setbacks; their estimation performance is parameter dependent, and they require a priori knowledge of the signal's support structure. In the Bayesian framework, an adaptive variational Bayes estimator has been recently proposed in [3], which is fully automatic, but also requires some grouping information.

In this paper we develop an adaptive variational Bayes estimator that has the ability to identify the grouping information of the signal. To this end, we adopt and extend the hierarchical Bayesian model presented in [4]. Specifically, we model sparsity on the signal coefficients using a multivariate Laplace distribution. The grouping information is then reflected on the parameters of the Laplace distribution, which are imposed to be interrelated in a Markov chain. To achieve this, we extend the formulation of the Gamma Markov random field, that has been proposed in [5] for modeling the positive variances of Gaussian audio signals, to define a generalized inverse Gaussian (GIG) random field. This formulation is important in order to retain the conjugacy of our model's prior distributions. A variational Bayes algorithm is then developed to perform approximate online inference on the model parameters. The advantages of the proposed estimator are that (a) it adjusts to any grouping pattern automatically, (b) it is fully automatic, (c) it has quadratic complexity, and (d) as a Bayesian technique, it provides entire approximate posterior distributions for the model parameters instead of point estimates. Experimental results are provided that validate the superior performance of the proposed method, when compared to state of the art methods.

Notation: Matrices are denoted by bold capital letters, e.g. $\mathbf{X}$, vectors are written with bold letters, e.g. $\mathbf{x}$. $\mathbf{X}_{\neg i}$ and $\mathbf{x}_{\neg i}$ denote matrix $\mathbf{X}$ and vector $\mathbf{x}$ after excluding its ith column or ith element, respectively. $\mathbf{I}_M$ is the $M \times M$ identity matrix. $\mathcal{N}(x; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. $\mathcal{GIG}(x; \zeta, \tau, \upsilon)$ is the generalized inverse Gaussian distribution defined as

$$\mathcal{GIG}(x;\zeta,\tau,\upsilon) = \frac{\exp\left[(\zeta-1)\log x - \tau x - \frac{\upsilon}{x}\right]\tau^{\frac{\zeta}{2}}}{2\upsilon^{\frac{\zeta}{2}}\mathcal{K}_\zeta(2\sqrt{\tau\upsilon})},$$

where $x \geq 0$, $\tau \geq 0$, $\upsilon \geq 0$, and $\mathcal{K}_\zeta(\cdot)$ denotes the modified Bessel function of second kind with ζ degrees of free-

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: ARISTEIA- HSI-MARS-1413.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

dom. 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 the inverse Gamma pdf has the form

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

2. PROBLEM FORMULATION

Consider a *group sparse* time-varying weight vector $\mathbf{w}(n) = [w_1(n), w_2(n), \ldots, w_N(n)]^T \in \mathbb{R}^N$, which has $\xi \ll N$ non-zero elements, and n is the time index. The sparsity structure of $\mathbf{w}(n)$ means that its nonzero elements occur in blocks rather than being independently distributed in random positions. In addition, we assume that there is no knowledge of the sparsity structure, neither in terms of the number and sizes of the blocks, nor in terms of their positions in the parameter vector. Our objective is to estimate the varying vector $\mathbf{w}(n)$ with the help of some noisy data, $\mathbf{y}(n) = [y(1), y(2), \ldots, y(n)]^T$, observed up to time n. The 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 $\mathbf{X}(n) = [\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(n)]^T$ is a known $n \times N$ input data matrix, and $\boldsymbol{\epsilon}(n)$ stands for additive noise which is assumed to be zero-mean Gaussian distributed, i.e., $\boldsymbol{\epsilon}(n) \sim \mathcal{N}(\boldsymbol{\epsilon}(n)|\mathbf{0}, \beta^{-1}\mathbf{I}_M)$, where β denotes the noise precision.

The cost function to be minimized for estimating $\mathbf{w}(n)$ is

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

where λ is the well known forgetting factor, $0 \ll \lambda < 1$, and $\boldsymbol{\Lambda} = \mathrm{diag}([\lambda^{n-1}, \lambda^{n-2}, \ldots, 1]^T)$. One way to solve this problem is via the celebrated RLS algorithm, which, however, cannot take advantage of the presence of the group sparsity property.

