Bayesian signal processing techniques for hyperspectral image unmixing

Themelis Konstantinos*

Department of Informatics and Telecommunications
National and Kapodistrian University of Athens
TYPA Buildings, University Campus, 15784 Athens, Greece
themelis@space.noa.gr

Abstract. In this dissertation the problem of semi-supervised spectral unmixing is studied. We describe a framework of Bayesian techniques that take into account the special characteristics of hyperspectral data and exploit the prior knowledge of the model's constraints. The assumption of sparsity on the unmixing process is introduced, which leads to the development of efficient schemes for spectral unmixing. Besides, a novel method for Bayesian inference is proposed, termed BI-ICE, which is computationally efficient, and can be considered as a first-moments approximation to variational inference methods. Experimental results conducted both on simulated and real hyperspectral data are presented that illustrate the robust estimation performance of the proposed sparsity-inducing methods.

Keywords: Hyperspectral imagery, Spectral unmixing, Sparse linear regression, Bayesian analysis, Variational inference.

1 Introduction

Over the past decades hyperspectral image analysis has emerged as one of the fastest growing technologies in the field of remote sensing. Hyperspectral imagery refers to the process of remotely obtaining data about an object (in our case a geographical area), with the use of hyperspectral sensors. Hyperspectral sensors have the ability to sample the electromagnetic spectrum in hundreds of continuous spectral bands. As a consequence, each pixel of a hyperspectral image is represented by a vector, where each coefficient is a measurement of reflectance at a respective wavelength.

The immense growth of hyperspectral imaging applications has also been accompanied by the development of numerous techniques to effectively process the overwhelming amount of collected data. Sophisticated algorithms have been proposed in the literature that either ameliorate the shortcomings of hyperspectral data, or focus on the exploitation of their informational content. Among all possible signal processing applications that exploit hyperspectral data, in this dissertation we are more interested in the thematic area of spectral unmixing.

The process of hyperspectral unmixing is described by two major steps: (a) the endmember extraction step, and (b) the inversion process. In the endmember

* Dissertation Advisor: Prof. Sergios Theodoridis.

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), [1], the N-FINDR algorithm, [2], and the vertex component analysis (VCA) method, [3]. 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 non-negative 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, [4], or numerical optimization methods, e.g., [5]. Bayesian methods have also been proposed for the problem, e.g., the Gibbs sampling scheme applied to the hierarchical Bayesian model of [6]. Semi-supervised unmixing, [6, 7], which is considered in this dissertation, assumes that the endmembers' spectral signatures are available. The objective of semi-supervised unmixing is (a) to determine how many and which endmembers are present in the mixed pixel under study and (b) to estimate their corresponding abundances.

2 Work contributions

The scientific contributions that appear in the present work exhibit a Bayesian treatment for the problem of supervised spectral unmixing. A major advantage of the Bayesian approach is that it provides a flexible framework to represent the probabilistic mechanisms of data generation and our prior information about it. To this end, parametric probabilistic models are considered, and appropriate prior distributions are employed to capture the uncertainties of the model parameters.

In the context of hyperspectral signal processing, we were greatly influenced by the wide applicability of Bayesian methods. Our first contribution is a Bayesian maximum a posteriori (MAP) estimator, published in [8], which is specifically designed to account for the convex constraints of the abundance estimation problem. Utilizing the simplicity of the Gaussian distribution and the symmetry of constraints, closed form expressions are derived for the modes of the posterior distribution of the abundances. Following this path an efficient estimator is proposed, which has almost similar estimation performance, but is computationally more efficient than quadratic programming methods that address the same constrained estimation problem.

The applicability of our proposed schemes has also been of concern. A representative example is the case study on the unmixing of real hyperspectral data collected from the OMEGA spectrometer, [9], aboard the Mars Express mission. The objective of the OMEGA spectrometer is to collect information which will help to determine the mineral composition of the Mars surface. This is an interesting application for spectral unmixing. Three methods are considered and compared in the unmixing process: (a) the ENVI-SVD method, which is commercially available through the popular remote sensing software ENVI, (b) the MAP estimator of the preceding paragraph, and (c) a quadratic programming method, [5], which is an iterative Newton method. The results of this comparison

are available in [10, 11], where it is seen that the MAP estimator can provide reliable results and outperform existing methods.

