A Novel Hierarchical Bayesian Approach for Sparse Semisupervised Hyperspectral Unmixing

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

***Abstract*—In this paper the problem of semisupervised hyperspectral unmixing is considered. More specifically, the unmixing process is formulated as a linear regression problem, where the abundance's physical constraints are taken into account. Based on this formulation, a novel hierarchical Bayesian model is proposed and suitable priors are selected for the model parameters such that, on the one hand, they ensure the nonnegativity of the abundances, while on the other hand they favor sparse solutions for the abundances' vector. Performing Bayesian inference based on the proposed hierarchical Bayesian model, a new low-complexity iterative method is derived, and its connection with Gibbs sampling and variational Bayesian inference is highlighted. Experimental results on both synthetic and real hyperspectral data illustrate that the proposed method converges fast, favors sparsity in the abundances' vector, and offers improved estimation accuracy compared to other related methods.**

***Index Terms*—Compressive sensing, constrained optimization, constrained sparse regression, hierarchical Bayesian analysis, hyperspectral imagery, sparse semisupervised unmixing.**

I. Introduction

HYPERSPECTRAL remote sensing has gained considerable attention in recent years, due to its wide range of applications, e.g., environmental monitoring and terrain classification [1]–[3] and the maturation of the required technology. Hyperspectral sensors are able to sample the electromagnetic spectrum in tens or hundreds of contiguous spectral bands from the visible to the near-infrared region. However, due to their low spatial resolution, more than one different materials can be mixed in a single pixel, which calls for spectral unmixing, [3]. In spectral unmixing, the measured spectrum of a mixed pixel is decomposed into a collection of constituent spectra, called *endmembers* and a set of corresponding fractions, called *abundances*, that indicate the percentage contribution of each endmember to the formation of the pixel.

Manuscript received December 07, 2010; revised May 26, 2011 and September 23, 2011; accepted October 12, 2011. Date of publication October 28, 2011; date of current version January 13, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mark Coates.

K. E. Themelis is with the Department of Informatics and Telecommunications, University of Athens, Ilissia, 157 84 Athens, Greece. He is also with the Institute for Space Applications and Remote Sensing, National Observatory of Athens, 152 36, P. Penteli, Greece (e-mail: themelis@noa.gr).

A. A. Rontogiannis and K. D. Koutroumbas are with the Institute for Space Applications and Remote Sensing, National Observatory of Athens, 152 36, P. Penteli, Greece (e-mail: 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.2011.2174052

The process of hyperspectral unmixing is described by two major steps: (a) the endmember extraction step, and (b) the inversion process. In the endmember extraction step the spectral signatures of the endmembers contributing to the hyperspectral image are determined. Popular endmember extraction algorithms include the pixel purity index (PPI), [4], the N-FINDR algorithm, [5] and the vertex component analysis (VCA) method, [6]. The inversion process determines the abundances corresponding to the estimated endmembers obtained in the previous step. The abundances should satisfy two constraints, in order to remain physically meaningful; they should be nonnegative and sum to one. Under these constraints, spectral unmixing is formulated as a convex optimization problem, which can be addressed using iterative methods, e.g., the fully constrained least squares method, [7], or numerical optimization methods, e.g., [8]. Bayesian methods have also been proposed for the problem, e.g., the Gibbs sampling scheme applied to the hierarchical Bayesian model of [9]. Semisupervised unmixing, [9], [10], which is considered in this paper, assumes that the endmembers' spectral signatures are available. The objective of semisupervised unmixing is to determine how many and which endmembers are present in the mixed pixel under study and to estimate their corresponding abundances.

An interesting perspective of the semisupervised spectral unmixing problem arises when the latent sparsity of the abundance vector is taken into account. A reasonable assumption is that only a small number of endmembers are mixed in a single pixel, and hence, the solution to the endmember determination and abundance estimation problem is inherently sparse. This lays the ground for the utilization of sparse signal representation techniques, e.g., [11]–[14], in semisupervised unmixing. A number of such semisupervised unmixing techniques has been recently proposed in [10], [15], and [16], based on the concept of ℓ_1 norm penalization to enhance sparsity. These methods assume that the spectral signatures of many different materials are available, in the form of a spectral library. Since only a small number of the available materials' spectra are expected to be present in the hyperspectral image, the abundance vector is expected to be sparse.

In this paper, a novel hierarchical Bayesian approach for semisupervised hyperspectral unmixing is presented, which is based on the sparsity hypothesis and the nonnegativity property of the abundances. In the proposed hierarchical model, appropriate prior distributions are assigned to the unknown parameters, which reflect prior knowledge about their natural characteristics. More specifically, to account for the nonnegativity of the abundances, a truncated nonnegative Gaussian distribution is used as a first level prior. The variance param-

eters of this distribution are then selected to be exponentially distributed. This two-level hierarchical prior formulates a Laplace type prior for the abundances, which is known to promote sparsity, [17], [18]. In addition, compared to other related hierarchical models, [14], [19], [20], which employ a single sparsity-controlling hyperparameter, the proposed model comprises multiple distinct sparsity-controlling hyperparameters. It is proven that this extension makes the model equivalent to a nonnegativity constrained variant of the adaptive least absolute shrinkage and selection operator (Lasso) criterion of [21], whose solution provides a consistent abundance estimator. The proposed hierarchical model also retains the conjugacy of the parameter distributions, which in the sequel is exploited to obtain closed form expressions for the parameters' posterior distributions.

As is usually the case in Bayesian analysis, the resulting joint posterior distribution of the proposed hierarchical model does not possess a tractable analytical form. To overcome this impediment, a novel iterative algorithm is developed, which can be considered as a deterministic approximation of the Gibbs sampler [22]. In this algorithmic scheme, the conditional posterior distributions of the model parameters are derived and their respective expectations are selected to replace the random samples used by the Gibbs sampler. More specifically, as far as the abundance vector is concerned, an efficient scheme is developed to update its posterior conditional expectation, while the conditional expectations of all remaining parameters are updated through simple, closed form expressions. The proposed Bayesian inference algorithm iterates through the derived conditional expectations, updating each one of them based on the current estimates of the remaining ones. To put the algorithm to its proper setting, its connection to other Bayesian inference methods, [23]–[26], is discussed. In particular, emphasis is given to show the affinity of the proposed algorithm with a variational Bayesian inference scheme, which is based on a suitable factorization of the corresponding variational posterior distribution.

Interestingly, the proposed algorithm produces a point estimate of the abundance vector, which is sparse and satisfies the nonnegativity constraint. As a by-product, estimates of all other parameters involved in the problem are also naturally produced; among them is the variance of the additive noise, which is assumed to corrupt the hyperspectral image. The proposed algorithm is computationally efficient and, as verified by extensive simulations, it converges very fast to the true model parameters. In addition, it offers enhanced estimation performance, as corroborated by the application of the proposed and other related methods for the unmixing of both simulated and real hyperspectral data.

The remaining of the paper is organized as follows. The sparse semisupervised hyperspectral unmixing problem is formulated in Section II. Section III describes the proposed hierarchical Bayesian model. In Section IV, the new iterative conditional expectations algorithm used to perform Bayesian inference is presented and analyzed. Simulation results both on artificial and real hyperspectral data are reported in Section V. Conclusions are provided in Section VI. Finally, the connection of the proposed algorithm to variational Bayesian inference and other methods is highlighted in Appendix E.

Notation: We use lowercase boldface and uppercase boldface letters to represent vectors and matrices, respectively. With $(\cdot)^T$ we denote transposition, and with $\|\cdot\|_1$ and $\|\cdot\|_2$ the ℓ_1 and ℓ_2 norm, respectively, ($\|\mathbf{x}\|_1 = \sum_{i=1}^{N} |x_i|$, $\|\mathbf{x}\|_2^2 = \mathbf{x}^T\mathbf{x}$). The determinant of a matrix or the absolute value of a scalar is denoted by $|\cdot|$, while $\mathrm{diag}(\mathbf{x})$ stands for a diagonal matrix, that contains the elements of vector $\mathbf{x}$ on its diagonal. Finally, $\mathcal{R}^N$ is the N-dimensional Euclidean space, $\mathbf{0}$ denotes the zero vector, $\mathbf{1}$ the all-ones vector, and $\mathbf{I}_K$ is the $K \times K$ identity matrix.

II. Problem Formulation

In this section, we provide definitions and formulate rigorously the sparse semisupervised unmixing problem. Let $\mathbf{y}$ be a $M \times 1$ hyperspectral image pixel vector, where M is the number of spectral bands. Also let $\mathbf{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_N]$ stand for the $M \times N$ signature matrix of the problem, with $M > N$, where the $M \times 1$ dimensional vector $\boldsymbol{\phi}_i$ represents the spectral signature (i.e., the reflectance values in all spectral bands) of the ith endmember and N is the total number of distinct endmembers. Finally, let $\mathbf{w} = [w_1, w_2, \ldots, w_N]^T$ be the $N \times 1$ abundance vector associated with $\mathbf{y}$, where w_i denotes the abundance fraction of $\boldsymbol{\phi}_i$ in $\mathbf{y}$.

In this work, the linear mixture model (LMM) is adopted, that is, the previous quantities are assumed to be interrelated as follows

$$\mathbf{y} = \mathbf{\Phi}\mathbf{w} + \mathbf{n}. \tag{1}$$

The additive noise $\mathbf{n}$ is assumed to be a zero-mean Gaussian distributed random vector, with independent and identically distributed (i.i.d.) elements, i.e., $\mathbf{n}|\beta \sim \mathcal{N}(\mathbf{n}|\mathbf{0}, \beta^{-1}\mathbf{I}_M)$, where β denotes the inverse of the noise variance (precision). Due to the nature of the problem, the abundance vector is usually assumed to satisfy the following two constraints

$$w_i \geq 0,\ i = 1, 2, \ldots, N, \text{ and } \sum_{i=1}^{N} w_i = 1, \tag{2}$$

namely, a nonnegativity constraint and a sum-to-one (additivity) constraint. Based on this formulation, a semisupervised hyperspectral unmixing technique is introduced, where the endmember matrix $\mathbf{\Phi}$ is assumed to be known a priori. As mentioned before, each column of $\mathbf{\Phi}$ contains the spectral signature of a single material, and its elements are nonnegative, since they represent reflectance values. The mixing matrix $\mathbf{\Phi}$ can either stem from a spectral library or it can be determined using an endmember extraction technique, e.g., [6]. However, the actual number of endmembers that compose a single pixel's spectrum, denoted as ξ, is unknown and may vary from pixel to pixel. Sparsity is introduced when $\xi \ll N$, that is by assuming that only few of the available endmembers are present in a single pixel. This is a reasonable assumption, that is in line with intuition, since it is likely for a pixel to comprise only a few different materials from a library of several available materials. Summarizing, in semisupervised unmixing, we are interested in estimating the abundance vector $\mathbf{w}$ for each image pixel,

which is nonnegative and sparse, with ξ out of its N entries being nonzero.

This problem can be solved using either one of the recently proposed compressive sensing techniques, e.g., [11], [13], [14], [19], that focus only on the sparsity issue, or quadratic programming techniques, e.g., [8], that successfully enforce the constraints given in (2), but do not exploit sparsity. In the following, a hierarchical Bayesian model is presented, that both (a) favors sparsity and (b) takes into account the nonnegativity constraint of the problem. Then, a novel algorithm that is suitable to perform Bayesian inference for this model is derived. Moreover, it is shown that by a simple modification of the initial problem, the additivity constraint could also be naturally embedded.

III. Hierarchical Bayesian Model

This section introduces a novel hierarchical Bayesian model to estimate the sparse abundance vector $\mathbf{w}$ from (1), subject to the nonnegativity constraint given in (2). In a Bayesian framework, all unknown quantities are assumed to be random variables, each one described by a prior distribution, which models our knowledge about its nature. Before we proceed, the definition of a truncated multivariate distribution is provided, which will be frequently used in the sequel to follow.