In this paper we examine the problem in (2) from a Bayesian viewpoint and propose a novel hierarchical Bayesian model that imposes *group sparsity*. To automatically capture the inherent correlation among adjacent coefficients of $\mathbf{w}(n)$, which will help us reveal the underlying sparsity structure, we consolidate a GIG Markov random chain in our hierarchical Bayesian model. A time adaptive variational Bayes algorithm is then developed to perform approximate inference in an online setting.

3. BAYESIAN MODELING

This section describes a structured sparsity imposing hierarchical Bayesian model. For notational expediency, we temporarily drop the dependency of the model parameters on the time index n. Time indexing will be re-introduced in Section 5, where online Bayesian inference is described.

Fig. 1. The proposed GIG Markov chain.

The Gaussian noise assumption combined with the weighted least squares function in (2) give rise to the data likelihood

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

Working within a Bayesian framework, we augment the likelihood function in (3) with suitable priors over the model parameters $\{\mathbf{w}, \beta\}$. Motivated by [4], we formulate a hierarchical Bayesian model, where, in the first level, $\mathbf{w}$ and β are distributed as

$$p(\mathbf{w}|\boldsymbol{\alpha},\beta) = \prod_{i=1}^{N} \mathcal{N}\left(w_i|0, \beta^{-1}\alpha_i^{-1}\right) \tag{4}$$

where α_i's are precision parameters, and

$$p(\beta) = \mathcal{G}(\beta;\rho,1/\delta) \tag{5}$$

where ρ and δ are fixed positive hyperparameters, set close to zero ($\rho = \delta = 10^{-6}$). In the second level of hierarchy, the precision parameters α_i's are given an inverse gamma distribution,

$$p(\boldsymbol{\alpha}) = \prod_{i=1}^{N} \mathcal{IG}\left(\alpha_i|1, b_i/2\right) \tag{6}$$

where b_i's are scale parameters.

We now delve into the distribution of b_i's, which will be added as a third hierarchical level to our Bayesian model. As shown in [6], the scale parameters b_i's can be interpreted as penalty parameters in the adaptive lasso formulation, and, hence, their estimation is of prime interest. A common approach is to assign a suitable non-informative prior over b_i's so as to infer them directly from the data, e.g., [7, 4]. An interesting idea, recently coined in [8], is that the grouping information can be properly embedded in the prior distribution of the penalties b_i's, in order to shrink the original signal $\mathbf{w}$ towards zero in a *structured* manner. However, in [8], an additional group-membership matrix is required that provides information on the grouping structure of the signal coefficients. In this work, we assume no prior knowledge on the sparsity structure and in order to detect it, we propose to impose correlation among adjacent vector coefficients by organizing the corresponding penalty parameters b_i's in a *generalized inverse Gaussian Markov chain*.

The proposed Markov chain is visualized in the undirected graph of Fig. 1. Each node in the graph corresponds

to a penalty parameter b_i, while the edges between adjacent variables encode the dependency between the penalties.

To build the Markov chain based on the generalized inverse Gaussian distribution, we consider the following dependencies, $i = 1, 2, \ldots, N$,

$$p(b_i|b_{i-1}) = \mathcal{G}(b_i|\kappa, \nu b_{i-1}) \tag{7}$$

where κ and ν are hyperparameters. Using the conditional distribution in (7) and Bayes law, the complete conditional for each penalty b_i is computed as

$$p(b_i|b_{i-1}, b_{i+1}) = \mathcal{GIG}\left(b_i; 0, \frac{2}{\nu b_{i-1}}, \frac{2b_{i+1}}{\nu}\right). \tag{8}$$

Moreover, the joint pdf of the Markov chain can be expressed as a product of the potential functions

$$p(\mathbf{b}|\nu) = \frac{1}{C}\prod_{n=1}^{N}\phi_b(b_n, \nu)\prod_{n=1}^{N}\phi_e(b_{n-1}, \nu b_n), \tag{9}$$

where C is the normalizing constant, $\phi_b(\zeta, \tau) = \exp[(\tau - 1)\log\zeta]$ and $\phi_e(\zeta, \tau) = \exp[-\tau\zeta - \frac{\tau}{\zeta}]$. The joint pdf in (9) is a GIG distribution and it is conjugate with respect to the priors of the second level of our Bayesian model. This leads to simple forms for the complete conditional posterior distributions of the model parameters and facilitates the use of a variational Bayes algorithm to perform inference.