In the sequel, a significant amount of efforts was invested to leverage sparse signal processing techniques for spectral unmixing. The primary assumption is that only a small number of endmembers will be mixed in a single pixel, and hence, the abundance estimation problem will inherently have a sparse solution. To the best of our knowledge, we were among the first to introduce the notion of sparsity to the problem of spectral unmixing. This consideration is exhibited in the publications [7, 12, 13]. In both publications a variant of the least absolute shrinkage and selection operator (lasso), [14], was selected to impose sparsity. In [7], the adaptively weighted lasso, [15], is utilized and special manipulation is provided for the constraints of the problem. To force the nonnegativity constraint, the optimization problem is solved using a modified LARS algorithm, which retrieves only nonnegative solutions. Moreover, the additivity constraint is included in the quadratic loss function of least squares, by means of an appropriate extra linear equation.

Another contribution of this thesis is the development of a Bayesian hierarchical model analogous to the adaptive lasso, [15]. In the proposed Bayesian setup, independent Laplace priors are employed by the model to correspond to the weighted ℓ_1 norm penalization of the adaptive lasso. Besides, the nonnegativity constraint is incorporated to the model by a truncation operation on the prior distributions. A novel method for Bayesian inference is then developed, termed as Bayesian inference iterative conditional expectations algorithm (BI-ICE). BI-ICE appears to be a first-moments approximation to variational approximation methods, and is summarized in Section 3. The proposed Bayesian approach is analytically described in [13].

Finally, a lot of effort has also been invested into the research and development of a sparse reconstruction algorithm, in the framework of Bayesian compressive sensing. The previous Bayesian set-up has been adopted to promote sparse solutions to an underdetermined system of linear equations. To perform Bayesian inference, a recently proposed sub-optimal, type-II maximum likelihood algorithm was adjusted to fit the needs of the present framework. The resulting incremental-type algorithm has superior performance when compared to other Bayesian compressive sensing methods, as illustrated in Chapter 4 of the dissertation.

3 Semi-supervised hyperspectral unmixing via a novel Bayesian approach

An interesting perspective of the semi-supervised 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., [16–18], in semi-supervised unmixing. A number of such semi-supervised unmixing techniques has been recently proposed in [7, 19, 20], 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.

Let $\mathbf{y}$ be a $M \times 1$ hyperspectral image pixel vector, where M is the number of spectral bands. Also let $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_N]$ 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 $\mathbf{y}$.

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

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

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 non-negativity constraint and a sum-to-one (additivity) constraint. Based on this formulation, a semi-supervised hyperspectral unmixing technique is introduced, where the endmember matrix $\boldsymbol{\Phi}$ is assumed to be known a priori. As mentioned before, each column of $\boldsymbol{\Phi}$ contains the spectral signature of a single material, and its elements are non-negative, since they represent reflectance values. The mixing matrix $\boldsymbol{\Phi}$ can either stem from a spectral library or it can be determined using an endmember extraction technique, e.g., [3]. 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 semi-supervised unmixing, we are interested in estimating the abundance vector $\mathbf{w}$ for each image pixel, which is non-negative and sparse, with ξ out of its N entries being non-zero.

3.1 Hierarchical Bayesian model

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

$$p(\mathbf{y}|\mathbf{w}, \beta) = \mathcal{N}(\mathbf{y}|\boldsymbol{\Phi}\mathbf{w}, \beta^{-1}\mathbf{I}_M) = (2\pi)^{-\frac{M}{2}}\, \beta^{\frac{M}{2}} \exp\left[-\frac{\beta}{2}\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}\|_2^2\right]. \tag{3}$$

