GROUP-SPARSE ADAPTIVE VARIATIONAL BAYES ESTIMATION

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

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

ABSTRACT

This paper presents a new variational Bayes algorithm for the adaptive estimation of signals possessing group structured sparsity. The proposed algorithm can be considered as an extension of a recently proposed variational Bayes framework of adaptive algorithms that utilize heavy tailed priors (such as the Student-t distribution) to impose sparsity. Variational inference is efficiently implemented via appropriate time recursive equations for all model parameters. Experimental results are provided that demonstrate the improved estimation performance of the proposed adaptive group sparse variational Bayes method, when compared to state-of-the-art sparse adaptive algorithms.

***Index Terms*—** variational Bayes, structured sparsity, adaptive estimation, group sparse Bayesian learning

1. INTRODUCTION

In recent years, compressive sensing (CS) theory has received considerable attention in the signal processing literature. CS is mainly concerned with the estimation of sparse signals via the solution of an underdetermined linear system of equations. Recent developments in CS theory have also sparked new research interest in the *adaptive* estimation of sparse signals. This is an interesting challenge, as sophisticated, fast adaptive algorithms are needed in various applications to perform sparse system identification in an online fashion. A number of deterministic algorithms have been recently proposed to address this task, e.g., [1, 2]. These algorithms can be considered as regularized variants of the recursive least squares (RLS) algorithm and utilize the well-documented ℓ_1 norm to promote sparsity.

It is also interesting to note that sparsity in natural and man-made signals and systems often comes with some form of structure, in the sense that nonzero signal coefficients are usually clustered together. This observation has led to some first attempts to exploit group sparsity in order to achieve better estimation performance. A typical example is the recursive $\ell_{1,\infty}$ group lasso, recently proposed in [3]. As its name suggests, the recursive $\ell_{1,\infty}$ group lasso utilizes the LS cost function penalized by the $\ell_{1,\infty}$ norm to impose group sparsity. In the same fashion, an RLS variant is proposed in [4], where the $\ell_{p,0}$ norm is used instead. However, both these algorithms are of deterministic nature and their estimation performance largely depends on parameter fine tuning.

In this paper we exploit the Bayesian framework to address the problem of group sparse system identification. In this context, we describe an adaptive variational Bayes algorithm that is specifically tailored to perform inference for group sparse, time varying signals. In our modeling, signal coefficients are assumed to be clustered in equally sized groups. To promote group sparsity, the proposed scheme utilizes a hierarchical Bayesian model based on the heavy-tailed multivariate Student-t distribution. Variational inference on the proposed model is then presented, both in batch and adaptive mode, by deriving appropriate recursive update equations. The proposed scheme can also be considered as an extension to the case of group sparse signals of a probabilistic family of adaptive algorithms recently presented in [5, 6]. Experimental results show that the new algorithm succeeds in exploiting group sparsity very effectively and exhibits better estimation performance in comparison to other related state-of-the-art sparse adaptive schemes.

2. PROBLEM FORMULATION

Let $\mathbf{w}(n) = [w_1(n), w_2(n), \ldots, w_N(n)]^T$ be an N - dimensional unknown signal vector that may be varying in time. We assume that $\mathbf{w}(n)$ has *structured sparsity*, in the sense that it has $\xi \ll N$ nonzero elements, which are aligned in equally sized groups within the N-dimensional vector, as explained in the next section. Let $\mathbf{y}(n) = [y(1), y(2), \ldots, y(n)]^T$ be a n-dimensional vector of observations at time instant n and $\mathbf{X}(n)$ denote a $n \times N$ *known* data matrix, i.e., $\mathbf{X}(n) = [\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(n)]^T$. The relationship between $\mathbf{y}(n)$, $\mathbf{X}(n)$ and $\mathbf{w}(n)$ is assumed to be linear, following the regression model

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

where $\boldsymbol{\epsilon}(n)$ is a n-dimensional vector of additive zero mean uncorrelated Gaussian noise. By utilizing the set of observations and data, $\{\mathbf{y}(n), \mathbf{X}(n)\}$, the signal vector $\mathbf{w}(n)$ can be