4. BAYESIAN INFERENCE

Unfortunately, the true posterior $p(\beta, \mathbf{w}, \boldsymbol{\alpha}, \mathbf{b}|\mathbf{y})$ of the proposed model parameters is too complex to be computed using Bayes rule. In the following we resort to a variational Bayes approach that entails (a) the approximation of the true posterior with an approximating pdf $q(\cdot)$, and (b) the minimization of the Kullback-Leibler divergence between the true posterior and $q(\cdot)$, for the estimation of the latter. We assume that $q(\cdot)$ is factorized as

$$q(\beta, \mathbf{w}, \boldsymbol{\alpha}, \mathbf{b}) = q(\beta)\prod_{i=1}^{N}q(w_i)\prod_{i=1}^{N}q(\alpha_i)\prod_{i=1}^{N}q(b_i). \tag{10}$$

The posterior independence in (10) renders the minimization problem tractable, in the sense that each approximating factor can be expressed in closed form, [4]. For the noise precision, we get a posterior approximating Gamma distribution,

$$q(\beta) = \mathcal{G}\left(\beta; \tilde{\rho}, \tilde{\delta}\right), \tag{11}$$

where $\tilde{\rho} = \rho + \frac{M+N}{2}$, $\tilde{\delta} = \left(\delta + \frac{\langle \mathbf{w}^T\mathbf{A}\mathbf{w}\rangle}{2} + \frac{\langle \|\mathbf{y}-\mathbf{X}\mathbf{w}\|^2\rangle}{2}\right)^{-1}$, $\mathbf{A} = \mathrm{diag}(\boldsymbol{\alpha})$, and $\langle\cdot\rangle$ denotes expectation with respect to the corresponding $q(\cdot)$. For the weight vector $\mathbf{w}$ we get

$$q(w_i) = \mathcal{N}(w_i; \mu_i, \sigma_i^2), i = 1, 2, \ldots, N, \tag{12}$$

with $\sigma_i^2 = \langle\beta\rangle^{-1}(\mathbf{x}_i^T\mathbf{x}_i + \langle\alpha_i\rangle)^{-1}$, and $\mu_i = \langle\beta\rangle\sigma_i^2\mathbf{x}_i^T(\mathbf{y} - \mathbf{X}_{\neg i}\boldsymbol{\mu}_{\neg i})$. The approximating posterior of the precision parameters α_i's, $i = 1, 2, \ldots, N$, is computed as

$$q(\alpha_i) = \mathcal{GIG}\left(\alpha_i; -1/2, \langle\beta\rangle\langle w_i^2\rangle, \langle b_i\rangle\right). \tag{13}$$

Finally, the penalty parameters b_i's, $i = 1, 2, \ldots, N-1$, are inferred via the GIG approximating distribution (due to space limitations analytic derivations are omitted)

$$q(b_i) = \mathcal{GIG}\left(b_i; 1, \varrho_i, \omega_i\right), \tag{14}$$

where $\varrho_i = \left\langle\frac{1}{\alpha_i}\right\rangle + \frac{2}{\nu}\left\langle\frac{1}{b_{i-1}}\right\rangle$ and $\omega_i = 2\langle b_{i+1}\rangle/\nu$. For the Nth node of the chain the posterior in (14) simplifies to

$$q(b_N) = \mathcal{G}\left(b_N; \kappa + 1, 2/\varrho_N\right). \tag{15}$$

At this point, it is interesting to notice the interdependency among the parameters of the approximating factors expressed in (11), (12), (13), (14), and (15). This mutual dependency gives rise to the alternating optimization scheme of the variational Bayes algorithm, where a single posterior parameter is updated while the remaining are kept fixed in their most recent values. The required moments of the posterior parameters are computed as

$$\langle\beta\rangle = \frac{2\rho + M + N}{2\delta + \langle\mathbf{w}^T\mathbf{A}\mathbf{w}\rangle + \langle\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2\rangle} \tag{16}$$