Definition 1: Let $\mathbf{R}^N$ be a subset of $\mathcal{R}^N$ $\left(\mathbf{R}^N \subseteq \mathcal{R}^N\right)$ with positive Lebesgue measure, $\mathcal{P}(\cdot|\boldsymbol{\zeta})$ a N-variate distribution, where $\boldsymbol{\zeta}$ is a vector of parameters, and $\mathcal{P}_{\mathbf{R}^N}(\cdot|\boldsymbol{\zeta})$ the truncated probability density function (pdf) resulting from the truncation of $\mathcal{P}(\cdot|\boldsymbol{\zeta})$ on $\mathbf{R}^N$. Then, $\mathbf{x} \sim \mathcal{P}_{\mathbf{R}^N}(\mathbf{x}|\boldsymbol{\zeta})$ denotes a random vector, whose pdf is *proportional* to $\mathcal{P}(\mathbf{x}|\boldsymbol{\zeta})\mathcal{I}_{\mathbf{R}^N}(\mathbf{x})$, where $\mathcal{I}_{\mathbf{R}^N}(\cdot)$ is the indicator function defined as,

$$\mathcal{I}_{\mathbf{R}^N}(\mathbf{x}) = \begin{cases} 1, & \mathbf{x} \in \mathbf{R}^N \\ 0, & \mathbf{x} \notin \mathbf{R}^N. \end{cases} \tag{3}$$

A. Likelihood

Considering the observation model defined in (1) and the Gaussian property of the additive noise, the likelihood function of $\mathbf{y}$ can be expressed as follows:

$$\begin{aligned} p(\mathbf{y}|\mathbf{w},\beta) &= \mathcal{N}(\mathbf{y}|\boldsymbol{\Phi}\mathbf{w}, \beta^{-1}\mathbf{I}_M) \\ &= (2\pi)^{-\frac{M}{2}}\beta^{\frac{M}{2}}\exp\left[-\frac{\beta}{2}||\mathbf{y}-\boldsymbol{\Phi}\mathbf{w}||_2^2\right]. \end{aligned} \tag{4}$$

B. Parameter Prior Distributions

The Bayesian formulation requires that both the *sparsity* and *nonnegativity* properties of $\mathbf{w}$ should emanate from a suitably selected prior distribution. A widely used prior that favors sparsity, [14], [17], [19], [20], [27], is the zero-mean Laplace probability density function, which, for a single w_i, is defined as

$$\mathcal{L}(w_i|\lambda) = \frac{\lambda}{2}\exp\left[-\lambda|w_i|\right], \tag{5}$$

where λ is the inverse of the Laplace distribution shape parameter, $\lambda \geq 0$. Assuming prior independence of the individual coefficients w_i's, the N-dimensional prior over $\mathbf{w}$ can be written as

$$\mathcal{L}(\mathbf{w}|\lambda) = \prod_{i=1}^{N}\mathcal{L}(w_i|\lambda) = \left(\frac{\lambda}{2}\right)^N \exp\left[-\lambda\left\|\mathbf{w}\right\|_1\right]. \tag{6}$$

It can be easily shown, [17], that under the Laplace prior, the maximum a posteriori (MAP) estimate of $\mathbf{w}$ is given by

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}}\left\{\frac{\beta}{2}\left\|\mathbf{y}-\boldsymbol{\Phi}\mathbf{w}\right\|_2^2 + \lambda\left\|\mathbf{w}\right\|_1\right\}, \tag{7}$$

which is, surprisingly enough, the solution of the Lasso criterion of [28]. However, if the Laplace prior was applied to the sparse vector $\mathbf{w}$ directly, conjugacy[1] would not be satisfied with respect to the Gaussian likelihood given in (4) and hence, the posterior probability density function of $\mathbf{w}$ could not be derived in closed form. As noted in [29], a key property of the Laplace distribution is that it can be expressed as a scaled mixture of normals, with an exponential mixing density, i.e.,

$$\begin{aligned} &\frac{\lambda}{2}\exp\left[-\lambda\left|w_i\right|\right] \\ &= \int_0^{+\infty}\frac{1}{\sqrt{2\pi s}}\exp\left[-\frac{w_i^2}{2s}\right]\frac{\lambda^2}{2}\exp\left[-\frac{\lambda^2 s}{2}\right]ds,\ \lambda > 0, \end{aligned} \tag{8}$$

In the framework of the problem at hand, (8) suggests that the Laplace prior is equivalent to a two-level hierarchical Bayesian model, where the vector of abundances $\mathbf{w}$ follows a Gaussian distribution (first level), with exponentially distributed variances (second level). This hierarchical Bayesian model, which is a type of a Gaussian scale mixture (GSM), [30], has been adopted in [14], [17], [19], [20], [27], [31]. The main advantage of this formulation is that it maintains the conjugacy of the involved parameters.

In this paper, a slightly different Bayesian model is developed. More specifically, in order to satisfy the nonnegativity constraint of the abundance vector $\mathbf{w}$, the proposed hierarchical Bayesian approach uses a *truncated* normal distribution[2] in the nonnegative orthant of $\mathcal{R}^N$ as a first-level prior for $\mathbf{w}$. Assuming that all w_i's are i.i.d. and γ_i's are the (normalized by β) variances of w_i's, the prior assigned to $\mathbf{w}$ is expressed as (see Appendix A)

$$p(\mathbf{w}|\boldsymbol{\gamma},\beta) = \mathcal{N}_{\mathbf{R}_+^N}(\mathbf{w}|0,\beta^{-1}\boldsymbol{\Lambda}^{-1}). \tag{9}$$

$\mathbf{R}_+^N$ is the nonnegative orthant of $\mathcal{R}^N$, $\mathcal{N}_{\mathbf{R}_+^N}(\cdot)$ stands for the N-variate truncated normal distribution in $\mathbf{R}_+^N$ according to Definition 1, and $\boldsymbol{\Lambda}$ is the $N \times N$ diagonal matrix with $\boldsymbol{\Lambda}^{-1} = \mathrm{diag}(\boldsymbol{\gamma})$, where $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \ldots, \gamma_N]^T$. Note that the use of β as a normalization parameter in (9), ensures the unimodality of the posterior distribution of $\mathbf{w}$, [20], [31].

For the second parameter, β, appearing in the likelihood function (4), a Gamma prior distribution is assumed, defined as

$$p(\beta|\kappa,\theta) = \Gamma(\beta|\kappa,\theta) = \frac{\theta^\kappa}{\Gamma(\kappa)}\beta^{\kappa-1}\exp\left[-\theta\beta\right], \tag{10}$$

where $\beta \geq 0$, κ is the shape parameter, $\kappa \geq 0$, and θ is the inverse of the scale parameter of the Gamma distribution, $\theta \geq 0$. The mean and variance of the Gamma distribution are $\mathrm{E}[p(\beta|\kappa,\theta)] = \frac{\kappa}{\theta}$ and $var[p(\beta|\kappa,\theta)] = \frac{\kappa}{\theta^2}$, respectively.

[1] In Bayesian probability theory, if the posterior $p(\theta|x)$ belongs to the same distribution family with the prior $p(\theta)$, (for instance if they are both Gaussians), the prior and posterior are then called conjugate distributions.

[2] Note that the truncation of the normal distribution preserves conjugacy.

C. Hyperparameters' Priors

Having defined the truncated Gaussian distribution for w_i's, we focus now on the definition of the exponential distributions for γ_i's, in the spirit of (8). Before we describe the model for the priors of the hyperparameters γ_i's proposed in this work, let us first describe the model adopted in [17], [19]. There, the following exponential priors on γ_i are used

$$p(\gamma_i|\lambda) = \Gamma\left(\gamma_i|1, \frac{\lambda}{2}\right) = \frac{\lambda}{2}\exp\left[-\frac{\lambda}{2}\gamma_i\right],\ i = 1, 2, \ldots, N, \tag{11}$$

where λ is a hyperparameter, which controls the level of sparsity, $\lambda \geq 0$. If these priors were used for the elements of $\boldsymbol{\gamma}$ in (9), the prior distribution of $\mathbf{w}$ would be given as follows

$$\begin{aligned} p(\mathbf{w}|\lambda, \beta) &= \int p(\mathbf{w}|\boldsymbol{\gamma}, \beta) p(\boldsymbol{\gamma}|\lambda)\, d\boldsymbol{\gamma} \\ &= \prod_{i=1}^{N} \int_0^{\infty} p(w_i|\gamma_i, \beta) p(\gamma_i|\lambda) d\gamma_i \\ &= (\beta\lambda)^{\frac{N}{2}} \exp\left[-\sqrt{\beta\lambda}\sum_{i=1}^{N}|w_i|\right] \mathcal{I}_{\mathbf{R}_+^N}(\mathbf{w}) \\ &= \mathcal{L}\left(\mathbf{w}|\sqrt{\beta\lambda}\right) \mathcal{I}_{\mathbf{R}_+^N}(\mathbf{w}). \end{aligned} \tag{12}$$

With respect to Definition 1, $\mathcal{L}\left(\mathbf{w}|\sqrt{\beta\lambda}\right)\mathcal{I}_{\mathbf{R}_+^N}(\mathbf{w})$ is denoted as $\mathcal{L}_{\mathbf{R}_+^N}(\mathbf{w}|\sqrt{\beta\lambda})$ and is a truncated Laplace distribution on $\mathbf{R}_+^N$. We have already pointed out the relationship between the Laplace density, shown in (6), and the Lasso criterion (7). In a similar way, it can be easily shown that under the truncated Laplace prior given in (12), the MAP estimator of $\mathbf{w}$ would be the solution of a nonnegativity constrained Lasso criterion. Moreover, from a Lasso point of view, [28], it is known that as λ increases, sparser solutions arise for $\mathbf{w}$.

After the previous parenthesis, we proceed with the description of the model for γ_i's proposed in this work. The latter is an extension of that given in (11), where instead of having a single λ for all γ_i's, a distinct λ_i is associated with each γ_i (the motivation for such a choice will become clear in the analysis to follow). Thus, in the second stage of our hierarchical model, N independent Gamma priors are assigned to the elements of $\boldsymbol{\gamma}$, each parameterized by a distinct λ_i, as follows

$$p(\gamma_i|\lambda_i) = \Gamma\left(\gamma_i|1, \frac{\lambda_i}{2}\right) = \frac{\lambda_i}{2}\exp\left[-\frac{\lambda_i}{2}\gamma_i\right],\ i = 1, 2, \ldots, N, \tag{13}$$

where $\lambda_i \geq 0,\ i = 1, 2, \ldots, N$. By assuming that all γ_i's are independent, the joint distribution of $\boldsymbol{\gamma}$ can now be written as

$$\begin{aligned} p(\boldsymbol{\gamma}|\boldsymbol{\lambda}) &= \prod_{i=1}^{N}\left[\frac{\lambda_i}{2}\exp\left[-\frac{\lambda_i}{2}\gamma_i\right]\right] \\ &= \left(\frac{1}{2}\right)^N |\boldsymbol{\Psi}| \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\lambda_i\gamma_i\right], \end{aligned} \tag{14}$$

where $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_N]^T$ and $\boldsymbol{\Psi} = \mathrm{diag}(\boldsymbol{\lambda})$.

The first two stages of the Bayesian model, summarized in (9) and (14), constitute a sparsity-promoting nonnegative (truncated) Laplace prior. This prior can be obtained by marginalizing the hyperparameter vector $\boldsymbol{\gamma}$ from the model. In the one dimensional case, we get

$$\begin{aligned} p(w_i|\lambda_i, \beta) &= \int_0^{\infty} p(w_i|\gamma_i, \beta) p(\gamma_i|\lambda_i) d\gamma_i \\ &= \sqrt{\beta\lambda_i}\exp\left[-\sqrt{\beta\lambda_i}\,|w_i|\right] \mathcal{I}_{\mathbf{R}_+^1}(w_i), \end{aligned} \tag{15}$$

whereas, for the full model, the truncated Laplace prior is given by

$$\begin{aligned} p(\mathbf{w}|\boldsymbol{\lambda}, \beta) &= \int p(\mathbf{w}|\boldsymbol{\gamma}, \beta) p(\boldsymbol{\gamma}|\boldsymbol{\lambda}) d\boldsymbol{\gamma} \\ &= \prod_{i=1}^{N} \int_0^{\infty} p(w_i|\gamma_i, \beta) p(\gamma_i|\lambda_i) d\gamma_i \\ &= \beta^{\frac{N}{2}} |\boldsymbol{\Psi}|^{\frac{1}{2}} \prod_{i=1}^{N}\left[\exp\left[-\sqrt{\beta\lambda_i}\,|w_i|\right] \mathcal{I}_{\mathbf{R}_+^1}(w_i)\right] \\ &= \beta^{\frac{N}{2}} |\boldsymbol{\Psi}|^{\frac{1}{2}} \exp\left[-\sqrt{\beta}\sum_{i=1}^{N}\sqrt{\lambda_i}\,|w_i|\right] \mathcal{I}_{\mathbf{R}_+^N}(\mathbf{w}). \end{aligned} \tag{16}$$

Our intention behind the use of a hyperparameter vector $\boldsymbol{\lambda}$ instead of a single λ for all γ_i's is to form a hierarchical Bayesian analogue to the adaptive Lasso, proposed in [21]. Indeed, as it is shown in Appendix B, the MAP estimator of $\mathbf{w}$ that follows the truncated Laplace prior of (16) coincides with the estimation of $\mathbf{w}$ resulting via the optimization of the nonnegativity constrained adaptive Lasso criterion, which is expressed as

$$\tilde{\mathbf{w}} = \arg\min_{\mathbf{w}}\left\{\frac{\beta}{2}\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}\|_2^2 + \sum_{i=1}^{N}\alpha_i w_i\right\},\ \text{s.t. } \mathbf{w} \in \mathbf{R}_+^N, \tag{17}$$

for $\alpha_i = \sqrt{\beta\lambda_i}, i = 1 \ldots N$. As shown in (17), the main feature of the adaptive Lasso is that each coordinate w_i of $\mathbf{w}$ is now weighted by a distinct positive parameter α_i. This modification results in a consistent estimator, [21], which is not the case for the original Lasso estimator (7).

It is obvious from (16) that the quality of the endmember selection procedure depends on the tuning parameter vector $\boldsymbol{\lambda}$. Typically, tuning parameters reflect one's prior knowledge about the estimation problem and they can either be manually set, or can be considered as random variables. We choose the latter alternative, by assuming a Gamma hyperprior for $\boldsymbol{\lambda}$,

$$\begin{aligned} p(\lambda_i|r, \delta) &= \Gamma(\lambda_i|r, \delta) \\ &= \frac{\delta^r}{\Gamma(r)}{\lambda_i}^{r-1}\exp\left[-\delta\lambda_i\right],\ i = 1, 2, \ldots, N \end{aligned} \tag{18}$$

where r and δ are hyperparameters, with $r \geq 0$ and $\delta \geq 0$. Both Gamma priors of β, in (10) and λ_i, in (18), are flexible enough to express prior information, by properly tuning their hyperparameters. In this paper, we use a noninformative Jeffrey's prior ($p(x) \propto \frac{1}{x}$) over these parameters, which is obtained from (10) and (18) by setting all hyperparameters $\kappa, \theta, r, \delta$ of the Gamma distributions to zero, as in [9], [18], [19]. A schematic representation of the proposed hierarchical Bayesian model in the form of a directed acyclic graph is shown in Fig. 1.