Accounting for the non-negativity property of $\mathbf{w}$, and assuming that all w_i's are i.i.d., a normal distribution truncated on the non-negative orthant $\mathbf{R}_+^N$ of the N-dimensional Euclidean space $\mathcal{R}^N$ is assigned to $\mathbf{w}$, i.e.,

$$\begin{aligned} p(\mathbf{w}|\boldsymbol{\gamma},\beta) &= \prod_{i=1}^{N}\left[\mathcal{N}(w_i|0,\frac{\gamma_i}{\beta})\,\mathbf{I}_{\mathbf{R}_+^1}(w_i)\right] \\ &= 2^N\,(2\pi)^{-\frac{N}{2}}\,\beta^{\frac{N}{2}}\,|\boldsymbol{\Lambda}|^{\frac{1}{2}}\exp\left[-\frac{\beta}{2}\mathbf{w}^T\boldsymbol{\Lambda}\mathbf{w}\right]\mathbf{I}_{\mathbf{R}_+^N}(\mathbf{w}) \end{aligned} \tag{4}$$

where $\mathbf{R}_+^1$ is the set of non-negative real numbers, $\mathbf{I}_{\mathbf{R}_+^N}(\cdot)$ is the indicator function[1] for $\mathbf{R}_+^N$, $\boldsymbol{\gamma}=[\gamma_1,\gamma_2,\ldots,\gamma_N]^T$ is a $N\times 1$ vector of hyperparameters, $\gamma_i\geq 0$, and $\boldsymbol{\Lambda}^{-1}=\mathrm{diag}(\boldsymbol{\gamma})$.

For β, a conjugate Gamma prior with respect to the Gaussian likelihood of (3) is selected, expressed as

$$p(\beta|\kappa,\theta)=\Gamma(\beta|\kappa,\theta)=\frac{\theta^\kappa}{\Gamma(\kappa)}\beta^{\kappa-1}\exp\left[-\theta\beta\right], \tag{5}$$

where $\beta\geq 0$, and $\kappa\geq 0$, $\theta\geq 0$ are the distribution parameters.

We extend the model of [21, 22], by assigning an independent Gamma distribution to every γ_i, each parameterized by a distinct hyperparameter λ_i, i.e.,

$$p(\gamma_i|\lambda_i)=\Gamma(\gamma_i|1,\frac{\lambda_i}{2})=\frac{\lambda_i}{2}\exp\left[-\frac{\lambda_i}{2}\gamma_i\right],\ i=1,2,\ldots,N, \tag{6}$$

Then, the combination of the hierarchical priors given in (4) and (6) leads to a sparsity-promoting, non-negative (truncated) Laplace distribution for $\mathbf{w}$ (this formulation gives rise to the so-called Bayesian lasso [22]). This distribution can be obtained by marginalizing the hyperparameter vector $\boldsymbol{\gamma}$ from the model, i.e.,

$$p(\mathbf{w}|\boldsymbol{\lambda},\beta)=\int p(\mathbf{w}|\boldsymbol{\gamma},\beta)p(\boldsymbol{\gamma}|\boldsymbol{\lambda})d\boldsymbol{\gamma}=\beta^{\frac{N}{2}}\,|\boldsymbol{\Psi}|^{\frac{1}{2}}\exp\left[-\sqrt{\beta}\sum_{i=1}^{N}\sqrt{\lambda_i}\,|w_i|\right]\mathbf{I}_{\mathbf{R}_+^N}(\mathbf{w}), \tag{7}$$

where $\boldsymbol{\lambda}=[\lambda_1,\lambda_2,\ldots,\lambda_N]^T$ and $\boldsymbol{\Psi}=\mathrm{diag}(\boldsymbol{\lambda})$. The motivation to use a hyperparameter vector $\boldsymbol{\lambda}$ instead of a single λ for all γ_i's as in [22, 21], is to form a hierarchical Bayesian analogue to the adaptive lasso, proposed in [15]. Indeed, it can be shown, that the maximum a posteriori (MAP) estimator of $\mathbf{w}$, which is distributed according to (7), is the solution to the following optimization problem,

$$\tilde{\mathbf{w}}=\arg\min_{\mathbf{w}}\left\{\frac{\beta}{2}\left\|\mathbf{y}-\boldsymbol{\Phi}\mathbf{w}\right\|_2^2+\sum_{i=1}^{N}\alpha_i|w_i|\right\},\ \text{s.t. }\mathbf{w}\in\mathbf{R}_+^N, \tag{8}$$

which, excluding the non-negativity constraint, coincides with the definition of the adaptive lasso, [15].