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.

sequentially estimated in time. To this end, the recursive least squares (RLS) algorithm solves the following minimization task over time,

$$\min_{\hat{\mathbf{w}}(n)} \sum_{k=1}^{n} \lambda^{n-k} |y(k) - \mathbf{x}^T(k)\hat{\mathbf{w}}(n)|^2, \tag{2}$$

where $0 \ll \lambda < 1$ is the forgetting factor. In vector notation, the previous LS cost function can be expressed as,

$$\min_{\hat{\mathbf{w}}(n)} \|\mathbf{\Lambda}^{1/2}(n)\mathbf{y}(n) - \mathbf{\Lambda}^{1/2}(n)\mathbf{X}(n)\hat{\mathbf{w}}(n)\|^2, \tag{3}$$

where $\mathbf{\Lambda}(n) = \mathrm{diag}([\lambda^{n-1}, \lambda^{n-2}, \dots, 1]^T)$.

In this paper we are interested in developing a time adaptive estimator for the group sparse weight vector $\mathbf{w}(n)$. Following the Bayesian approach, (c.f. [5]), we employ a hierarchical Bayesian model emanating from the sparse Bayesian learning framework, [7], and we establish an adaptive variational Bayes scheme that performs online approximate inference for all model parameters.

3. BAYESIAN MODELING

We consider first the Bayesian modeling of the batch optimization problem first and, to simplify notation, we temporarily drop the time index n from all model parameters. Time indexing is restored in Section 5, where the adaptive variational Bayes scheme is presented.

Accounting for the presence of the forgetting factor λ in (3), the additive noise in (1) is assumed to be distributed as $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \beta^{-1}\mathbf{\Lambda}^{-1})$. This gives rise to the following likelihood function

$$p(\mathbf{y}|\mathbf{w}, \beta) = \frac{\beta^{\frac{n}{2}}|\mathbf{\Lambda}|^{\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}} \exp\left[-\frac{\beta}{2}\|\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{y} - \mathbf{\Lambda}^{\frac{1}{2}}\mathbf{X}\mathbf{w}\|^2\right], \tag{4}$$

where $|\cdot|$ stands for matrix determinant. Notice that the maximum likelihood (ML) estimate of $\mathbf{w}$ based on (4) is exactly the LS solution obtained by solving (3). To further perform inference on the model parameters β and $\mathbf{w}$ and to impose group sparsity on $\mathbf{w}$, we equip our Bayesian model with appropriate *conjugate* priors. First, for the precision parameter β we assume a conjugate Gamma prior that guarantees the positivity of this parameter, i.e.,

$$p(\beta; \rho, \delta) = \mathcal{G}(\beta; \rho, \delta) = \frac{\delta^\rho}{\Gamma(\rho)}\beta^{\rho-1}\exp\left[-\delta\beta\right]. \tag{5}$$

Next, we assume that $\mathbf{w}$ consists of M groups of D coefficients each, (i.e., $N = MD$), where M is known a priori. Let us consider the grouping $\mathbf{w} = [\mathbf{w}_1^T, \mathbf{w}_2^T, \dots, \mathbf{w}_M^T]^T$, where $\mathbf{w}_m$ is the $D \times 1$ weight component corresponding to the m-th block of $\mathbf{w}$. We assign an independent zero-mean Gaussian prior $\mathcal{N}(\mathbf{w}_m|\mathbf{0}, \beta^{-1}\alpha_m^{-1}\mathbf{I}_D)$ to each $\mathbf{w}_m$, i.e,

$$p(\mathbf{w}|\boldsymbol{\alpha}, \beta) = \prod_{m=1}^{M} (2\pi)^{-\frac{D}{2}} \beta^{\frac{D}{2}} \alpha_m^{\frac{D}{2}} \exp\left[-\frac{\beta}{2}\alpha_m\|\mathbf{w}_m\|^2\right], \tag{6}$$

with $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_M]^T$. To define a conjugate form of a multivariate Student-t prior for each group of coefficients $\mathbf{w}_m$, we select a Gamma distribution for the precision parameters α_m in the second level of hierarchy, i.e.,

$$p(\alpha_m) = \mathcal{G}(\alpha_m; c, \frac{a}{2}) = \frac{(\frac{a}{2})^c}{\Gamma(c)}\alpha_m^{c-1}\exp\left[-\frac{a}{2}\alpha_m\right]. \tag{7}$$