$$\langle w_i\rangle = (\mathbf{x}_i^T\mathbf{x}_i + \langle\alpha_i\rangle)^{-1}\mathbf{x}_i^T(\mathbf{y} - \mathbf{X}_{\neg i}\langle\mathbf{w}_{\neg i}\rangle), \tag{17}$$

$$\langle\alpha_i\rangle = \sqrt{\frac{\langle b_i\rangle}{\langle\beta\rangle\langle w_i^2\rangle}}, \qquad \left\langle\frac{1}{\alpha_i}\right\rangle \equiv \breve{\alpha}_i = \frac{1}{\langle\alpha_i\rangle} + \frac{1}{\langle b_i\rangle}, \tag{18}$$

$$\langle b_i\rangle = \sqrt{\frac{\omega_i}{\varrho_i}}\frac{K_2(\sqrt{\omega_i\varrho_i})}{K_1(\sqrt{\omega_i\varrho_i})}, \quad \langle b_N\rangle = (\kappa + 1)\frac{2}{\varrho_N}, \tag{19}$$

$$\left\langle\frac{1}{b_i}\right\rangle \equiv \breve{b}_i = \sqrt{\frac{\varrho_i}{\omega_i}}\frac{K_0(\sqrt{\omega_i\varrho_i})}{K_1(\sqrt{\omega_i\varrho_i})}, \tag{20}$$

$$\left\langle\mathbf{w}^T\mathbf{A}\mathbf{w}\right\rangle = \sum_{i=1}^{N}\langle\alpha_i\rangle\left\langle w_i^2\right\rangle, \quad \left\langle w_i^2\right\rangle = \langle w_i\rangle^2 + \sigma_i^2, \tag{21}$$

and

$$\left\langle\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2\right\rangle = \|\mathbf{y} - \sum_{i=1}^{N}\mathbf{x}_i\langle w_i\rangle\|^2 + \sum_{i=1}^{N}\sigma_i^2\mathbf{x}_i^T\mathbf{x}_i. \tag{22}$$

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. The final estimate for the weight vector is given by the expectation $\langle\mathbf{w}\rangle$, and it is *group sparse*. In the next Section, we show how the proposed variational Bayes algorithm can be extended over to the time-adaptive domain.

5. ADAPTIVE 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 the previous Section. To achieve this, we propose the recursive computation of appropriate fixed-size quantities and map batch iterations to time iterations. As in [4], we define the quantities, $\mathbf{R}(n) = \mathbf{X}^T(n)\mathbf{\Lambda}(n)\mathbf{X}(n) + \mathbf{A}(n-1)$, $\mathbf{z}(n) = \mathbf{X}^T(n)\mathbf{\Lambda}(n)\mathbf{y}(n)$, $d(n) = \mathbf{y}^T(n)\mathbf{\Lambda}(n)\mathbf{y}(n)$, which are easily time-updated as

$$\mathbf{R}(n) = \lambda\mathbf{R}(n-1) + \mathbf{x}(n)\mathbf{x}^T(n) - \lambda\mathbf{A}(n-2) + \mathbf{A}(n-1), \tag{23}$$

$$\mathbf{z}(n) = \lambda\mathbf{z}(n-1) + \mathbf{x}(n)y(n), \tag{24}$$

$$d(n) = \lambda d(n-1) + y^2(n). \tag{25}$$

Utilizing (23) and (24), the weight estimate $\hat{w}_i(n) \equiv \mu_i(n)$ can be calculated recursively as, [4],

$$\hat{w}_i(n) = \frac{1}{r_{ii}(n)}\left(z_i(n) - \mathbf{r}_{\neg i}^T(n)\hat{\mathbf{w}}_{\neg i}(n)\right), \tag{26}$$

where $\mathbf{r}_{\neg i}^T(n) = \mathbf{x}_i^T(n)\mathbf{\Lambda}(n)\mathbf{X}_{\neg i}(n)$ is the i-th row of $\mathbf{R}(n)$ after removing its i-th element $r_{ii}(n)$, and $\hat{\mathbf{w}}_{\neg i}(n)$ is a $(N-1)$-dimensional vector containing all but the ith latest estimate $\hat{w}_i(n)$. Utilizing (25), the noise precision estimate can be efficiently approximated by, [4],