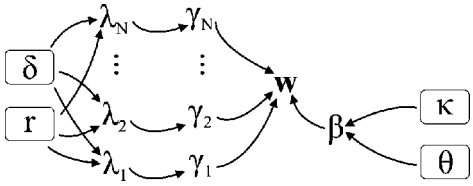

Fig. 1. Directed acyclic graph of the proposed Bayesian model. The deterministic model parameters appear in boxes.

IV. The Proposed Bayesian Inference Methodology

As it is common in Bayesian inference, the estimation of the parameters is based on their joint posterior distribution. This posterior for the model presented in Section III is expressed as

$$p\left(\mathbf{w},\beta,\boldsymbol{\gamma},\boldsymbol{\lambda}|\mathbf{y}\right)=\frac{p\left(\mathbf{y}|\mathbf{w},\beta\right)p\left(\mathbf{w}|\beta,\boldsymbol{\gamma}\right)p\left(\boldsymbol{\gamma}|\boldsymbol{\lambda}\right)p\left(\boldsymbol{\lambda}\right)p\left(\beta\right)}{p(\mathbf{y})}, \tag{19}$$

which is intractable, in the sense that the integral

$$p(\mathbf{y})=\int\int\int\int p\left(\mathbf{y},\mathbf{w},\beta,\boldsymbol{\gamma},\boldsymbol{\lambda}\right)d\mathbf{w}d\boldsymbol{\gamma}d\boldsymbol{\lambda}d\beta \tag{20}$$

cannot be expressed in closed form. In such cases, the Gibbs sampler [22] provides an alternative method for overcoming this impediment. The Gibbs sampler generates random samples from the conditional posterior distributions of the associated model parameters iteratively. As explained in [32], this sampling procedure generates a Markov chain of random variables, which converges to the joint distribution (19) (usually the first few iterations, also called burn-in, are ignored). In the sequel, we compute first the conditional posterior distributions, which are vital for the proposed Bayesian inference algorithm, and we explain the difficulty of utilizing Gibbs sampling in the present application. Then the proposed algorithm is discussed in detail.

A. Posterior Conditional Distributions

In this subsection, in accordance with the Gibbs sampler spirit, we derive the conditional posterior distributions of the model parameters $\mathbf{w}$, $\boldsymbol{\gamma}$, $\boldsymbol{\lambda}$ and β. Starting with $\mathbf{w}$, it is easily shown (utilizing (4) and (9)) that its posterior conditional density is a truncated multivariate Gaussian in $\mathbf{R}_+^N$,

$$\begin{aligned}p(\mathbf{w}|\mathbf{y},\boldsymbol{\gamma},\boldsymbol{\lambda},\beta)&=\frac{p(\mathbf{y}|\mathbf{w},\beta)p(\mathbf{w}|\boldsymbol{\gamma},\beta)}{\int p(\mathbf{y}|\mathbf{w},\beta)p(\mathbf{w}|\boldsymbol{\gamma},\beta)d\mathbf{w}}\\&=\mathcal{N}_{\mathbf{R}_+^N}(\mathbf{w}|\boldsymbol{\mu},\boldsymbol{\Sigma})\end{aligned} \tag{21}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are expressed as follows, [33, theorem 10.3]

$$\boldsymbol{\mu}=\beta\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{y} \tag{22}$$

$$\boldsymbol{\Sigma}=\beta^{-1}\left[\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\boldsymbol{\Lambda}\right]^{-1}. \tag{23}$$

The posterior conditional for the precision parameter β, after eliminating the terms which are independent of β, is expressed as

$$p(\beta|\mathbf{y},\mathbf{w},\boldsymbol{\gamma},\boldsymbol{\lambda})=\frac{p(\mathbf{y}|\mathbf{w},\beta)p(\mathbf{w}|\boldsymbol{\gamma},\beta)p(\beta)}{\int_0^\infty p(\mathbf{y}|\mathbf{w},\beta)p(\mathbf{w}|\boldsymbol{\gamma},\beta)p(\beta)d\beta}. \tag{24}$$

Utilizing (4), (9) and (10), it is easily shown that β is Gamma distributed as follows

$$\begin{aligned}&p(\beta|\mathbf{y},\mathbf{w},\boldsymbol{\gamma},\boldsymbol{\lambda})\\&=\Gamma\left(\beta\left|\frac{M+N}{2}+\kappa,\frac{1}{2}||\mathbf{y}-\boldsymbol{\Phi}\mathbf{w}||_2^2+\theta+\frac{1}{2}\mathbf{w}^T\boldsymbol{\Lambda}\mathbf{w}\right.\right).\end{aligned} \tag{25}$$

Straightforward computations, reported in Appendix C, yield that the conditional pdf of γ_i given $\mathbf{y}$, w_i, λ_i, β_i is the following generalized inverse Gaussian distribution [34]

$$\begin{aligned}&p(\gamma_i|\mathbf{y},w_i,\lambda_i,\beta)\\&\quad=\frac{p(\mathbf{y}|w_i,\beta)p(w_i|\gamma_i,\beta)p(\gamma_i|\lambda_i)p(\lambda_i)p(\beta)}{\int p(\mathbf{y}|w_i,\beta)p(w_i|\gamma_i,\beta)p(\gamma_i|\lambda_i)p(\lambda_i)p(\beta)d\gamma_i}\\&\quad=\left(\frac{\lambda_i}{2\pi}\right)^{\frac{1}{2}}\gamma_i^{-\frac{1}{2}}\exp\left[-\frac{\beta w_i^2}{2\gamma_i}-\frac{\lambda_i}{2}\gamma_i+\sqrt{\beta\lambda_i}\,|w_i|\right],\\&\quad i=1,2,\ldots,N.\end{aligned} \tag{26}$$

Finally, the conditional posterior of λ_i given $\mathbf{y}$, w_i, γ_i, β is expressed as

$$p(\lambda_i|\mathbf{y},w_i,\gamma_i,\beta)=\frac{p(\gamma_i|\lambda_i)p(\lambda_i)}{\int_0^\infty p(\gamma_i|\lambda_i)p(\lambda_i)d\lambda_i}, \tag{27}$$

which, using (13) and (18), is shown to be a Gamma pdf,

$$p(\lambda_i|\mathbf{y},w_i,\gamma_i,\beta)=\Gamma\left(\lambda_i|1+r,\frac{\gamma_i}{2}+\delta\right),\ i=1,2,\ldots,N. \tag{28}$$

The Gibbs sampler generates a sequence of samples $\mathbf{w}^{(t)}$, $\beta^{(t)}$, $\gamma_i^{(t)}$, and $\lambda_i^{(t)}$, $i=1,2,\ldots,N$, by sampling the conditional pdfs (21), (25), (26), and (28), respectively.

In this paper, a different procedure is followed. Specifically, we propose a deterministic approximation of the Gibbs sampler, where the randomly generated samples of the Gibbs sampler are replaced by the *means* of the corresponding conditional distributions, (21), (25), (26), and (28). Thus, a novel iterative scheme among the conditional means of $\mathbf{w}$, β, γ_i, and λ_i arises, termed *Bayesian inference iterative conditional expectations (BI-ICE)* algorithm. It should be emphasized that by following this approach, we depart from the statistical framework implied by the Gibbs sampler and we end up with a new deterministic algorithm for estimating the parameters of the proposed hierarchical model. Besides avoiding the complexity of sampling (26), BI-ICE is expected to converge faster than the original Gibbs sampler and, as a result, is expected to be much less computationally demanding. Also, as verified by extensive simulations, BI-ICE leads to sparse solutions and offers robust estimation performance under various experimental settings.

B. The BI-ICE Algorithm

As mentioned previously, BI-ICE needs the conditional expectations of the model parameters. These are computed analytically as described below.

1) Expectation of $p(\mathbf{w}|\mathbf{y},\boldsymbol{\gamma},\boldsymbol{\lambda},\beta)$: As shown in (21), $p(\mathbf{w}|\mathbf{y},\boldsymbol{\gamma},\boldsymbol{\lambda},\beta)$ is a truncated Gaussian distribution in $\mathbf{R}_+^N$. We

know from [35] that in the one-dimensional case, the expectation of a random variable x modeled by the truncated Gaussian distribution in $\mathbf{R}_+^1$ can be computed as

$$x \sim \mathcal{N}_{\mathbf{R}_+^1}(x|\mu^*, \sigma^{*2}) \Rightarrow \mathrm{E}[x] = \mu^* + \frac{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{\mu^{*2}}{\sigma^{*2}}\right)}{1 - \frac{1}{2}\mathrm{erfc}\left(\frac{\mu^*}{\sqrt{2}\sigma^*}\right)}\sigma^* \tag{29}$$

where $\mathrm{erfc}(\cdot)$ is the complementary error function. Unfortunately, to the best of our knowledge, there is no analogous closed form expression for the N-dimensional case. However, as shown in [36] and [37], the distribution of the ith element of $\mathbf{w}$ conditioned on the remaining elements $\mathbf{w}_{\neg i} = [w_1, \ldots, w_{i-1}, w_{i+1}, \ldots, w_N]^T$ can be expressed as

$$w_i|\mathbf{w}_{\neg i} \sim \mathcal{N}_{\mathbf{R}_+^1}(w_i|\mu_i^*, \sigma_{ii}^*) \tag{30}$$

with

$$\mu_i^* = \mu_i + \boldsymbol{\sigma}_{\neg i}^T \boldsymbol{\Sigma}_{\neg i \neg i}^{-1}(\mathbf{w}_{\neg i} - \boldsymbol{\mu}_{\neg i}) \tag{31}$$

$$\sigma_{ii}^* = \sigma_{ii} - \boldsymbol{\sigma}_{\neg i}^T \boldsymbol{\Sigma}_{\neg i \neg i}^{-1} \boldsymbol{\sigma}_{\neg i}. \tag{32}$$

Recalling that $\mathbf{w} \sim \mathcal{N}_{\mathbf{R}_+^N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, μ_i and σ_{ii} represent the ith and iith elements of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively. The $(N-1)\times(N-1)$ matrix $\boldsymbol{\Sigma}_{\neg i \neg i}$ is formed by removing the ith row and the ith column from $\boldsymbol{\Sigma}$, while the $(N-1)\times 1$ vector $\boldsymbol{\sigma}_{\neg i}$ is the ith column of $\boldsymbol{\Sigma}$ after removing its ith element. By applying (29) and utilizing (31)–(32), the expected values of all random variables $w_i|\mathbf{w}_{\neg i}, i = 1, 2, \ldots, N$ can be analytically computed. Based on this result, an iterative procedure is proposed in order to compute the mean of the posterior $p(\mathbf{w}|\mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \beta)$. Specifically, the jth iteration, $j = 1, 2, \ldots$, of this procedure is described as follows[3]

$$\begin{aligned}
1.\ w_1^{(j)} &= \mathrm{E}\left[p(w_1|w_2^{(j-1)}, w_3^{(j-1)}, \ldots, w_N^{(j-1)})\right] \\
2.\ w_2^{(j)} &= \mathrm{E}\left[p(w_2|w_1^{(j)}, w_3^{(j-1)}, \ldots, w_N^{(j-1)})\right] \\
&\vdots \\
N.\ w_N^{(j)} &= \mathrm{E}\left[p(w_N|w_1^{(j)}, w_2^{(j)}, \ldots, w_{N-1}^{(j)})\right].
\end{aligned} \tag{33}$$

This procedure is repeated iteratively until convergence. Experimental results have shown that the iterative scheme in (33) converges to the mean of $\mathbf{w} \sim \mathcal{N}_{\mathbf{R}_+^N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ after a few iterations.

2) Expectation of $p(\beta|\mathbf{y}, \mathbf{w}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$: The mean value of the Gamma distribution in (25) is given by

$$\mathrm{E}[p(\beta|\mathbf{y}, \mathbf{w}, \boldsymbol{\gamma}, \boldsymbol{\lambda})] = \frac{\frac{M+N}{2} + \kappa}{\frac{1}{2}||\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}||_2^2 + \theta + \frac{1}{2}\mathbf{w}^T\boldsymbol{\Lambda}\mathbf{w}} \tag{34}$$

3) Expectation of $p(\gamma_i|\mathbf{y}, w_i, \lambda_i, \beta)$: As shown in Appendix C, this expectation is expressed as

$$\begin{aligned}
&\mathrm{E}[p(\gamma_i|\mathbf{y}, w_i, \lambda_i, \beta)] \\
&= \left(\frac{2\lambda_i}{\pi}\right)^{\frac{1}{2}} \left(\frac{\beta w_i^2}{\lambda_i}\right)^{\frac{3}{4}} \exp\left[\sqrt{\beta\lambda_i}\,|w_i|\right] K_{\frac{3}{2}}\left(\sqrt{\beta\lambda_i}\,|w_i|\right),
\end{aligned} \tag{35}$$

[3]In the following, for notational simplicity, the expectation $E_{p(x|y)}[x]$ of a random variable x with conditional distribution $p(x|y)$ is denoted as $E[p(x|y)]$.