By integrating out the precision parameters $\boldsymbol{\alpha}$, it can be easily shown that the prior on $\mathbf{w}$ is expressed as,

$$p(\mathbf{w}) = \int p(\mathbf{w}|\boldsymbol{\alpha})p(\boldsymbol{\alpha}; c, \frac{a}{2})d\boldsymbol{\alpha} = \prod_{m=1}^{M} \mathcal{S}t_{2c}(\mathbf{0}, \frac{a}{2c\beta}\mathbf{I}_D), \tag{8}$$

where $\mathcal{S}t_\nu(\boldsymbol{\zeta}, \boldsymbol{\Upsilon})$ denotes the standard *multivariate* Student-t distribution with location $\boldsymbol{\zeta}$, scale matrix $\boldsymbol{\Upsilon}$ and ν degrees of freedom. Notice that the heavy tailed Student-t distribution is known to promote sparsity and has been widely used in the sparse Bayesian learning framework, [8]. Thus, our Bayesian model favors group sparsity on the weight vector $\mathbf{w}$, and it can be considered as a natural extension of the sparse Bayesian learning (SBL) model, [8], in the case of structured sparsity.

4. VARIATIONAL BAYESIAN INFERENCE

Due to the complexity of the proposed Bayesian model, the posterior of interest, $p(\mathbf{w}, \beta, \boldsymbol{\alpha}|\mathbf{y})$, cannot be explicitly computed. In this paper we rely on the variational methodology to perform Bayesian inference, extending the approach followed in [5,6] to a group sparsity scenario. In the variational framework it is common to use the mean field approximation for the posterior $p(\mathbf{w}, \beta, \boldsymbol{\alpha}|\mathbf{y})$, utilizing a distribution $q(\mathbf{w}, \beta, \boldsymbol{\alpha})$ of the factorized form

$$q(\mathbf{w}, \beta, \boldsymbol{\alpha}) = q(\beta)\prod_{m=1}^{M} q(\mathbf{w}_m)\prod_{m=1}^{M} q(\alpha_m). \tag{9}$$

Then, our goal is to minimize the Kullback-Leibler distance between the true posterior $p(\mathbf{w}, \beta, \boldsymbol{\alpha}|\mathbf{y})$ and the approximating distribution $q(\mathbf{w}, \beta, \boldsymbol{\alpha})$. For this task, the assumed posterior independence between the grouped model parameters makes the approximating factors in the right hand side of (9) tractable. Let $\boldsymbol{\theta}$ be the vector containing all model parameters, i.e., $\boldsymbol{\theta} = [\mathbf{w}_1^T, \dots, \mathbf{w}_M^T, \beta, \alpha_1, \dots, \alpha_M]^T$, and θ_i denote either a $\mathbf{w}_j^T$, or a α_j, $j = 1, \dots, M$, or β. Then, it is known from the variational Bayes theory, [9], that

$$q(\theta_i) = \frac{\exp\left(\mathbb{E}_{j\neq i}\left[\log p(\mathbf{y}, \boldsymbol{\theta})\right]\right)}{\int \exp\left(\mathbb{E}_{j\neq i}\left[\log p(\mathbf{y}, \boldsymbol{\theta})\right]\right)d\theta_i}, \tag{10}$$

where $\mathbb{E}_{j\neq i}[\cdot]$ denotes expectation w.r.t. all $q(\theta_j)$'s except for $q(\theta_i)$. Applying (10) for all parameters of interest, we get for each $\mathbf{w}_m$ that