$$\beta(n) = \frac{(1-\lambda)^{-1} + N + 2\rho}{2\delta + d(n) - \mathbf{z}^T(n)\hat{\mathbf{w}}(n-1) + \mathbf{r}^T(n)\boldsymbol{\sigma}(n-1)}, \tag{27}$$

where $\mathbf{r}(n) = \mathrm{diag}(\mathbf{R}(n))$ and $\boldsymbol{\sigma}(n-1)$ is the vector of posterior weight variances at time $n-1$ with $\sigma_i^2(n-1) = 1/(\beta(n-1)r_{ii}(n-1))$. According to (18), α_i's and $\breve{\alpha}_i$'s are time updated as

$$\alpha_i(n) = \sqrt{\frac{b(n-1)}{\beta(n)\hat{w}_i^2(n) + r_{ii}^{-1}(n)}}, \tag{28}$$

and

$$\breve{\alpha}_i(n) = \frac{1}{\alpha_i(n)} + \frac{1}{b_i(n-1)}. \tag{29}$$

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

$$b_i(n) = \sqrt{\frac{\omega_i(n)}{\varrho_i(n)}}\frac{K_2(\sqrt{\omega_i(n)\varrho_i(n)})}{K_1(\sqrt{\omega_i(n)\varrho_i(n)})}, \tag{30}$$

while for the last chain node we have that $b_N(n) = (\kappa+1)2/\varrho_N(n)$. Finally, with respect to (20), $\breve{b}_i$'s are updated as

$$\breve{b}_i(n) = \sqrt{\frac{\varrho_i(n)}{\omega_i(n)}}\frac{K_0(\sqrt{\omega_i(n)\varrho_i(n)})}{K_1(\sqrt{\omega_i(n)\varrho_i(n)})}, i = 1, \ldots, N-1. \tag{31}$$

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

Table 1. The proposed AGSVB-MC algorithm.

The proposed adaptive group sparse variational Bayes Markov chain (AGSVB-MC) algorithm is summarized in Table 1. The proposed algorithm converges in a few iterations and produces group sparse solutions, as shown in the experimental results section. Its computational complexity is $\mathcal{O}(N^2)$.

6. EXPERIMENTAL RESULTS

In this section we assess the performance of the proposed algorithm by considering a time-varying channel estimation problem. We simulate a Rayleigh fading channel of length 128, where the nonzero coefficients are generated using Jakes model, [9], and follow a Rayleigh distribution with normalized Doppler frequency $f_d T_s = 5 \times 10^{-5}$. Moreover, we assume that the channel's sparsity structure pattern is *random*, which is a very challenging setting. Specifically, the nonzero coefficients are randomly organized in groups of length 8 to 10, while the total number of groups in each channel realization is also randomly selected from 2 to 4. The forgetting factor is set to $\lambda = 0.99$. The channel input is a random sequence of binary ± 1 symbols, organized in packets of length 2000. Noise is assumed to be white Gaussian, and its variance is adjusted so as to get an SNR level of 15dB.

The proposed algorithm is compared with (a) the recently proposed AGSVB-S algorithm, [3], which requires

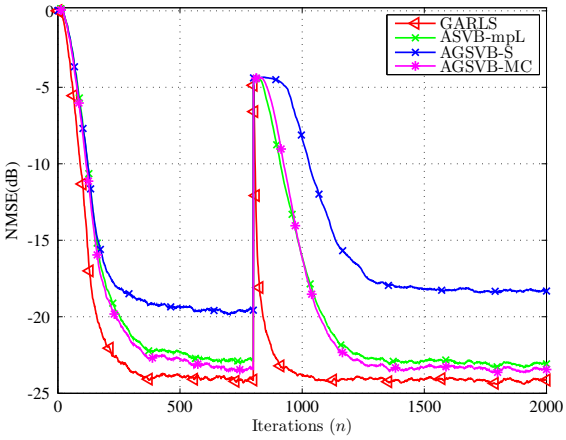

Fig. 2. NMSE curves under slow fading and a sudden channel change.