TABLE I
THE BI-ICE ALGORITHM

Input $\boldsymbol{\Phi}$, $\mathbf{y}$, $\kappa, \theta, r, \delta$
Initialize $\boldsymbol{\gamma}^{(0)} = \boldsymbol{\lambda}^{(0)} = \mathbf{1}, \beta^{(0)} = 0.01\,\|\mathbf{y}\|_2$
for $t = 1, 2, \ldots$ **do**
- Compute $\mathbf{w}^{(t)}$ as follows
  - Compute $\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}$ using (22), (23)
  - Set $\mathbf{v}^{(0)} = \boldsymbol{\mu}^{(t)}$
  - Compute $v_1^{(1)} = \mathrm{E}\left[p(v_1|v_2^{(0)}, \ldots, v_N^{(0)})\right]$, using (31), (32), and (29)
  - Compute $v_2^{(1)} = \mathrm{E}\left[p(v_2|v_1^{(1)}, v_3^{(0)}, \ldots, v_N^{(0)})\right]$, using (31), (32), and (29)
  - $\vdots$
  - Compute $v_N^{(1)} = \mathrm{E}\left[p(v_N|v_1^{(1)}, v_2^{(1)}, \ldots, v_{N-1}^{(1)})\right]$, using (31), (32), and (29)
  - Set $\mathbf{w}^{(t)} = \mathbf{v}^{(1)}$
- Compute $\beta^{(t)} = \mathrm{E}\left[p(\beta|\mathbf{y}, \mathbf{w}^{(t)}, \boldsymbol{\gamma}^{(t-1)}, \boldsymbol{\lambda}^{(t-1)}\right]$, using (34)
- Compute $\gamma_i^{(t)} = \mathrm{E}\left[p(\gamma_i|\mathbf{y}, w_i^{(t)}, \lambda_i^{(t-1)}, \beta^{(t)})\right],\ i = 1, 2, \ldots, N$, using (35)
- Compute $\lambda_i^{(t)} = \mathrm{E}\left[p(\lambda_i|\mathbf{y}, w_i^{(t)}, \gamma_i^{(t)}, \beta^{(t)})\right],\ i = 1, 2, \ldots, N$, using (36)

endfor

where $K_\nu(\cdot)$ stands for the modified Bessel function of second kind of order ν.

4) Expectation of $p(\lambda_i|\mathbf{y}, w_i, \gamma_i, \beta)$: Again, the mean value of the Gamma distribution in (28) is given by

$$\mathrm{E}[p(\lambda_i|\mathbf{y}, w_i, \gamma_i, \beta)] = \frac{1+r}{\frac{1}{2}\gamma_i + \delta}. \tag{36}$$

Based on the previous expressions, the proposed BI-ICE algorithm is summarized in Table I. As shown in the Table, the algorithm is initialized with $\boldsymbol{\lambda}^{(0)} = \mathbf{1}, \boldsymbol{\gamma}^{(0)} = \mathbf{1}$ and as in [19], $\beta^{(0)} = 0.01\,\|\mathbf{y}\|_2$.

Regarding the updating of parameter $\mathbf{w}^{(t)}$, an auxiliary variable $\mathbf{v}$ has been utilized in Table I. This is initialized with $\boldsymbol{\mu}^{(t)}$ (the value of $\boldsymbol{\mu}$ at iteration t) and is updated by performing a *single* iteration of the scheme described in (33). The resulting value of $\mathbf{v}$ is assigned to $\mathbf{w}^{(t)}$. The rationale behind this choice is that for a diagonal $\boldsymbol{\Sigma}$ (which happens when the columns of $\boldsymbol{\Phi}$ are orthogonal), it easily follows from (31), (32) that the w_i's in (33) are uncorrelated. Thus, a single iteration is sufficient to obtain the mean of $\mathcal{N}_{\mathbf{R}_+^N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Although, this is not valid when $\boldsymbol{\Sigma}$ is not diagonal, experimental results have evidenced that the estimation of the mean of $\mathcal{N}_{\mathbf{R}_+^N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ resulting after the execution of a single iteration of the scheme in (33) is also sufficient in the framework of the BI-ICE algorithm.

Due to the fact that the BI-ICE algorithm springs out from the hierarchical Bayesian model described in Section III, it leads to sparse estimations for $\mathbf{w}$, and the endmembers present in the pixel are identified by the nonzero entries of $\mathbf{w}$. In addition, all parameters of the model are naturally estimated from the data, as a consequence of the Bayesian Lasso approach followed in this paper. This is in contrast to deterministic algorithms for solving the Lasso, e.g., [11], [21], or adaptive methods, [16], which face the problem of fine-tuning specific parameters, (corresponding to $\boldsymbol{\lambda}$ of our model), that control the sparsity of the solution. Besides, useful by-products of the BI-ICE algorithm are the estimates of (a) the variance of the additive noise of the linear model, as in [9], and (b) the variance of the abundance vector. The latter, coupled with the estimate of $\boldsymbol{\mu}$, provides the posterior distribution of the abundance vector, which can be used to provide confidence intervals to assess the reliability of the proposed estimator.

Concerning the computational complexity, as it is clear from Table I, the BI-ICE algorithm requires the evaluation of simple closed form formulas. The main computational burden is due to the calculation of the N inverse matrices $\boldsymbol{\Sigma}_{\neg i\neg i},\ i=1,2,\ldots,N$ appearing in (31) and (32). As shown in Appendix D, all these matrices can be derived very efficiently from $\boldsymbol{\Sigma}^{-1}$, and thus only one matrix inversion per iteration (related to the computation of $\boldsymbol{\Sigma}$ in (23)) is required. This results in a reduction of the computational complexity of the BI-ICE algorithm by one order of magnitude per iteration.

Thus far, the proposed BI-ICE algorithm has been described as a deterministic approximation of the Gibbs sampler. An alternative view of the BI-ICE algorithm in the framework of variational Bayesian inference is provided in Appendix E. As shown in the Appendix, the adoption of a proper factorization of an approximation of the posterior $p(\mathbf{w},\boldsymbol{\gamma},\boldsymbol{\lambda},\beta|\mathbf{y})$ results to a variational Bayesian inference scheme that exploits the same type of distributions and updates the same form of parameters. From this point of view, BI-ICE can be thought of as a first moments approximation to a variational Bayesian inference scheme.

C. Embedding the Sum-to-one Constraint

The sparsity-promoting hierarchical Bayesian model presented in the previous sections takes into consideration the nonnegativity of the abundance vector $\mathbf{w}$. However, the abundances' sum-to-one constraint has not yet been considered. As noted in [38], the sum-to-one constraint is prone to strong criticisms. In real hyperspectral images the spectral signatures are usually defined up to a scale factor, and thus, the sum-to-one constraint should be replaced by a generalized constraint of the form $\sum c_i w_i = 1$, in which the weights c_i denote the pixel-dependent scale factors. Moreover, it is known that the sparse solution of a linear system with $\boldsymbol{\Phi}$ having nonnegative entries already admits a generalized sum-to-one constraint, [39]. Thus, it can be safely assumed that the impact of not enforcing the sum-to-one constraint on the performance of the algorithm is not expected to be severe. Despite this fact, in this section we describe an efficient way to enforce this constraint, although through a regularization parameter.

Note that direct incorporation of this constraint to the proposed Bayesian framework would require truncation of the prior normal distribution of $\mathbf{w}$ over a simplex, rendering the derivation of closed form expressions for the conditional posterior distributions intractable. To alleviate this, we choose, as in [7], [10], [40, p. 586], to impose the sum-to-one constraint deterministically, by augmenting the initial LMM of (1) with an extra equation as follows:

$$\begin{bmatrix}\mathbf{y}\\ \alpha\end{bmatrix}=\begin{bmatrix}\boldsymbol{\Phi}\\ \alpha\mathbf{1}^T\end{bmatrix}\mathbf{w}+\begin{bmatrix}\mathbf{n}\\ 0\end{bmatrix} \tag{37}$$

where α is a scalar parameter, which controls the effect of the sum-to-one constraint on the estimation of $\mathbf{w}$. Specifically, the larger the value of α is, the closer the sum of the estimated w_i's will be to one. It should be noticed that the augmentation of the LMM as in (37) does not affect the proposed hierarchical Bayesian model and the subsequent analysis.

V. Experimental Results

A. Simulation Results on Synthetic Data

This section illustrates the effectiveness of the proposed BI-ICE algorithm, by a series of experiments related to the unmixing of a synthetic hyperspectral image. Following the experimental settings of [38], where a thorough comparison of several sparse semisupervised unmixing algorithms is presented, we consider two spectral data sets for the simulated hyperspectral scene: (a) $\boldsymbol{\Phi}_1 \in \mathcal{R}^{453\times 220}$, which is a matrix containing the spectral signatures of 220 endmembers selected from the USGS spectral library, [41], and (b) $\boldsymbol{\Phi}_2 \in \mathcal{R}^{453\times 220}$, which is a matrix of i.i.d. components uniformly distributed in the interval $[0\ 1]$. As expected, the spectral signatures of the materials of $\boldsymbol{\Phi}_1$ are highly correlated. The condition number and the mutual coherence, [38], of $\boldsymbol{\Phi}_1$ are 36.182×10^6 and 0.999933, respectively, whereas, for $\boldsymbol{\Phi}_2$, the same measures are equal to 82 and 0.8373, respectively.

The abundance fractions of the simulated image and the number of different endmembers composing a single pixel are generated according to a Dirichlet distribution, [6]. In all simulations, the observations are considered to be corrupted by either white Gaussian or colored noise. Colored noise is produced by filtering a sequence of white noise using a low-pass filter with a normalized cutoff frequency of $5\pi/M$. The variance of the additive noise is determined by the SNR level.

First, the fast convergence and sparse estimations of $\mathbf{w}$ exhibited by the new algorithm are depicted in Fig. 2. In this experiment, a pixel with three nonzero abundances (0.1397, 0.2305, 0.6298) is considered, and white noise is added to the model, such that the SNR is equal to 25dB. The curves in Fig. 2 are the average of 50 noise realizations. We observe that less than 15 iterations are sufficient for the BI-ICE algorithm to converge to the correct sparse solution of $\mathbf{w}$. That is, it determines correctly the abundance fractions of the endmembers present in the pixel, while all remaining abundance fractions converge to zero.

Next, the BI-ICE algorithm was compared to: (a) the least squares (LS) algorithm, (b) a quadratic programming (QP) technique, which enforces the constraints, but does not specifically exploit the problem's sparsity, [8], (c) the orthogonal matching pursuit (OMP) algorithm, [12], which is a widely used, greedy, sparsity promoting algorithm, (d) the sparse unmixing by variable splitting and augmented Lagrangian (SUnSAL) algorithm, [16], [38], which is based on the alternating direction method of multipliers to solve the ℓ_1 penalization problem of (7) subject to the physical constraints of the unmixing problem, and (e) the constrained version of SUnSAL, CSUnSAL, which solves the constrained version of the problem in (7), (see also [38] for details). In our experiments, the parameters used for SUnSAL are $\mu=1$ and $\lambda=1$, while for CSUnSAL we used $\mu=1$, $\lambda=10^{-3}$ and $\delta=10^{-6}$, see also [16]. Based on the following metric:

$$\mathrm{MSE}=\mathrm{E}\left[\frac{\|\mathbf{w}-\hat{\mathbf{w}}\|_2^2}{\|\mathbf{w}\|_2^2}\right] \tag{38}$$

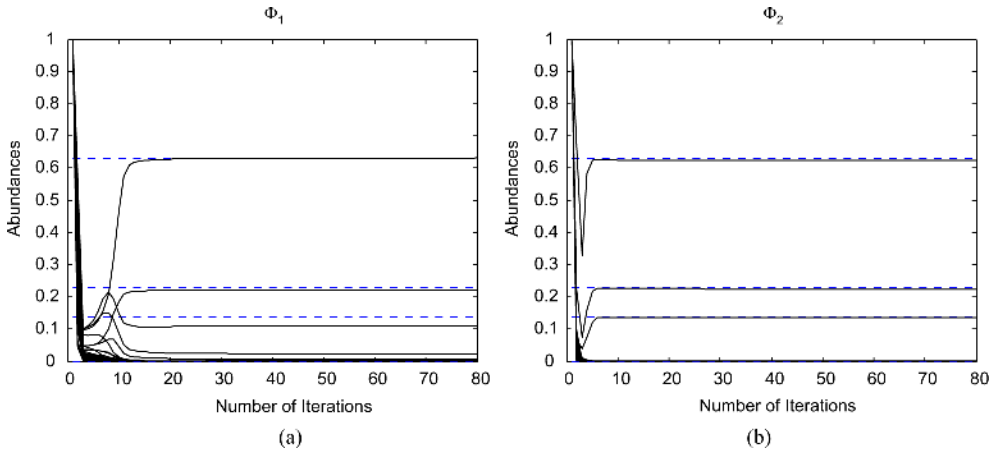

Fig. 2. Estimation of the entries of the sparse vector $\mathbf{w}$, as BI-ICE progresses. The algorithm is applied to simulated data, generated using (a) a highly correlated matrix of spectral data (b) a matrix of i.i.d uniform data. White noise is added (SNR = 25 dB). Dashed lines: True values. Solid lines: Estimated values.