$$q(\mathbf{w}_m) = \mathcal{N}(\mathbf{w}_m; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = (2\pi)^{-\frac{D}{2}}|\boldsymbol{\Sigma}_m|^{-\frac{1}{2}}$$

$$\exp\left[-\frac{1}{2}(\mathbf{w}_m-\boldsymbol{\mu}_m)^T\boldsymbol{\Sigma}_m^{-1}(\mathbf{w}_m-\boldsymbol{\mu}_m)\right], \quad (11)$$

where

$$\boldsymbol{\Sigma}_m = \langle\beta\rangle^{-1}(\mathbf{X}_m^T\boldsymbol{\Lambda}\mathbf{X}_m+\langle\alpha_m\rangle\mathbf{I}_D)^{-1}, \quad (12)$$

$$\boldsymbol{\mu}_m = \langle\beta\rangle\boldsymbol{\Sigma}_m\mathbf{X}_m^T\boldsymbol{\Lambda}(\mathbf{y}-\mathbf{X}_{\neg m}\boldsymbol{\mu}_{\neg m}), \quad (13)$$

where $\langle\cdot\rangle$ denotes expectation with respect to $q(\cdot)$. Letting the columns of $\mathbf{X}$ to be separated in M $n\times D$ groups according to the partitioning of $\mathbf{w}$, i.e., $\mathbf{X}=[\mathbf{X}_1,\mathbf{X}_2,\dots,\mathbf{X}_M]$, $\mathbf{X}_{\neg m}$, $\boldsymbol{\mu}_{\neg m}$ result from $\mathbf{X}$, $\mathbf{w}$ after removing $\mathbf{X}_m$, $\boldsymbol{\mu}_m$ respectively. For the approximating posterior of the noise precision β we get the conjugate posterior distribution

$$q(\beta)=\mathcal{G}(\beta;\tilde{\rho},\tilde{\delta}), \quad (14)$$

with $\tilde{\rho}=\frac{n+MD}{2}+\rho$ and

$$\tilde{\delta}=\delta+\frac{1}{2}\left\langle\|\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{y}-\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{X}\mathbf{w}\|^2\right\rangle+\frac{1}{2}\left\langle\sum_{m=1}^{M}\alpha_m\|\mathbf{w}_m\|^2\right\rangle. \quad (15)$$

Next, for the precision parameters of the groups $\mathbf{w}_m$'s we get the approximating Gamma distribution,

$$q(\alpha_m)=\mathcal{G}(\alpha_m;c+\frac{D}{2},\frac{\langle\beta\rangle\langle\|\mathbf{w}_m\|^2\rangle}{2}+\frac{a}{2}). \quad (16)$$

The means of the parameters of $\boldsymbol{\theta}$ are :
$\langle\mathbf{w}_m\rangle\equiv\boldsymbol{\mu}_m, m=1,\dots,M,$

$$\langle\beta\rangle=\frac{n+MD+2\rho}{2\delta+\left\langle\|\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{y}-\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{X}\mathbf{w}\|^2\right\rangle+\sum_{m=1}^{M}\langle\alpha_m\rangle\langle\|\mathbf{w}_m\|^2\rangle} \quad (17)$$

$$\langle\alpha_m\rangle=\frac{2c+D}{\langle\beta\rangle\langle\|\mathbf{w}_m\|^2\rangle+a}, m=1,\dots,M. \quad (18)$$

Clearly, these quantities are interrelated. This results to a cyclic updating mechanism, termed as variational Bayes algorithm, [9]. This algorithm progresses as follows: we initialize all parameter means first, and we update each mean in a cyclic manner, by keeping all the remaining fixed.

We observe from (11), (14) and (16) that some second order moments are required for the updating of the posterior means, i.e.,

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

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

Using these moments, the updating of the proposed variational Bayes scheme converges to a sparse estimate $\boldsymbol{\mu}=[\boldsymbol{\mu}_1^T,\boldsymbol{\mu}_2^T,\dots,\boldsymbol{\mu}_M^T]^T$ for the unknown signal vector $\mathbf{w}$ in a few iterations.