[1] $\mathbf{I}_{\mathbf{R}_+^N}(\mathbf{x})=1(0),\text{if }\mathbf{x}\in\mathbf{R}_+^N(\mathbf{x}\notin\mathbf{R}_+^N)$.

It is obvious from (7) that the quality of the endmember selection procedure depends on the tuning parameter vector $\boldsymbol{\lambda}$. We choose to infer the hyperparameter vector $\boldsymbol{\lambda}$ from the data, by assuming a Gamma hyperprior for each element of $\boldsymbol{\lambda}$,

$$p(\lambda_i|r,\delta) = \Gamma(\lambda_i|r,\delta) = \frac{\delta^r}{\Gamma(r)}{\lambda_i}^{r-1}\exp\left[-\delta\lambda_i\right], \quad i = 1,2,\ldots,N \tag{9}$$

where r and δ are hyperparameters, with $r \geq 0$ and $\delta \geq 0$. Both Gamma priors of β, in (5), and λ_i, in (9), are flexible enough to express prior information, by properly tuning their hyperparameters. In this paper, the hyperparameters $\kappa, \theta, r, \delta$ are set to zero as in [6, 21], forming non-informative (Jeffreys') priors, although other values can, in principle, be selected.

3.2 Bayesian inference

As it is common in Bayesian inference, the estimation procedure is based on the computation of the joint posterior distribution of the parameters. For the model presented in Section 3.1, this posterior is

$$p\left(\mathbf{w},\beta,\boldsymbol{\gamma},\boldsymbol{\lambda}|\mathbf{y}\right) = \frac{p\left(\mathbf{y}|\mathbf{w},\beta\right)p\left(\mathbf{w}|\beta,\boldsymbol{\gamma}\right)p\left(\boldsymbol{\gamma}|\boldsymbol{\lambda}\right)p\left(\boldsymbol{\lambda}\right)p\left(\beta\right)}{p(\mathbf{y})}, \tag{10}$$

which is intractable, because $p(\mathbf{y})$ cannot be computed analytically. To overcome this obstacle, a Markovian Gibbs sampling strategy can be followed, in which the conditional posterior distributions of the associated parameters are utilized.

Posterior conditional distributions In the following, analytical expressions are derived for the posterior conditional distributions of the model parameters $\mathbf{w}$, $\boldsymbol{\gamma}$, $\boldsymbol{\lambda}$ and β. Starting with $\mathbf{w}$, it can be easily shown that its posterior conditional density is the multivariate Gaussian, truncated in $\mathbf{R}_+^N$,

$$p(\mathbf{w}|\mathbf{y},\boldsymbol{\gamma},\boldsymbol{\lambda},\beta) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu},\boldsymbol{\Sigma})\mathbf{I}_{\mathbf{R}_+^N} = \mathcal{N}_{\mathbf{R}_+^N}(\mathbf{w}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \tag{11}$$

where $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ are respectively expressed as follows,

$$\boldsymbol{\Sigma} = \beta^{-1}\left[\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{\Lambda}\right]^{-1}, \qquad \boldsymbol{\mu} = \beta\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{y}. \tag{12}$$

The posterior conditional for the precision parameter β, is easily shown to be a Gamma distribution, i.e.,

$$p(\beta|\mathbf{y},\mathbf{w},\boldsymbol{\gamma},\boldsymbol{\lambda}) = \Gamma\left(\beta|\frac{M+N}{2} + \kappa, \frac{1}{2}\|\mathbf{y}-\boldsymbol{\Phi}\mathbf{w}\|_2^2 + \theta + \frac{1}{2}\mathbf{w}^T\boldsymbol{\Lambda}\mathbf{w}\right) \tag{13}$$

Straightforward computations yield that the conditional distribution of γ_i given $\mathbf{y}, w_i, \lambda_i, \beta$ is expressed as

$$p(\gamma_i|\mathbf{y},w_i,\lambda_i,\beta) = \left(\frac{\lambda_i}{2\pi}\right)^{\frac{1}{2}}{\gamma_i}^{-\frac{1}{2}}\exp\left[-\frac{\beta w_i^2}{2\gamma_i} - \frac{\lambda_i}{2}\gamma_i + \sqrt{\beta\lambda_i}\,|w_i|\right], i = 1,2,\ldots,N \tag{14}$$