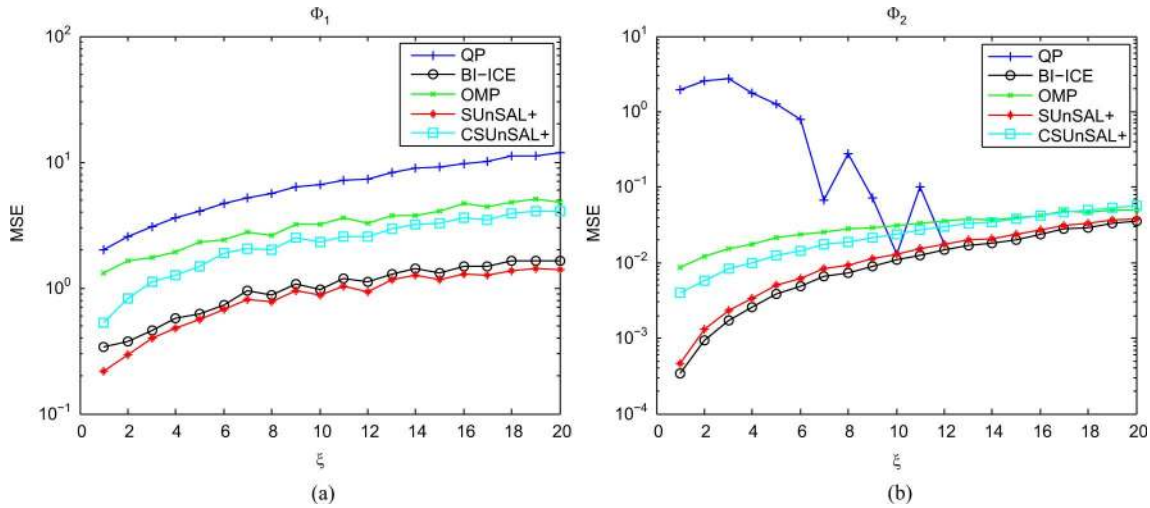

Fig. 3. MSE as a function of the level of sparsity obtained by different unmixing methods when applied to simulated data with white additive noise (SNR = 20 dB) and using two different spectral libraries.

where $\mathbf{w}$ and $\hat{\mathbf{w}}$ are the true and the estimated abundance vectors, respectively, the corresponding MSE curves for different sparsity levels ranging from 1 (pure pixel) to 20 are shown in Fig. 3, for both spectral libraries $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$. Due to poor results, the MSE curve of the LS algorithm is not shown in the figure. It can be seen that the proposed algorithm outperforms the OMP, QP, and CSUnSAL algorithms and has similar performance to the SUnSAL algorithm. In comparison to BI-ICE, the adaptive methods SUnSAL and CSUnSAL are of lower computational complexity. However, it should be pointed out that the comparable performance, in terms of MSE, of the alternating direction algorithms SUnSAL and CSUnSAL with BI-ICE comes at the additional expense of manually fine-tuning nontrivial parameters, such as the sparsity promoting parameter λ, (see (7), and [38]). Thus, an advantage of the proposed BI-ICE algorithm over SUnSAL and CSUnSAL algorithms is that all unknown parameters are directly inferred from the data. Besides that, BI-ICE bears interesting byproducts such as: (a) estimates of all model parameters; a useful parameter in many applications is the noise variance; (b) estimates for the variances of the estimated parameters, which may serve as confidence intervals; and (c) approximate posterior distributions for the estimated parameters. In contrast, all other algorithms considered are iterative algorithms that return point estimates of the parameters of interest.

A quick view of Fig. 3 also reveals that the OMP and QP algorithms attain the worst performance, in terms of MSE. OMP adds one endmember to its active set in each iteration, and subtracts its contribution from the residual signal, until the correlation coefficient of the remaining signal vector drops below a certain threshold, or the maximum of 20 selected endmembers is reached. However, due to its greedy nature and the high conditioning of $\mathbf{\Phi}_1$, OMP fails to detect the correct endmembers that compose the pixel. This is the reason for the algorithm's poor performance, shown in Fig. 3. Note also that, in the cases of high sparsity, the QP algorithm fails to detect the correct support of the sparse vector $\mathbf{w}$, resulting in poor MSE performance. This may not come as a surprise, since the QP algorithm is not specifically designed for sparse regression problems.

In Fig. 4 the MSE values of the various sparse unmixing algorithms versus the SNR are displayed. For this experiment, the spectral libraries $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$ were used to simulate two different

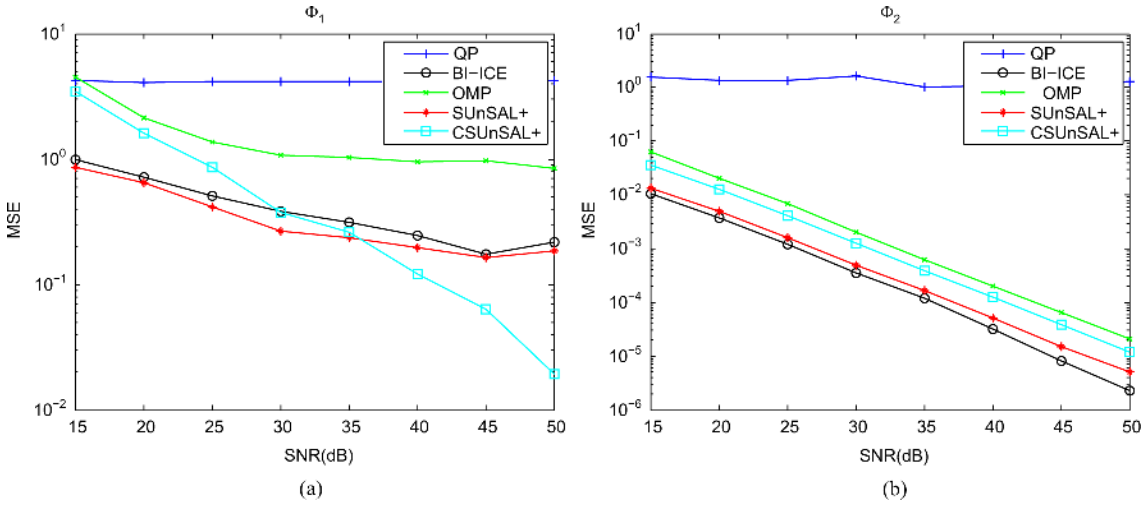

Fig. 4. MSE as a function of the SNR obtained by different sparse unmixing methods when applied to simulated data with white additive noise and using different spectral libraries for sparsity level $\xi = 5$.

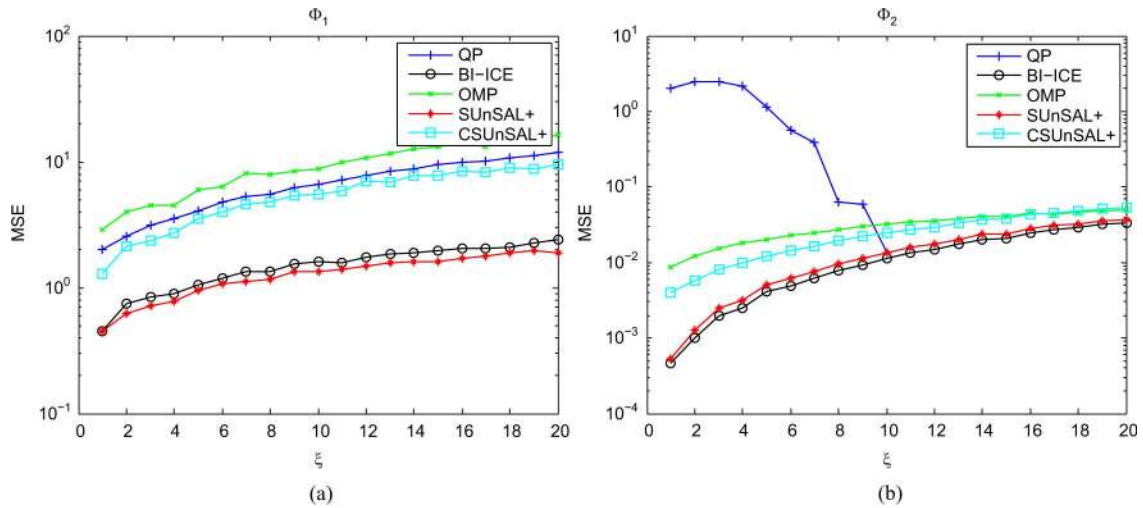

Fig. 5. MSE as a function of the level of sparsity obtained by different unmixing methods when applied to simulated data with colored additive noise (SNR = 20 dB) and using two different spectral libraries.

hyperspectral scenes, each having 100 pixels. The level of sparsity for the abundance vectors of all pixels is held fixed and equal to five. As expected, the MSE values of all algorithms decrease as the SNR increases. This is not the case for the QP algorithm though, which completely fails to retrieve the correct support of the sparse abundance vector $\mathbf{w}$, and its MSE is almost constant. Again, the performance of SUnSAL and BI-ICE is comparable, with BI-ICE having slightly better performance in the case of the i.i.d. mixing matrix $\mathbf{\Phi}_2$. In Figs. 5 and 6 the same experimental results are provided in the scenario where the simulated pixels are contaminated with colored noise. We observe that the performance pattern of the various algorithms is not affected by the presence of colored noise, apart form the fact that the MSE values are now slightly increased. Although our hierarchical Bayesian model assumes i.i.d. noise, these figures provide us with enough evidence to conclude that the proposed BI-ICE algorithm can also provide reliable results in colored noise environments.

Finally, in Fig. 7 the MSE performance of the proposed BI-ICE algorithm is shown, when the sum-to-one constraint is incorporated to the regression problem, as explained earlier in Section IV-C, with $\alpha = 10^3$. It can be seen that the performance of the algorithm is particularly enhanced in the case of high sparsity, i.e., when the image pixel is either pure $(\xi = 1)$ or it is composed of a few $(\xi < 5)$ endmembers. As verified by experiments, the BI-ICE with the sum-to-one constraint correctly detects the support of the sparse signal with a probability close to one, which accounts for a significant decrease of the MSE. The experiment has been conducted for both spectral libraries $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$. The higher MSE improvement is observed for the case of i.i.d. spectral data.

B. Simulation Results on Real Data

This section describes the application of the proposed BI-ICE algorithm to real hyperspectral image data. The real data were

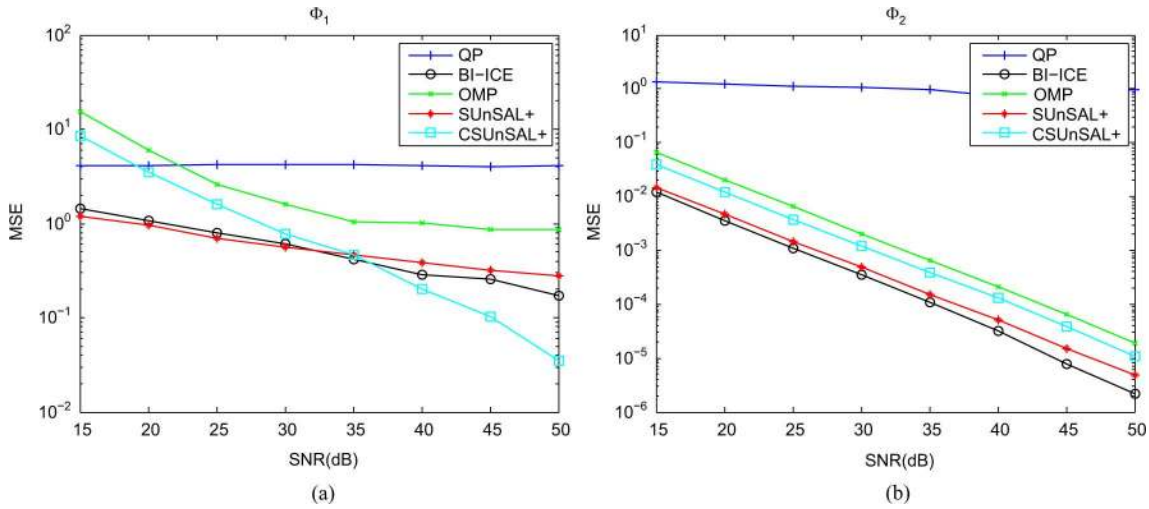

Fig. 6. MSE as a function of the SNR obtained by different sparse unmixing methods when applied to simulated data with colored additive noise and using different spectral libraries for sparsity level $\xi = 5$.

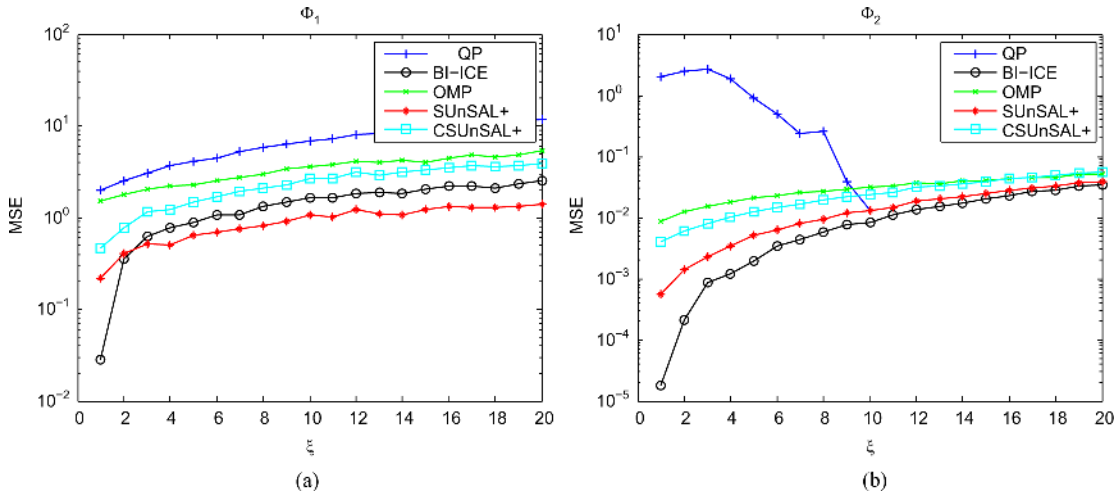

Fig. 7. MSE as a function of the level of sparsity obtained by different unmixing methods when applied to simulated data with white additive noise (SNR = 20 dB) and using two different spectral libraries. The sum-to-one constraint is incorporated to the BI-ICE algorithm, as explained in Section IV-C.

collected by the airborne visible/infrared imaging spectrometer (AVIRIS) flight over the Cuprite mining site, Nevada, in 1997, [42]. The AVIRIS sensor is a 224-channel imaging spectrometer with approximately 10-nm spectral resolution covering wavelengths ranging from 0.4 to 2.5 μm. The spatial resolution is 20 m. This data set has been widely used for remote sensing experiments [6], [43]–[45]. The spectral bands 1–2, 104–113, 148–167, and 221–224 were removed due to low SNR and water-vapor absorption. Hence, a total of 188 bands were considered in this experiment. The subimage of the 150th band, including 200 vertical lines with 200 samples per line (200×200) is shown in Fig. 8.