Initialize $\beta(0),\hat{\mathbf{w}}(0),\mathbf{A}(-1),\mathbf{A}(0),\mathbf{R}(0),\mathbf{z}(0),d(0)$
Set a,b,ρ,δ to very small values
for $n=1,2,\dots$
  $\mathbf{R}(n)=\lambda\mathbf{R}(n-1)+\mathbf{x}(n)\mathbf{x}^T(n)$
    $-\lambda\mathbf{A}(n-1)\otimes\mathbf{I}_D+\mathbf{A}(n)\otimes\mathbf{I}_D$
  $\mathbf{z}(n)=\lambda\mathbf{z}(n-1)+\mathbf{x}(n)y(n)$
  $d(n)=\lambda d(n-1)+y^2(n)$
  $\beta(n)=((1-\lambda)^{-1}+N+2\rho)/(2\delta+d(n)$
    $-\mathbf{z}^T(n)\hat{\mathbf{w}}(n-1)+\sum_{m=1}^{M}\mathrm{tr}[\frac{\mathbf{R}_m^{-1}(n-1)}{\beta(n-1)}\mathbf{R}_m(n)])$
  for $m=1,2,\dots,M$
    $\hat{\mathbf{w}}_m(n)=\mathbf{R}_m^{-1}(n)\left(\mathbf{z}_m(n)-\mathbf{R}_{m\neg m}(n)\hat{\mathbf{w}}_{\neg m}(n)\right)$
    $\alpha_m(n)=\frac{2c+D}{a+\beta(n)\|\hat{\mathbf{w}}_m(n)\|^2+\mathrm{tr}(\mathbf{R}_m^{-1}(n))}$
  end for
end for

Table 1. The AGSVB-S algorithm

5. ADAPTIVE VARIATIONAL GROUP SBL

In the previous section we have described a variational Bayes scheme based on the Bayesian modeling of Section 3 that performs inference for the time invariant signal vector $\mathbf{w}$ in a batch mode. Let us now restore time indexing and extend the variational Bayes scheme in an adaptive setting, where the weight vector $\mathbf{w}(n)$ is now time varying. To facilitate computations we define the following time dependent quantities,

$$\mathbf{R}(n)=\mathbf{X}^T(n)\boldsymbol{\Lambda}(n)\mathbf{X}(n)+\mathbf{A}(n-1)\otimes\mathbf{I}_D, \quad (21)$$

$$\mathbf{z}(n)=\mathbf{X}^T(n)\boldsymbol{\Lambda}(n)\mathbf{y}(n), \quad (22)$$

$$d(n)=\mathbf{y}^T(n)\boldsymbol{\Lambda}(n)\mathbf{y}(n), \quad (23)$$

where $\mathbf{A}(n)=\mathrm{diag}(\boldsymbol{\alpha}(n))$ and $\otimes$ denotes the Kronecker product. These quantities can be updated recursively, i.e.,

$$\mathbf{R}(n)=\lambda\mathbf{R}(n-1)+\mathbf{x}(n)\mathbf{x}^T(n) - \lambda\mathbf{A}(n-2)\otimes\mathbf{I}_D+\mathbf{A}(n-1)\otimes\mathbf{I}_D \quad (24)$$

$$\mathbf{z}(n)=\lambda\mathbf{z}(n-1)+\mathbf{x}(n)y(n), \quad (25)$$

$$d(n)=\lambda d(n-1)+y^2(n). \quad (26)$$

We can identify $\mathbf{R}(n)$ as the sample auto-correlation matrix of $\mathbf{x}(n)$ regularized by the diagonal matrix $\mathbf{A}(n-1)\otimes\mathbf{I}_D$, $\mathbf{z}(n)$ as the sample cross-correlation vector between $\mathbf{x}(n)$ and $y(n)$, and $d(n)$ as the energy of the observation vector $\mathbf{y}(n)$. Substituting (12) in (13) and utilizing (21) and (22), the adaptive weights $\hat{\mathbf{w}}_m(n)(=\boldsymbol{\mu}_m(n))$ can be efficiently time updated as follows [1]