Finally, the conditional posterior of λ_i given $\mathbf{y}, w_i, \gamma_i, \beta$ also turns out to be a Gamma distribution,

$$p(\lambda_i|\mathbf{y}, w_i, \gamma_i, \beta) = \mathit{\Gamma}\left(\lambda_i|1+r, \frac{\gamma_i}{2}+\delta\right),\ i = 1, 2, \ldots, N. \tag{15}$$

In the sequel, 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, (11), (13), (14) and (15), as explained in Section 3.2. Thus, a novel scheme iterating among the conditional means of $\mathbf{w}$, β, γ_i and λ_i arises, which will be 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.

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

Expectation of $p(\mathbf{w}|\mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \beta)$ $\mathbf{w}$: As shown in (11), $p(\mathbf{w}|\mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \beta)$ is a multivariate Gaussian distribution, truncated in $\mathbf{R}_+^N$. In the one-dimensional case, the expectation of the truncated Gaussian distribution in $\mathbf{R}_+^1$ can be computed as

$$x \sim \mathcal{N}_{\mathbf{R}_+^1}(x|\mu^*, \sigma^*) \Rightarrow \mathrm{E}[x] = \mu^* + \frac{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{\mu^{*2}}{\sigma^{*2}}\right)}{1-\frac{1}{2}\mathrm{erfc}\left(\frac{\mu^*}{\sqrt{2}\sigma^*}\right)}\sigma^*, \tag{16}$$

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 [23], 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

$$w_i|\mathbf{w}_{\neg i} \sim \mathcal{N}_{\mathbf{R}_+^1}(w_i|\mu_i^*, \sigma_{ii}^*) \tag{17}$$

$$\mu_i^* = \mu_i + \boldsymbol{\sigma}_{\neg i}^T \boldsymbol{\Sigma}_{\neg i \neg i}^{-1}(\mathbf{w}_{\neg i} - \boldsymbol{\mu}_{\neg i}) \tag{18}$$

$$\sigma_{ii}^* = \sigma_{ii} - \boldsymbol{\sigma}_{\neg i}^T \boldsymbol{\Sigma}_{\neg i \neg i}^{-1} \boldsymbol{\sigma}_{\neg i}, \tag{19}$$

where matrix $\boldsymbol{\Sigma}_{\neg i \neg i}$ is formed by removing the ith row and the ith column from $\boldsymbol{\Sigma}$, the $(N-1)\times 1$ vector $\boldsymbol{\sigma}_{\neg i}$ is the ith column of $\boldsymbol{\Sigma}$ after removing its ith element σ_{ii} and μ_i is the ith element of $\boldsymbol{\mu}$. 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 j-th iteration, $j = 1, 2, \ldots$, of this procedure is

described as follows[2]

$$\begin{aligned} &1.\ w_1^{(j)} = \mathrm{E}[p(w_1|w_2^{(j-1)}, w_3^{(j-1)}, \ldots, w_N^{(j-1)})] \\ &2.\ w_2^{(j)} = \mathrm{E}[p(w_2|w_1^{(j)}, w_3^{(j-1)}, \ldots, w_N^{(j-1)})] \\ &\qquad\vdots \\ &N.\ w_N^{(j)} = \mathrm{E}[p(w_N|w_1^{(j)}, w_2^{(j)}, \ldots, w_{N-1}^{(j)})] \end{aligned} \tag{20}$$

This procedure is repeated iteratively until convergence. Experimental results have shown that the iterative scheme in (20) converges to the mean of $\mathbf{w} \sim \mathcal{N}_{\mathbf{R}_+^N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ after a few iterations.