The VCA algorithm was used to extract 14 endmembers present in the image, as in [6]. Using these spectral signatures, three algorithms are tested to estimate the abundances, namely the LS algorithm, the QP method, and the proposed BI-ICE algorithm. The unmixing process generates an output image for each endmember, depicting the endmember's estimated abundance fraction for each pixel. The darker the pixel, the smaller the contribution of this endmember in the pixel is. On the other hand, a light pixel indicates that the proportion of the endmember in the specific pixel is high. The abundance fractions of four endmembers, estimated using the LS, QP, and

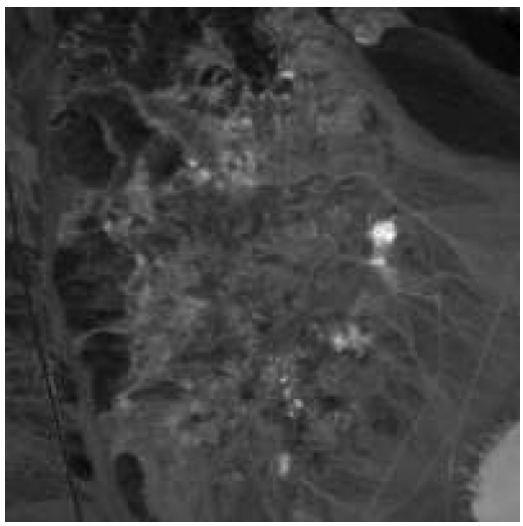

Fig. 8. Band 150 of a subimage of the Cuprite Aviris hyperspectral data set.

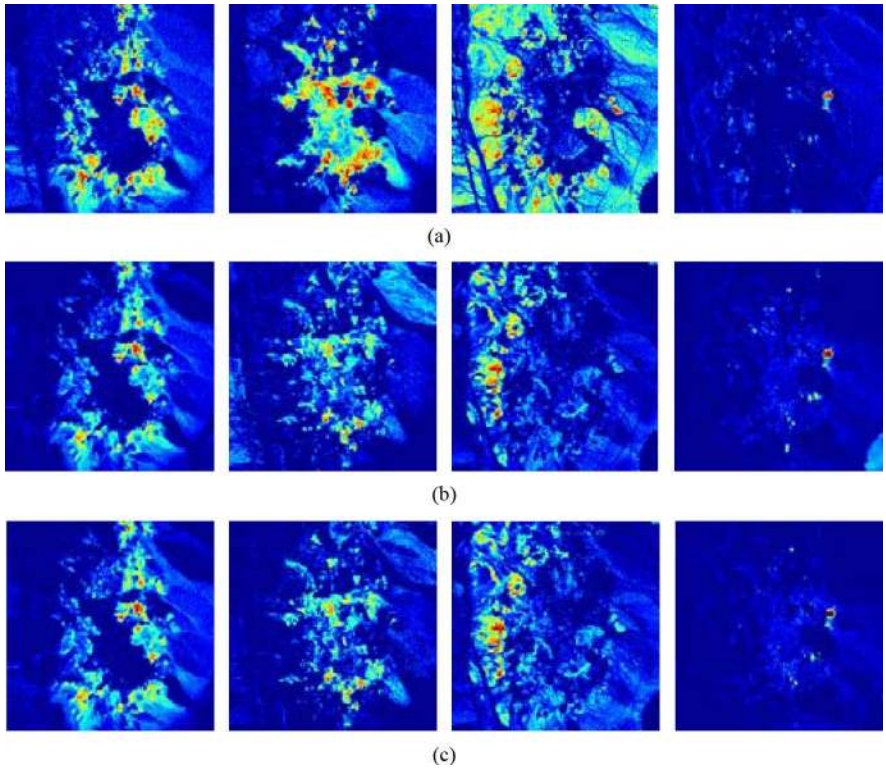

Fig. 9. Estimated abundance values of four endmembers using: (a) the LS algorithm; (b) the QP algorithm; (c) the proposed BI-ICE algorithm.

BI-ICE algorithms, are shown in Fig. 9(a)–(c), respectively. Note that, for the sake of comparison, a necessary linear scaling in the range $[0\ 1]$ has been performed for the LS abundance images. By simple inspection, it can be observed that the images taken using the LS algorithm clearly deviate from the images of the other two methods. The LS algorithm imposes no constraints on the estimated abundances, and hence the scaling has a major impact on the abundance fractions, resulting in performance degradation. On the contrary, the images obtained by QP and BI-ICE share a high degree of similarity and are in full agreement with previous results concerning the selected abundances and reported in [6], [45], as well as with the conclusions derived in Section V-A.

VI. Conclusion

A novel perspective for sparse semisupervised hyperspectral unmixing has been presented in this paper. The unmixing problem has been expressed in the form of a hierarchical Bayesian model, where the problem constraints and the parameters' properties were incorporated by suitably selecting the priors' and hyperpriors' distributions of the model. Then, a new Bayesian inference iterative scheme has been developed for estimating the model parameters. The proposed algorithm is computationally efficient, converges very fast and exhibits enhanced estimation performance compared to other related methods. Moreover, it provides sparse solutions, without necessitating the tuning of any parameters, which are naturally estimated from the algorithm. As it is also the case for other Bayesian inference methods, the theoretical proof of convergence of the proposed algorithm turns out to be a cumbersome task. Such a theoretical analysis is currently under investigation.

Appendix A
Derivation of the Truncated Gaussian Prior Distribution of $\mathbf{w}$

Assuming that all w_i's are i.i.d., the prior of the abundance vector $\mathbf{w}$ can be analytically expressed as

$$\begin{aligned} p(\mathbf{w}|\boldsymbol{\gamma},\beta) &= \prod_{i=1}^{N}\left[\mathcal{N}_{\mathbf{R}_+^1}(w_i|0,\frac{\gamma_i}{\beta})\right] \\ &= \prod_{i=1}^{N}\left[2\,(2\pi)^{-\frac{1}{2}}\,\beta^{\frac{1}{2}}\gamma_i^{-\frac{1}{2}}\exp\left[-\frac{\beta}{2}\frac{w_i^2}{\gamma_i}\right]\mathcal{I}_{\mathbf{R}_+^1}(w_i)\right] \\ &= 2^N\,(2\pi)^{-\frac{N}{2}}\,\beta^{\frac{N}{2}}\,|\boldsymbol{\Lambda}|^{\frac{1}{2}}\exp\left[-\frac{\beta}{2}\mathbf{w}^T\boldsymbol{\Lambda}\mathbf{w}\right]\mathcal{I}_{\mathbf{R}_+^N}(\mathbf{w}) \\ &= \mathcal{N}_{\mathbf{R}_+^N}(\mathbf{w}|0,\beta^{-1}\boldsymbol{\Lambda}^{-1}) \end{aligned} \tag{39}$$

where $\mathbf{R}_+^1$ is the set of nonnegative real numbers and $\mathbf{R}_+^N$ is the nonnegative orthant of $\mathcal{R}^N$, $\mathcal{N}_{\mathbf{R}_+^N}(\cdot)$ stands for the N-variate truncated normal distribution in $\mathbf{R}_+^N$ according to Definition 1, $\boldsymbol{\gamma}=[\gamma_1,\gamma_1,\ldots,\gamma_N]^T$ is the $N\times 1$ vector containing the hyperparameters, $\gamma_i\geq 0, i=1,2,\ldots,N$ and $\boldsymbol{\Lambda}$ is the $N\times N$ diagonal matrix, with $\boldsymbol{\Lambda}^{-1}=\mathrm{diag}(\boldsymbol{\gamma})$.

Appendix B
The Non-Negativity Constrained Bayesian Adaptive Lasso

The MAP estimator of $\mathbf{w}$ is defined as

$$\mathbf{w}_{\mathrm{MAP}}=\arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{y}). \tag{40}$$

From Bayes' theorem, the MAP estimator can be expressed as

$$\begin{aligned}\mathbf{w}_{\text{MAP}} &= \arg\max_{\mathbf{w}} p(\mathbf{y}|\mathbf{w},\beta)p(\mathbf{w}|\boldsymbol{\lambda},\beta) \\ &= \arg\min_{\mathbf{w}} \left\{-\log\left[p(\mathbf{y}|\mathbf{w},\beta)p(\mathbf{w}|\boldsymbol{\lambda},\beta)\right]\right\}. \quad (41)\end{aligned}$$

Then, substituting in (41) the likelihood function from (4) and the truncated Laplace prior from (16), the MAP estimator can be expressed as shown in (42) at the bottom of the page. Note that $-\log\left(\mathcal{I}_{\mathbf{R}_+^N}(\mathbf{w})\right) = \infty$, for $\mathbf{w} \notin \mathbf{R}_+^N$ and $-\log\left(\mathcal{I}_{\mathbf{R}_+^N}(\mathbf{w})\right) = 0$, for $\mathbf{w} \in \mathbf{R}_+^N$, i.e., this term severely penalizes $\mathbf{w}$'s with negative elements. Thus, it is established that the MAP estimation of $\mathbf{w}$, given the truncated Laplace prior of (16), is equivalent to solving the adaptive Lasso criterion of (17), for $\alpha_i = \sqrt{\beta\lambda_i}, i = 1,\ldots,N$, subject to $\mathbf{w}$ being nonnegative, i.e., $\mathbf{w} \in \mathbf{R}_+^N$.

Appendix C
The Conditional Posterior Distribution $p(\gamma_i|\mathbf{y},w_i,\lambda_i,\beta)$ and its Mean

Using (9) and (13) the posterior conditional distribution $p(\gamma_i|\mathbf{y},w_i,\lambda_i,\beta)$ for $w_i \geq 0$ can be computed as

$$\begin{aligned}&p(\gamma_i|\mathbf{y},w_i,\lambda_i,\beta) \\ &= \frac{p(\mathbf{y}|w_i,\beta)p(w_i|\gamma_i,\beta)p(\gamma_i|\lambda_i)p(\lambda_i)p(\beta)}{\int p(\mathbf{y}|w_i,\beta)p(w_i|\gamma_i,\beta)p(\gamma_i|\lambda_i)p(\lambda_i)p(\beta)d\gamma_i} \\ &= \frac{p(w_i|\gamma_i,\beta)p(\gamma_i|\lambda_i)}{\int p(w_i|\gamma_i,\beta)p(\gamma_i|\lambda_i)d\gamma_i} \\ &= \frac{2(2\pi)^{-\frac{1}{2}}\beta^{\frac{1}{2}}\gamma_i^{-\frac{1}{2}}\exp\left[-\frac{\beta}{2}\frac{w_i^2}{\gamma_i}\right]\mathcal{I}_{\mathbf{R}_+^1}(w_i)\frac{\lambda_i}{2}\exp\left[-\frac{\lambda_i}{2}\gamma_i\right]}{\int_0^\infty 2(2\pi)^{-\frac{1}{2}}\beta^{\frac{1}{2}}\gamma_i^{-\frac{1}{2}}\exp\left[-\frac{\beta}{2}\frac{w_i^2}{\gamma_i}\right]\mathcal{I}_{\mathbf{R}_+^1}(w_i)\frac{\lambda_i}{2}\exp\left[-\frac{\lambda_i}{2}\gamma_i\right]d\gamma_i} \\ &= \frac{\gamma_i^{-\frac{1}{2}}\exp\left[-\frac{\beta}{2}\frac{w_i^2}{\gamma_i}-\frac{\lambda_i}{2}\gamma_i\right]}{\int_0^\infty \gamma_i^{-\frac{1}{2}}\exp\left[-\frac{\beta}{2}\frac{w_i^2}{\gamma_i}-\frac{\lambda_i}{2}\gamma_i\right]d\gamma_i} \\ &= \frac{\gamma_i^{-\frac{1}{2}}\exp\left[-\frac{\beta}{2}\frac{w_i^2}{\gamma_i}-\frac{\lambda_i}{2}\gamma_i\right]}{\sqrt{\frac{2\pi}{\lambda_i}}\exp\left[-\sqrt{\beta\lambda_i w_i^2}\right]} \\ &= \left(\frac{\lambda_i}{2\pi}\right)^{\frac{1}{2}}\gamma_i^{-\frac{1}{2}}\exp\left[-\frac{\beta w_i^2}{2\gamma_i}-\frac{\lambda_i}{2}\gamma_i+\sqrt{\beta\lambda_i}\,|w_i|\right] \quad (43)\end{aligned}$$

where we used [46, eq. 3.471.15] to compute the integral. The mean of (43) is computed as

$$\begin{aligned}&\mathrm{E}\left[p(\gamma_i|\mathbf{y},w_i,\lambda_i,\beta)\right] \\ &= \int_0^\infty \gamma_i\, p(\gamma_i|\mathbf{y},w_i,\lambda_i,\beta)d\gamma_i \\ &= \int_0^\infty \left(\frac{\lambda_i}{2\pi}\right)^{\frac{1}{2}}\gamma_i^{\frac{1}{2}}\exp\left[-\frac{\beta w_i^2}{2\gamma_i}-\frac{\lambda_i}{2}\gamma_i+\sqrt{\beta\lambda_i}\,|w_i|\right]d\gamma_i \\ &= \left(\frac{\lambda_i}{2\pi}\right)^{\frac{1}{2}}\exp\left[\sqrt{\beta\lambda_i}\,|w_i|\right]\int_0^\infty \gamma_i^{\frac{1}{2}}\exp\left[-\frac{\beta w_i^2}{2\gamma_i}-\frac{\lambda_i}{2}\gamma_i\right]d\gamma_i \\ &= \left(\frac{2\lambda_i}{\pi}\right)^{\frac{1}{2}}\left(\frac{\beta w_i^2}{\lambda_i}\right)^{\frac{3}{4}}\exp\left[\sqrt{\beta\lambda_i}\,|w_i|\right]K_{\frac{3}{2}}\left(\sqrt{\beta\lambda_i}\,|w_i|\right) \quad (44)\end{aligned}$$

where we used [46, eq. 3.471.9] for the integral computation. Finally, we set $p(\gamma_i|\mathbf{y},w_i,\lambda_i,\beta) = 0$, for $w_i < 0$. Note that this does not affect the BI-ICE algorithm, since w_i's are guaranteed to be nonnegative (the fact $w_i < 0$ is impossible by the formulation of the problem).