$$\hat{\mathbf{w}}_m(n)=\mathbf{R}_m^{-1}(n)\left(\mathbf{z}_m(n)-\mathbf{R}_{m\neg m}(n)\hat{\mathbf{w}}_{\neg m}(n)\right), \quad (27)$$

where $\mathbf{z}_m$ is the m-th $D\times 1$ block of $\mathbf{z}$, $\mathbf{R}_m(n)$ is the m-th $D\times D$ diagonal block of $\mathbf{R}(n)$, $\mathbf{R}_{m\neg m}(n)$ is the $D\times(N-D)$

[1] It can be shown that (27) represents a block coordinate descent updating rule, [10].

matrix resulting from the m-th row block of $\mathbf{R}(n)$ after removing its m-th group of D columns, and $\hat{\mathbf{w}}_{\neg m}(n) = [\hat{\mathbf{w}}_1(n), \ldots, \hat{\mathbf{w}}_{m-1}(n), \hat{\mathbf{w}}_{m+1}(n-1), \ldots, \hat{\mathbf{w}}_M(n-1)]^T$. Moreover, based on [6], it can be shown that noise precision is efficiently computed as follows,

$$\beta(n) = ((1-\lambda)^{-1} + N + 2\rho)/(2\delta + d(n) - \mathbf{z}^T(n)\hat{\mathbf{w}}(n-1) + \sum_{m=1}^{M} \mathrm{tr}[\mathbf{\Sigma}_m(n-1)\mathbf{R}_m(n)]), \quad (28)$$

where $\mathbf{\Sigma}_m(n-1) = \beta^{-1}(n-1)\mathbf{R}_m^{-1}(n-1)$ according to (12). Finally, straightforward computations using (18), (13) and (19) yield that the updating of the precisions α_m is performed as

$$\alpha_m(n) = \frac{2c + D}{a + \beta(n)\|\hat{\mathbf{w}}_m(n)\|^2 + \mathrm{tr}(\mathbf{R}_m^{-1}(n))}. \quad (29)$$

The resulting algorithm is termed *Adaptive Group Sparse Variational Bayes based on a multivariate Student-t prior (AGSVB-S)* and is summarized in Table 1. To the best of our knowledge, AGSVB-S is the first group-sparsity promoting adaptive algorithm that originates from a Bayesian framework. As it will be shown in the next section, the proposed algorithm exploits structured sparsity very effectively and, at the cost of a slight increase in computational complexity, it offers the best estimation performance compared to related state-of-the-art techniques. Notably, this is achieved in a fully automated manner, by entirely alleviating the need for parameter cross-validation and fine-tuning.

6. EXPERIMENTAL RESULTS

In this section we consider the adaptive estimation of a *group sparse* multipath wireless channel. The AGSVB-S algorithm is compared with state-of-the-art adaptive algorithms, e.g., the time-norm weighted lasso (TNWL), [2], and the recently proposed adaptive sparse variational Bayes with a Student-t prior (ASVB-S), [6], in order to validate its performance. Additionally, the genie-aided RLS (GARLS), which operates only on the nonzero coefficients of the parameter vector $\mathbf{w}$, is also included as a benchmark.

In our experiments we consider a group sparse Rayleigh fading channel partitioned in $M = 16$ groups. Each group contains $D = 4$ coefficients for a total of 64 weight coefficients. Only $T = 3$ groups are assumed to be nonzero. The nonzero coefficients vary in time according to Jakes model, [11]. The input sequence is a random ± 1 sequence of length 1000 and the forgetting factor λ is set to 0.99. All the hyperparameters a, c, ρ and δ are set to 10^{-6}. Noise is assumed to be white Gaussian, and its variance is adjusted so as to get an SNR level of 12dB in all experiments. The normalized mean square error, defined as $\mathrm{NMSE} = \mathbb{E}\left[\|\mathbf{w} - \hat{\mathbf{w}}\|^2\right] / \mathbb{E}\left[\|\mathbf{w}\|^2\right]$, is used to assess the performance of the algorithms. All performance curves are ensemble averages of 200 transmission packets, channels, and noise realizations