Expectation of $p(\beta|\mathbf{y}, \mathbf{w}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$: The mean value of the Gamma distribution of (13) is given by

$$\mathrm{E}\left[p(\beta|\mathbf{y}, \mathbf{w}, \boldsymbol{\gamma}, \boldsymbol{\lambda})\right] = \frac{\frac{M+N}{2} + \kappa}{\frac{1}{2}\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}\|_2^2 + \theta + \frac{1}{2}\mathbf{w}^T\boldsymbol{\Lambda}\mathbf{w}} \tag{21}$$

Expectation of $p(\gamma_i|\mathbf{y}, w_i, \lambda_i, \beta)$: It can be shown that the expectation of (14) is expressed as

$$\mathrm{E}\left[p(\gamma_i|\mathbf{y}, w_i, \lambda_i, \beta)\right] = \left(\frac{2\lambda_i}{\pi}\right)^{\frac{1}{2}} \left(\frac{\beta w_i^2}{\lambda_i}\right)^{\frac{3}{4}} \exp\left[\sqrt{\beta\lambda_i}\,|w_i|\right] K_{3/2}\left(\sqrt{\beta\lambda_i}\,|w_i|\right), \tag{22}$$

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

Expectation of $p(\lambda_i|\mathbf{y}, w_i, \gamma_i, \beta)$: Finally, the mean value of the Gamma distribution (15) is

$$\mathrm{E}\left[\lambda_i|\mathbf{y}, w_i, \gamma_i, \beta)\right] = \frac{1+r}{\frac{1}{2}\gamma_i + \delta}. \tag{23}$$

The basic steps of the proposed BI-ICE algorithm are summarized in Table 1. Regarding the updating of parameter $\mathbf{w}^{(t)}$, an auxiliary variable $\mathbf{v}$ has been utilized in Table 1. 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 (20). The resulting value of $\mathbf{v}$ is assigned to $\mathbf{w}^{(t)}$. The rationale behind this choice is that for $\boldsymbol{\Sigma}$ diagonal (which happens when the columns of $\boldsymbol{\Phi}$ are orthogonal), it easily follows from (18), (19) that the w_i's in (20) 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 (20) is also sufficient in the framework of the BI-ICE algorithm.

A basic advantage of the proposed Bayesian approach, which is the Bayesian analogue to the adaptive lasso, is that all parameters are naturally estimated

[2] 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 1. 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 $\boldsymbol{\Sigma}^{(t)}, \boldsymbol{\mu}^{(t)}$ using (12)
  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 (18), (19), and (16)
  Compute $v_2^{(1)} = \mathrm{E}\left[p(v_2|v_1^{(1)}, v_3^{(0)}, \ldots, v_N^{(0)})\right]$,
  using (18), (19), and (16)
  $\vdots$
  Compute $v_N^{(1)} = \mathrm{E}\left[p(v_N|v_1^{(1)}, v_2^{(1)}, \ldots, v_{N-1}^{(1)})\right]$,
  using (18), (19), and (16)
  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 (21)
- 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 (22)
- 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 (23)
endfor

from the data. In contrast, deterministic algorithms for solving the lasso, e.g. [15], face the problem of fine-tuning specific parameters (corresponding to $\boldsymbol{\lambda}$ of our model), that control the sparsity of the solution. As shown in the simulations presented in the next section, the BI-ICE algorithm converges very fast, and retains the sparsity of the solution. It has been further observed that by initializing $\mathbf{w}$ with $\boldsymbol{\mu}$, a single cycle is sufficient for the inner sampler to converge. The computational complexity of the proposed method can be further reduced by avoiding the explicit computation of the matrices $\boldsymbol{\Sigma}_{\neg i \neg i}^{-1}$ in (18), (19).

4 Experimental results and discussion

The performance of the BI-ICE algorithm is illustrated by unmixing a synthetic hyperspectral image, using data from the USGS spectral library. Specifically, 30 endmembers were selected from the library, to construct a 453×220 endmember matrix, having condition number 36.182×10^6. The number of disparate endmembers composing a single pixel varied between one (pure pixel) and five, whereas the abundances were generated according to a Dirichlet distribution, so as to satisfy the positivity and sum-to-one constraints. The observations were corrupted by Gaussian noise, whose variance was determined by the SNR level. First, the fast convergence and sparse estimations of $\mathbf{w}$ exhibited by the new algorithm are depicted in Fig. 1a. In this experiment, a pixel with three non-