Appendix D
Fast Computation of (31) and (32)

Let us define $\mathbf{V} = \boldsymbol{\Sigma}^{-1}$. In [36], the formula $\boldsymbol{\Sigma}_{\neg i\neg i}^{-1} = \mathbf{V}_{\neg i\neg i} - \mathbf{v}_{\neg i}\mathbf{v}_{\neg i}^T/\mathbf{V}_{ii}$, $i = 1,2,\ldots,N$, has been utilized for computing all matrices $\boldsymbol{\Sigma}_{\neg i\neg i}^{-1}$, $i = 1,2,\ldots,N$ from $\boldsymbol{\Sigma}^{-1}$, where $\mathbf{V}_{\neg i\neg i}$ and $\mathbf{v}_{\neg i}$ are related to $\mathbf{V}$ in the same way $\boldsymbol{\Sigma}_{\neg i\neg i}$ and $\boldsymbol{\sigma}_{\neg i}$ are related to $\boldsymbol{\Sigma}$. It has been seen in simulations that this rank-one downdate formula is numerically susceptible. In the following, an alternative method is proposed, which avoids direct computation of $\boldsymbol{\Sigma}_{\neg i\neg i}^{-1}$ and has exhibited numerical robustness in all simulations performed. Let $\mathbf{T}_i$ be an $N \times N$ permutation matrix, which when it premultiplies a matrix, moves its ith row to the Nth position, after upshifting rows $i+1,\ldots,N$. Then, by defining $\boldsymbol{\Sigma}_i = \mathbf{T}_i\boldsymbol{\Sigma}\mathbf{T}_i^T$, it is easily verified that

$$\boldsymbol{\Sigma}_i = \begin{bmatrix}\boldsymbol{\Sigma}_{\neg i\neg i} & \boldsymbol{\sigma}_{\neg i}^T \\ \boldsymbol{\sigma}_{\neg i} & \sigma_{ii}\end{bmatrix}. \quad (45)$$

Moreover, due to the orthogonality of $\mathbf{T}_i$, $\boldsymbol{\Sigma}_i^{-1} = \mathbf{T}_i\boldsymbol{\Sigma}^{-1}\mathbf{T}_i^T$, $i = 1,2,\ldots,N$, i.e., all $\boldsymbol{\Sigma}_i^{-1}$, $i = 1,2,\ldots,N$, are obtained from $\boldsymbol{\Sigma}^{-1}$ by simple permutations. From [47, p. 54] and (45), we get

$$\begin{aligned}\boldsymbol{\Sigma}_i^{-1} = &\begin{bmatrix}\boldsymbol{\Sigma}_{\neg i\neg i}^{-1} & \mathbf{0}^T \\ \mathbf{0} & 0\end{bmatrix} \\ &+ \frac{1}{\sigma_{ii} - \boldsymbol{\sigma}_{\neg i}^T\boldsymbol{\Sigma}_{\neg i\neg i}^{-1}\boldsymbol{\sigma}_{\neg i}}\begin{bmatrix}-\boldsymbol{\Sigma}_{\neg i\neg i}^{-1}\boldsymbol{\sigma}_{\neg i} \\ 1\end{bmatrix}\left[-\boldsymbol{\sigma}_{\neg i}^T\boldsymbol{\Sigma}_{\neg i\neg i}^{-1} \quad 1\right]. \quad (46)\end{aligned}$$

Let

$$\begin{aligned}\mathbf{w}_{\text{MAP}} &= \text{argmin}_{\mathbf{w}}\left\{-\log\left[(2\pi)^{-\frac{M}{2}}\beta^{\frac{M}{2}}\exp\left[-\frac{\beta}{2}\|\mathbf{y}-\boldsymbol{\Phi}\mathbf{w}\|_2^2\right]\beta^{\frac{N}{2}}|\boldsymbol{\Psi}|^{\frac{1}{2}}\exp\left[-\sqrt{\beta}\sum_{i=1}^N\sqrt{\lambda_i}\,|w_i|\right]\mathcal{I}_{\mathbf{R}_+^N}(\mathbf{w})\right]\right\} \\ &= \text{argmin}_{\mathbf{w}}\left[\frac{\beta}{2}\|\mathbf{y}-\boldsymbol{\Phi}\mathbf{w}\|_2^2 + \sum_{i=1}^N\sqrt{\beta\lambda_i}\,|w_i| - \log\left(\mathcal{I}_{\mathbf{R}_+^N}(\mathbf{w})\right)\right]. \quad (42)\end{aligned}$$

$$q_i = [\,\boldsymbol{\sigma}_{\neg i}^T \quad 0\,]\,\boldsymbol{\Sigma}_i^{-1}\begin{bmatrix}\boldsymbol{\sigma}_{\neg i}^T \\ 0\end{bmatrix} = \boldsymbol{\sigma}_{\neg i}^T\boldsymbol{\Sigma}_{\neg i\neg i}^{-1}\boldsymbol{\sigma}_{\neg i} + \frac{\left(\boldsymbol{\sigma}_{\neg i}^T\boldsymbol{\Sigma}_{\neg i\neg i}^{-1}\boldsymbol{\sigma}_{\neg i}\right)^2}{\sigma_{ii} - \boldsymbol{\sigma}_{\neg i}^T\boldsymbol{\Sigma}_{\neg i\neg i}^{-1}\boldsymbol{\sigma}_{\neg i}}. \tag{47}$$

Then, by rearranging (47) the term $\boldsymbol{\sigma}_{\neg i}^T\boldsymbol{\Sigma}_{\neg i\neg i}^{-1}\boldsymbol{\sigma}_{\neg i}$ can be written as

$$\boldsymbol{\sigma}_{\neg i}^T\boldsymbol{\Sigma}_{\neg i\neg i}^{-1}\boldsymbol{\sigma}_{\neg i} = \frac{q_i\sigma_{ii}}{q_i + \sigma_{ii}}, \tag{48}$$

and from (32)

$$\sigma_{ii}^* = \sigma_{ii} - \frac{q_i\sigma_{ii}}{q_i + \sigma_{ii}}. \tag{49}$$

Define $\boldsymbol{\upsilon}_{\neg i} = \mathbf{w}_{\neg i} - \boldsymbol{\mu}_{\neg i}$ and

$$p_i = [\,\boldsymbol{\sigma}_{\neg i}^T \quad 0\,]\,\boldsymbol{\Sigma}_i^{-1}\begin{bmatrix}\boldsymbol{\upsilon}_{\neg i} \\ 0\end{bmatrix} = \boldsymbol{\sigma}_{\neg i}^T\boldsymbol{\Sigma}_{\neg i\neg i}^{-1}\boldsymbol{\upsilon}_{\neg i} + \frac{\boldsymbol{\sigma}_{\neg i}^T\boldsymbol{\Sigma}_{\neg i\neg i}^{-1}\boldsymbol{\sigma}_{\neg i}\boldsymbol{\sigma}_{\neg i}^T\boldsymbol{\Sigma}_{\neg i\neg i}^{-1}\boldsymbol{\upsilon}_{\neg i}}{\sigma_{ii} - \boldsymbol{\sigma}_{\neg i}^T\boldsymbol{\Sigma}_{\neg i\neg i}^{-1}\boldsymbol{\sigma}_{\neg i}}. \tag{50}$$

Then, solving for $\boldsymbol{\sigma}_{\neg i}^T\boldsymbol{\Sigma}_{\neg i\neg i}^{-1}\boldsymbol{\upsilon}_{\neg i}$, we get

$$\boldsymbol{\sigma}_{\neg i}^T\boldsymbol{\Sigma}_{\neg i\neg i}^{-1}\boldsymbol{\upsilon}_{\neg i} = \frac{p_i\sigma_{ii}}{q_i + \sigma_{ii}} \tag{51}$$

and (31) becomes

$$\mu_i^* = \mu_i + \frac{p_i\sigma_{ii}}{q_i + \sigma_{ii}}. \tag{52}$$

In summary, after obtaining $\boldsymbol{\Sigma}_i^{-1}$ from $\boldsymbol{\Sigma}^{-1}$, q_i, and p_i are computed from the first equations in (47) and (50), respectively. Then, σ_{ii}^* μ_i^*, $i = 1, 2, \ldots, N$ are efficiently computed from (49) and (52), respectively.

Appendix E
Relation to Variational Bayesian Inference and Other Methods

In this Appendix, we highlight the relation of the proposed BI-ICE algorithm with other known Bayesian inference methods and primarily with variational Bayesian inference, [23]–[25], [48]. To this end, we first apply the variational inference method to the proposed Bayesian model described in Section II. In variational inference, the joint posterior distribution of the model parameters $p(\mathbf{w}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \beta|\mathbf{y})$ is approximated by a variational distribution $Q(\mathbf{w}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \beta)$. Assuming posterior independence among the model parameters, this variational distribution factorizes as follows

$$Q(\mathbf{w}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \beta) = q(\mathbf{w})q(\boldsymbol{\gamma})q(\boldsymbol{\lambda})q(\beta) = q(\mathbf{w})\left(\prod_{i=1}^N q(\gamma_i)\right)\left(\prod_{i=1}^N q(\lambda_i)\right)q(\beta). \tag{53}$$

According to the variational Bayes methodology, [48, pp. 466], the factors in (53) can be computed by minimizing the Kullback—Leibler divergence between the approximate distribution $Q(\mathbf{w}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \beta)$ and the target distribution $p(\mathbf{w}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \beta|\mathbf{y})$. After some straightforward algebraic manipulations, it turns out that $q(\mathbf{w})$ is expressed as

$$q(\mathbf{w}) = \mathcal{N}_{\mathbf{R}_+^N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{54}$$

with

$$\boldsymbol{\mu} = \langle\beta\rangle_{q(\beta)}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{y} \text{ and } \boldsymbol{\Sigma} = \left(\langle\beta\rangle_{q(\beta)}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \langle\boldsymbol{\Lambda}\rangle_{q(\boldsymbol{\gamma})}\right)^{-1} \tag{55}$$

where $\langle\beta\rangle_{q(\beta)}$ denotes the mean value of β with respect to the distribution $q(\beta)$. For the rest factors, we have (56)–(57) shown at the bottom of the page, and

$$q(\lambda_i) = \Gamma\left(\lambda_i|r + 1, \frac{1}{2}\langle\gamma_i\rangle_{q(\gamma_i)} + \delta\right). \tag{58}$$

Equations (54)–(58) do not provide an explicit solution, since they depend on each other's factors. However, in principle, a solution may be reached iteratively, by initializing the required moments and then cycling through the model parameters, updating each distribution in turn. It may come as a surprise, but, although a different approach is used, the derived expressions resemble the conditional posterior distributions (21), (25), (26), and (28) employed in the iterative scheme of BI-ICE. Notice that both approaches share (a) the same type of distributions and (b) the updating of the same form of parameters. The only difference is that, in a variational Bayesian framework, the computation of the mean values of the model parameters require a blend of their first and second moments with respect to the approximate posterior distributions given in (54), (56)–(58), while this is not the case with BI-ICE (see (33), (34), (35) and (36)). As a result, the proposed BI-ICE can be considered as a first moments approximation of the variational Bayesian inference scheme, which is based on the factorization given in (53).