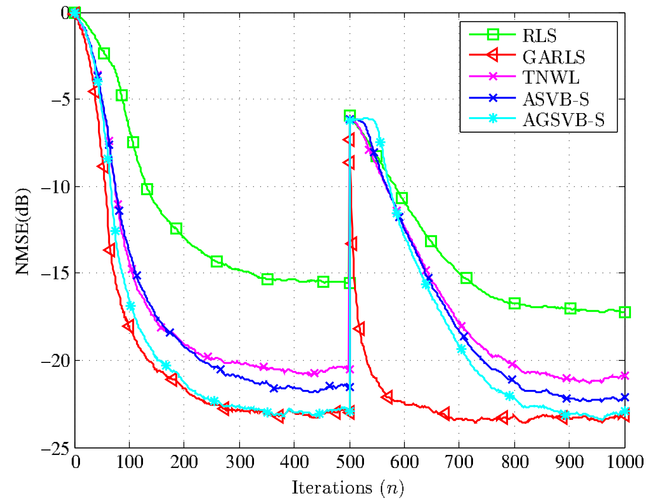

Fig. 1. NMSE curves under slow fading and a sudden channel change.

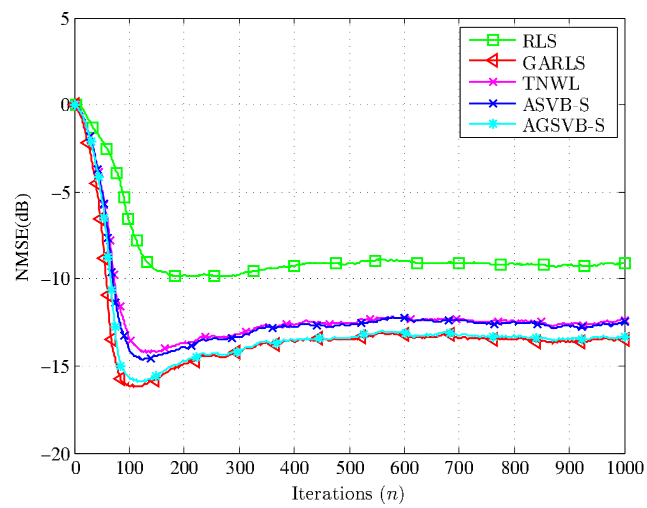

Fig. 2. NMSE curves in a fast fading scenario.

In the first experiment, we consider a Rayleigh fading channel with an abrupt change at time instant $n = 500$; an extra nonzero group is added to the channel coefficients. For the remaining time period, a slow fading environment is simulated by setting the normalized Doppler frequency to $f_d T_s = 5 \times 10^{-5}$. Fig. 1 shows the NMSE curves of the considered algorithms. It is observed that the performance of the proposed AGSVB-S before the abrupt change is near optimum, i.e., it converges as fast as the GARLS algorithm at approximately the same error floor. This performance improvement with regard to ASVB-S is theoretically expected and is justified by the fact that AGSVB-S exploits the sparsity structure of the

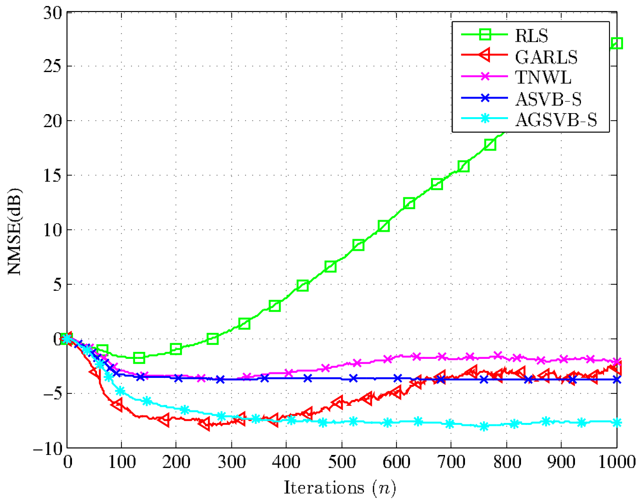

Fig. 3. NMSE curves for correlated input.

weight coefficients. In Fig. 1 it is also observed that AGSVB-S has the ability to track the channel's abrupt change, since after a sudden NMSE fluctuation, it converges faster that the other algorithms to the error floor of the GARLS, again.