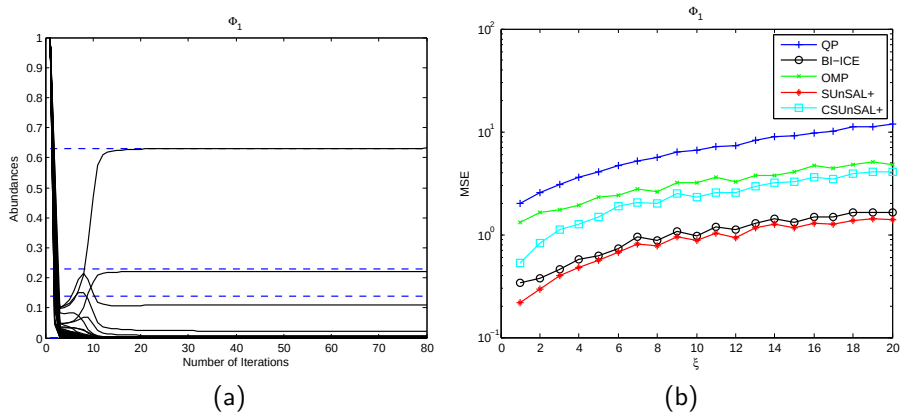

Fig. 1. (a) Estimation of the entries of the sparse vector $\mathbf{w}$, as BI-ICE progresses. The algorithm is applied to simulated data, generated using a highly correlated matrix of spectral data. White noise is added (SNR = 25 dB). Dashed lines: true values. Solid lines: estimated values. (b) MSE 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)

zero 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. 1a 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, [5], (c) the orthogonal matching pursuit (OMP) algorithm, [17], which is a widely used, greedy, sparsity promoting algorithm, (d) the sparse unmixing by variable splitting and augmented Lagrangian (SUnSAL) algorithm, [20, 24], and (e) the constrained version of SUnSAL, CSUnSAL, which solves the constrained version of the lasso operator, (see also [24] 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 [20]. Based on the following metric, $\text{MSE} = \text{E}\left[\frac{\|\mathbf{w}-\tilde{\mathbf{w}}\|_2^2}{\|\mathbf{w}\|_2^2}\right]$, where $\mathbf{w}$ and $\tilde{\mathbf{w}}$ 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. 1b. 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 [24]). 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 by-products 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.

5 Conclusions

In this dissertation we have presented a general framework for semi-supervised Bayesian learning, comprised of soft-constraint Bayesian estimation methods, sparsity-promoting hierarchical Bayesian models and novel methods for Bayesian inference. Under the assumption of the linear mixing model, spectral unmixing has been formulated as a standard linear regression problem, where the parameters of interest are the abundance fractions of the endmembers spectra in each pixel of a hyperspectral scene. A Bayesian MAP estimator has been suitably adjusted to address the abundance estimation problem, taking into account the constraints of nonnegativity and full additivity. To the best of our knowledge, we were among the first to introduce sparsity in the context of spectral unmixing. In order to induce sparsity, our research efforts have focused around the lasso and its variants. First, we have successfully applied the adaptively weighted lasso to the unmixing problem. A fully Bayesian treatment for the unmixing problem was further developed, wherein a suitably selected hierarchical Bayesian model and an efficient method to perform Bayesian inference are proposed. Finally, we attempt to demonstrate the potential of the hierarchical Bayesian model by considering a sparse image reconstruction problem. To this end, a fast, suboptimal, type-II maximum likelihood algorithm has been developed.

References

1. J. W. Boardman, Automating spectral unmixing of AVIRIS data using convex geometry concepts, in: Proc. Summaries 4th Annu. JPL Airborne Geosci. Workshop, Vol. 1, Washington, DC, 1993, pp. 11–14.
2. M. E. Winter, N-FINDR: an algorithm for fast autonomous spectral end-member determination in hyperspectral data, in: Proc. SPIE Imaging Spectrometry V, Vol. 3753, 1999, pp. 266 –275.
3. J. M. P. Nascimento, J. M. Bioucas-Dias, Vertex component analysis: A fast algorithm to unmix hyperspectral data, IEEE Trans. Geosci. Remote Sensing 4 (2005) 898–910.
4. D. C. Heinz, C. I. Chang, Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery, IEEE Trans. Geosci. Remote Sensing 39 (2001) 529–545.