$$q(\beta) = \Gamma\left(\beta|\frac{M+N}{2} + \kappa, \frac{1}{2}\langle\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}\|^2\rangle_{q(\mathbf{w})} + \frac{1}{2}\langle\mathbf{w}^T\boldsymbol{\Lambda}\mathbf{w}\rangle_{q(\mathbf{w})q(\gamma)} + \theta\right) \tag{56}$$

$$q(\gamma_i) = \sqrt{\frac{\langle\lambda_i\rangle_{q(\lambda_i)}}{2\pi}}\gamma_i^{-\frac{1}{2}}\exp\left[-\frac{\langle\beta\rangle_{q(\beta)}\langle w_i^2\rangle_{q(\mathbf{w})}}{2}\frac{1}{\gamma_i} - \frac{1}{2}\langle\lambda_i\rangle_{q(\lambda_i)}\gamma_i + \sqrt{\langle\lambda_i\rangle_{q(\lambda_i)}\langle\beta\rangle_{q(\beta)}\langle w_i^2\rangle_{q(\mathbf{w})}}\right] \tag{57}$$

To elaborate further on the relation of BI-ICE to variational Bayes approximation, let us assume that in the variational framework $q(\mathbf{w})$ is factorized as $q(\mathbf{w}) = \prod_i^N q(w_i)$. Then, it can be shown that the posterior approximate distributions $q(\boldsymbol{\gamma})$, $q(\boldsymbol{\lambda})$ and $q(\beta)$ of the variational Bayes scheme remain exactly the same as in (57), (58) and (56), respectively, while $q(w_i)$ is expressed as

$$q(w_i) = \mathcal{N}_{\mathbf{R}_+^1}(w_i|\dot{\mu}_i, \dot{\sigma}_{ii}) \tag{59}$$

$$\dot{\mu}_i = \left(\phi_i^T\phi_i + \langle\frac{1}{\gamma_i}\rangle\right)^{-1}\phi_i^T\left(\mathbf{y} - \boldsymbol{\Phi}_{\neg i}\langle\mathbf{w}_{\neg i}\rangle\right) \tag{60}$$

$$\dot{\sigma}_{ii} = \langle\beta\rangle^{-1}\left(\phi_i^T\phi_i + \langle\frac{1}{\gamma_i}\rangle\right)^{-1} \tag{61}$$

where $\boldsymbol{\Phi}_{\neg i}$ is the matrix resulting from $\boldsymbol{\Phi}$ after removing its ith column. By superimposing (59)–(61) and (30)–(32) reveals that the posterior independence of w_i's assumed in the variational framework leads to a different updating mechanism compared to BI-ICE, in which such an assumption is not made. This means that the proposed scheme in (33) cannot result from a factorized approximation of the form $\prod_i^N q(w_i)$.

It is also worth noting that the motivation for the derivation of the BI-ICE algorithm has been the so-called Rao-Blackwellized Gibbs sampling scheme [49], [50]. In a Rao-Blackwellized Gibbs sampler with two random variables X, Z, the sequences $x_1, x_2, \ldots$ and $z_1, z_2, \ldots$ are generated first by sampling the conditional distributions $p(X|Z)$ and $p(Z|X)$, respectively, as in the conventional Gibbs sampler. Then, the conditional expectations $\mathrm{E}[X|z_i]$ and $\mathrm{E}[Z|x_i]$, $\forall i$ are computed and the sample means $\frac{1}{K}\sum_{i=1}^K \mathrm{E}[X|z_i]$ and $\frac{1}{K}\sum_{i=1}^K \mathrm{E}[Z|x_i]$ for large K are obtained. According to the Rao-Blackwell theorem [51], these estimates improve upon the original Gibbs sampler estimates $\frac{1}{K}\sum_{i=1}^K x_i$ and $\frac{1}{K}\sum_{i=1}^K z_i$, [32], [49]. Note that in the proposed iterative scheme, the conditional expectations of all involved parameters are computed as well. However, each one of them is now evaluated directly in each iteration, conditioned on the current values of the remaining conditional expectations.

Finally, it should be mentioned that the proposed BI-ICE algorithm resembles the iterative conditional modes (ICM) algorithm presented in [26]. As noted in [48, pp. 546], the ICM algorithm can be viewed as a "greedy" approximation to the Gibbs sampler, where instead of drawing a sample from each conditional distribution, the maximum of the conditional distribution is selected. The difference with the ICM method is that in BI-ICE the first order moment of the conditional posterior distributions is used instead of the maximum.

REFERENCES

[1] D. Landgrebe, "Hyperspectral image data analysis," *IEEE Signal Process. Mag.*, vol. 19, pp. 17–28, Jan. 2002.

[2] G. Shaw and D. Manolakis, "Signal processing for hyperspectral image exploitation," *IEEE Signal Process. Mag.*, vol. 19, pp. 12–16, Jan. 2002.

[3] N. Keshava and J. F. Mustard, "Spectral unmixing," *IEEE Trans. Signal Process.*, vol. 19, pp. 44–57, Jan. 2002.

[4] J. W. Boardman, "Automating spectral unmixing of AVIRIS data using convex geometry concepts," in *Proc. Summaries 4th Ann. JPL Airborne Geosci. Workshop*, Wash., DC, 1993, vol. 1, pp. 11–14.

[5] M. E. Winter, "N-FINDR: An algorithm for fast autonomous spectral end-member determination in hyperspectral data," *Proc. SPIE Imaging Spectrometry V*, vol. 3753, pp. 266–275, Jul. 1999.

[6] J. M. Nascimento and J. M. Bioucas-Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, pp. 898–910, Apr. 2005.

[7] D. C. Heinz and C. I. Chang, "Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, pp. 529–545, Mar. 2001.

[8] T. F. Coleman and Y. Li, "A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables," *SIAM J. Optimiz.*, vol. 6, pp. 1040–1058, 1996.

[9] N. Dobigeon, J.-Y. Tourneret, and C.-I. Chang, "Semisupervised linear spectral unmixing using a hierarchical Bayesian model for hyperspectral imagery," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 2684–2695, Jul. 2008.

[10] K. Themelis, A. A. Rontogiannis, and K. D. Koutroumbas, "Semisupervised hyperspectral unmixing via the weighted Lasso," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP'10)*, Dallas, TX, Mar. 2010.

[11] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, pp. 407–499, Feb. 2002.

[12] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, pp. 4655–4666, Dec. 2007.

[13] D. L. Donoho, Y. Tsaig, I. Drori, and J. L. Starck, Sparse Solution of Underdetermined Linear Equations by Stagewise Orthogonal Matching Pursuit Dep. Statist., Stanford Univ., CA, 2006.

[14] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2346–2356, Jun. 2008.

[15] M.-D. Iordache, J. Bioucas-Dias, and A. Plaza, "Unmixing sparse hyperspectral mixtures," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Cape Town, South Africa, Jul. 2009, vol. 4, pp. 85–88.

[16] J. Bioucas-Dias and M. Figueiredo, "Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing," in *Proc. IEEE Int. Workshop on Hyperspectral Image and Signal Process.: Evolution in Remote Sens. (WHISPERS'10)*, Reykjavik, Iceland, Jun. 2010.

[17] M. Figueiredo, "Adaptive sparseness for supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1150–1159, Sep. 2003.

[18] N. Dobigeon, A. Hero, and J.-Y. Tourneret, "Hierarchical Bayesian sparse image reconstruction with application to MRFM," *IEEE Trans. Image Process.*, vol. 18, no. 9, pp. 2059–2070, Sep. 2009.

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

[20] T. Park and C. George, "The Bayesian Lasso," *J. Amer. Statist. Assoc.*, vol. 103, no. 482, pp. 681–686, Jun. 2008.

[21] H. Zou, "The adaptive Lasso and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 101, pp. 1418–1429, Dec. 2006.

[22] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.

[23] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, pp. 183–233, Jan. 1999.

[24] H. Attias, "A variational Bayesian framework for graphical models," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2000, vol. 12, pp. 209–215.

[25] T. S. Jaakkola and M. I. Jordan, "Bayesian parameter estimation via variational methods," *Statist. Comput.*, vol. 10, pp. 25–37, Jan. 2000.

[26] J. Besag, "On the statistical analysis of dirty pictures," *J. Royal Statist. Soc. Ser. B (Methodological)*, vol. 48, pp. 259–302, Mar. 1986.

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

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

[29] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *J. Royal Statist. Soc., Ser. B*, vol. 36, no. 1, pp. 99–102, 1974.

[30] J. Bioucas-Dias, "Bayesian wavelet-based image deconvolution: a GEM algorithm exploiting a class of heavy-tailed priors," *IEEE Trans. Image Process.*, vol. 15, no. 4, pp. 937–951, Apr. 2006.

[31] M. Kyung, J. Gilly, M. Ghoshz, and G. Casella, "Penalized regression, standard errors, and Bayesian Lassos," *Bayesian Anal.*, vol. 5, pp. 369–412, Feb. 2010.

[32] G. Casella and E. I. George, "Explaining the Gibbs sampler," *Amer. Statist.*, vol. 46, pp. 167–174, Aug. 1992.
[33] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[34] H. Snoussi and J. Idier, "Bayesian blind separation of generalized hyperbolic processes in noisy and underdeterminate mixtures," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3257–3269, Sep. 2006.
[35] N. L. Johnson and S. Kotz, *Continuous Univariate Distributions-1*. New York: Wiley , 1970.
[36] C. P. Robert, "Simulation of truncated normal variables," *Statist. Comput.*, vol. 5, pp. 121–125, 1995.
[37] G. Rodriguez-Yam, R. Davis, and L. Scharf, "Efficient Gibbs sampling of truncated multivariate normal with application to constrained linear regression," Columbia Univ., New York, 2004.
[38] M.-D. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Sparse unmixing of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 6, pp. 2014–2039, Jun. 2011.
[39] A. Bruckstein, M. Elad, and M. Zibulevsky, "On the uniqueness of nonnegative sparse solutions to underdetermined systems of equations," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 4813–4820, Nov. 2008.
[40] G. H. Golub and C. F. Van Loan, *Matrix Computations (Johns Hopkins Studies in Mathematical Sciences)*, 3rd ed. Baltimore, MD: The Johns Hopkins Univ. Press, 1996.
[41] R. N. Clark, G. A. Swayze, R. Wise, K. E. Livo, T. M. Hoefen, R. F. Kokaly, and S. J. Sutley, USGS Digital Spectral Library, 2007 [Online]. Available: http://speclab.cr.usgs.gov/spectral.lib06/ds231/datatable.html
[42] AVIRIS Free Standard Data Products [Online]. Available: http://aviris.jpl.nasa.gov/html/aviris.freedata.html
[43] R. N. Clark *et al.*, "Imaging Spectroscopy: Earth and Planetary Remote Sensing With the Usgs Tetracorder and Expert Systems," *J. Geophys. Res.*, vol. 108, no. E12, pp. 5-1–5-44, Dec. 1993.
[44] L. Miao and H. Qi, "Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 3, pp. 765–777, Mar. 2007.
[45] T.-H. Chan, C.-Y. Chi, Y.-M. Huang, and W.-K. Ma, "A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4418–4432, Nov. 2009.
[46] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*. New York: Academic, 1980.
[47] L. L. Scharf, *Statistical Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1991.
[48] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. New York: Springer-Verlag, 2006.
[49] A. E. Gelfand and A. F. M. Smith, "Sampling-based approaches to calculating marginal densities," *J. Amer. Statist. Assoc.*, vol. 85, pp. 398–409, Jun. 1990.
[50] J. S. Liu, W. H. Wong, and A. Kong, "Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes," *Biometrika*, vol. 81, no. 1, pp. 27–40, 1994.
[51] E. L. Lehmann and G. Casella, *Theory of Point Estimation*. New York: Springer, 1998.

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.

He is currently pursuing the Ph.D. degree in signal processing at the University of Athens. Since 2007 he is a research associate with the Institute for Space Applications and Remote Sensing of the National Observatory of Athens, Greece. His research interests are in the area of Bayesian analysis with application to hyperspectral image processing.

Athanasios A. Rontogiannis (M'93) was born in Lefkada Island, Greece, in 1968. He received the Diploma degree in electrical engineering from the National Technical University of Athens (NTUA), Greece, in 1991, the M.A.Sc. degree in electrical and computer engineering from the University of Victoria, Canada, in 1993, and the Ph.D. degree in communications and signal processing from the University of Athens, Greece, in 1997.

From 1998 to 2003, he was with the University of Ioannina, where he was a lecturer in informatics since June 2000. In 2003, he joined the Institute for Space Applications and Remote Sensing (ISARS) of the National Observatory of Athens (NOA), where he is currently a Senior Researcher. His research interests are in the general areas of signal processing and wireless communications.

Dr. Rontogiannis has been a graduate and a postgraduate scholar of the Greek State Scholarship Foundation from 1994 to 1999. Currently, he serves at the Editorial Boards of the EURASIP *Journal on Advances in Signal Processing*, Springer (since 2008) and the EURASIP *Signal Processing Journal*, Elsevier (since 2011). He is a member of the IEEE Signal Processing and Communication Societies and the Technical Chamber of Greece.

Konstantinos D. Koutroumbas received the B.Sc. degree from the Department of Computer Engineering and Informatics of the University of Patras in 1989, the M.Sc. degree in advanced methods in computer science from the Queen Mary College of the University of London in 1990, and the Ph.D. degree from the Department of Informatics and Telecommunications from the University of Athens in 1995.

Since 2001, he has been with the Institute for Space Applications and Remote Sensing of the National Observatory of Athens, Greece, where he currently is a Senior Researcher. His research interests include mainly pattern recognition, time series estimation and their application to remote sensing and to the estimation of characteristic quantities of the upper atmosphere. He is a coauthor of the books *Pattern Recognition* (1st, 2nd, 3rd, 4th editions) and *Introduction to Pattern Recognition: A MATLAB Approach*. He has more than 2500 citations in his work.