In the second experiment, the tracking capability of the proposed AGSVB-S algorithm is explored, this time in a fast fading environment. In this setup, we set the normalized Doppler frequency to $f_d T_s = 8.35 \times 10^{-4}$ and the forgetting factor to $\lambda = 0.98$. The NMSE curves of the adaptive algorithms considered are shown in Fig. 2. It is easy to verify that AGSVB-S has, again, the overall best performance. Its convergence speed is similar to that of GARLS, and it reaches the same best achievable error floor after convergence.

In the next experiment we evaluate the performance of the algorithms in the case of correlated input. To this end, we lowpass filter a Gaussian sequence of zero mean and unit variance in order to generate a colored input sequence. In our experiments, a 5th order Butterworth filter is used, with a cut-off frequency $1/4$ the sampling rate. The remaining settings are the same as in the first experiment. In Fig. 3 the resulting NMSE curves for all adaptive algorithms are depicted. In comparison to Fig. 1, a considerable degradation in terms of NMSE performance is observed, owning to the worse conditioning of the autocorrelation matrix $\mathbf{R}(n)$. Interestingly, both RLS and GARLS seem to diverge. This is not surprising, since, RLS is known to be sensitive to input signal correlation, as noted in [12]. On the contrary, the proposed AGSVB-S has now a slower convergence rate but achieves the lowest estimation error among the considered algorithms.

7. CONCLUSION

We have presented an adaptive variational group sparse Bayesian learning algorithm. The proposed scheme is fully automated but requires a priori knowledge of the structure of the signal sparsity, as is common in most group sparse schemes. Experimental results have demonstrated the robustness of the proposed variational scheme under different circumstances. A further development of the proposed scheme that is currently under investigation is the incorporation of the group length as a parameter in our model, perhaps via an appropriate discrete prior distribution.

REFERENCES

[1] B. Babadi, N. Kalouptsidis, and V. Tarokh, "SPARLS: The sparse RLS algorithm," *Signal Processing, IEEE Transactions on*, vol. 58, pp. 4013 –4025, Aug. 2010.

[2] D. Angelosante, J.A. Bazerque, and G.B. Giannakis, "Online adaptive estimation of sparse signals: Where RLS meets the ℓ_1-norm," *Signal Processing, IEEE Transactions on*, vol. 58, pp. 3436 –3447, July 2010.

[3] 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.

[4] E. M. Eksioglu, "Group sparse RLS algorithms," *International Journal of Adaptive Control and Signal Processing*, pp. 1099–1115, 2013.

[5] K.E. Themelis, A.A. Rontogiannis, and K.D. Koutroumbas, "Variational Bayesian sparse adaptive filtering using a Gauss-Seidel recursive approach," in *EUSIPCO*, Sep. 2013.

[6] K.E. Themelis, A.A. Rontogiannis, and K.D. Koutroumbas, "A variational Bayes framework for sparse adaptive estimation," arXiv, 1401.2771, 2014.

[7] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[8] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.

[9] T. J. Jaakkola and M. I. Jordan, "Bayesian parameter estimation via variational methods," *Statistics and Computing*, vol. 10, no. 1, pp. 25–37, 2000.

[10] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *Journal of Optimization Theory and Applications*, vol. 109, no. 3, pp. 475–494, 2001.

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

[12] S.H. Ardalan and T. Adali, "Sensitivity analysis of transversal RLS algorithms with correlated inputs," in *Circuits and Systems, 1989., IEEE International Symposium on*, May 1989, pp. 1744–1747 vol.3.