5. T. F. Coleman, Y. Li, A reflective Newton method for minimizing a quadratic function subject to bounds on some of the variables, SIAM Journal on Optimization 6 (1996) 1040–1058.
6. N. Dobigeon, J.-Y. Tourneret, C.-I. Chang, Semi-supervised linear spectral unmixing using a hierarchical Bayesian model for hyperspectral imagery, IEEE Trans. Signal Process. 56 (7) (2008) 2684–2695.
7. K. E. Themelis, A. A. Rontogiannis, K. D. Koutroumbas, Semi-supervised hyperspectral unmixing via the weighted Lasso, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'10), Dallas, Texas, 2010.
8. K. E. Themelis, A. A. Rontogiannis, A soft constrained MAP estimator for supervised hyperspectral signal unmixing, in: Proc. of European Signal Processing Conference, (EUSIPCO), 2008.
9. J. P. Bibring, A. Soufflot, M. Berthe', Y. Langevin, B. Gondet, P. Drossart, M. Bouye', M. Combes, P. Puget, G. Bellucci, et al., OMEGA: Observatoire pour la Mine'ralogie, l'Eau, les Glaces et l'Activit, Vol. SP-1240, ESA, 2004, pp. 37–50.
10. K. E. Themelis, A. A. Rontogiannis, O. Sykioti, K. D. Koutroumbas, I. Daglis, F. Schmidt, On the unmixing of MEx/OMEGA hyperspectral data, in: Proc. European Planetary Science Congress (EPSC), 2010.
11. K. E. Themelis, F. Schmidt, O. Sykioti, A. A. Rontogiannis, K. D. Koutroumbas, I. A. Daglis, On the unmixing of MEx/OMEGA hyperspectral data, Planetary and Space Science, Elsevier (2011) –Accepted.
12. K. E. Themelis, A. A. Rontogiannis, K. D. Koutroumbas, Sparse semi-supervised hyperspectral unmixing using a novel iterative bayesian inference algorithm, in: 19th European Signal Processing Conference (EUSIPCO), 2011, pp. 1165–1169.
13. K. E. Themelis, A. A. Rontogiannis, K. D. Koutroumbas, A novel hierarchical Bayesian approach for sparse semi-supervised hyperspectral unmixing, IEEE Trans. Signal Process. 60 (2) (2012) 585–599.
14. R. Tibshirani, Regression shrinkage and selection via the Lasso, Journal of the Royal Statistical Society 58 (1) (1996) 267–288.
15. H. Zou, The adaptive Lasso and its oracle properties, Journal of the American Statistical Association 101 (2006) 1418–1429.
16. B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Annals of Statistics 32 (2002) 407–499.
17. J. Tropp, A. Gilbert, Signal recovery from random measurements via orthogonal matching pursuit, IEEE Trans. Inf. Theory 53 (2007) 4655 –4666.
18. S. Ji, Y. Xue, L. Carin, Bayesian compressive sensing, IEEE Trans. Signal Process. 56 (6) (2008) 2346–2356.
19. M.-D. Iordache, J. Bioucas-Dias, A. Plaza, Unmixing sparse hyperspectral mixtures, in: Geoscience and Remote Sensing Symposium, 2009 IEEE International, IGARSS 2009, Vol. 4, Cape Town, 2009, pp. 85 –88.
20. J. Bioucas-Dias, M. Figueiredo, Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing, in: Proc. IEEE International Workshop on Hyperspectral Image and Signal Processing: evolution in Remote Sensing (WHISPERS'10), Reykjavik, Iceland, 2010.
21. S. D. Babacan, R. Molina, A. Katsaggelos, Bayesian compressive sensing using Laplace priors, IEEE Trans. Image Process. 19 (1) (2010) 53 –63.
22. T. Park, C. G., The Bayesian Lasso, Journal of the American Statistical Association 103 (482) (2008) 681–686.
23. G. Rodriguez-Yam, R. Davis, L. Scharf, A Bayesian model and Gibbs sampler for hyperspectral imaging, in: Proc. IEEE Sensor Array and Multichannel Signal Processing Workshop, 2002, pp. 105–109.
24. M.-D. Iordache, J. Bioucas-Dias, A. Plaza, Sparse unmixing of hyperspectral data, Geoscience and Remote Sensing, IEEE Transactions on 49 (6) (2011) 2014 –2039.