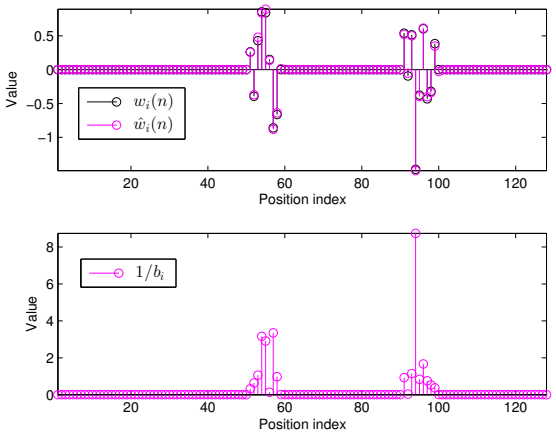

Fig. 3. Estimated channel and inverse penalty coefficients.

prior knowledge of the group signal structure, (b) the ASVB-mpL algorithm, [4], which is sparsity structure-ignorant, and (c) a genie-aided RLS that operates only on the correct support set, and it is thus used as a benchmark. To assess the estimation performance of the comparing algorithms, we use the normalized mean square error, defined as $\text{NMSE} = \mathbb{E}\left[\|\mathbf{w}(n) - \hat{\mathbf{w}}(n)\|^2\right] / \mathbb{E}\left[\|\mathbf{w}(n)\|^2\right]$, where $\hat{\mathbf{w}}(n)$ is the estimate of the channel vector $\mathbf{w}(n)$.

Fig. 2 shows the NMSE curves of the considered algorithms, which are ensemble averages of 100 transmission packets, channels, and noise realizations. It is easily observed that all algorithms converge to their respective error floor after almost 300 time iterations. At time instant $n = 800$, a sudden change is simulated on the channel, caused by the addition of an extra group of nonzero coefficients. This causes a burst on the NMSE curves, which need about 300 iterations to reach an error floor again. The worst performance is obtained by the AGSVB-S algorithm. This is to be expected since AGSVB-S algorithm is designed to operate on a fixed group size (set at D = 8), which is not the case in this experiment. Moreover, the proposed AGSVB-MC algorithm has the overall best performance and it is shown to outperform its structure-ignorant counterpart ASVB-mpL.

Let us now have a closer look at AGSVB-MC's channel estimate, produced at the last iteration of Fig. 2. Fig. 3 displays both the estimated channel coefficients (at the top) and the corresponding *inverse* penalty parameters b_i's (at the bottom). Notice that the estimated weight vector is indeed group sparse, and that its nonzero entries are positioned at the exact indexes where the penalty parameters have the lowest value. Hence, it can be concluded that the structural information is definitely captured by the penalty parameters, which, in our algorithm, have the role of detecting the support set.

REFERENCES

[1] Y. Chen and A.O. Hero, "Recursive $\ell_{1,\infty}$ group lasso," *Signal Processing, IEEE Transactions on*, vol. 60, no. 8, pp. 3978–3987, Aug 2012.

[2] E. M. Eksioglu, "Group sparse RLS algorithms," *International Journal of Adaptive Control and Signal Processing*, vol. 28, no. 12, pp. 1398–1412, 2014.

[3] K.E. Themelis, A.A. Rontogiannis, and K.D. Koutroumbas, "Group-sparse adaptive variational Bayes estimation," in *Signal Processing Conference EUSIPCO*, Sept 2014, pp. 1342–1346.

[4] K.E. Themelis, A.A. Rontogiannis, and K.D. Koutroumbas, "A variational Bayes framework for sparse adaptive estimation," *Signal Processing, IEEE Transactions on*, vol. 62, no. 18, pp. 4723–4736, Sept 2014.

[5] O. Dikmen and A.T. Cemgil, "Gamma Markov random fields for audio source modeling," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 589–601, March 2010.

[6] K. E. Themelis, A. A. Rontogiannis, and K. D. Koutroumbas, "A novel hierarchical Bayesian approach for sparse semisupervised hyperspectral unmixing," *Signal Processing, IEEE Transactions on*, vol. 60, no. 2, pp. 585–599, 2012.

[7] T. Park and G. Casella, "The Bayesian lasso," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681–686, 2008.

[8] V. Rockova and E. Lesaffre, "Incorporating grouping information in Bayesian variable selection with applications in genomics," *Bayesian Analysis*, vol. 9, no. 1, pp. 221–258, 2014.

[9] W. C. Jakes and D. C. Cox, Eds., *Microwave Mobile Communications*, Wiley-IEEE Press, 1994.