Spectral unmixing algorithms

K. Themelis, A. Rontogiannis, and K. Koutroumbas

July 21, 2013

Abstract

This document presents the sparse Bayesian unmixing algorithms recently developed in the framework of the "HSI-MARS" research project. 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 hierarchical Bayesian model is presented and suitable priors are selected for the model parameters such that, on the one hand, they ensure the non-negativity of the abundances, while on the other hand they favor sparse solutions for the abundances' vector. To perform Bayesian inference based on the proposed hierarchical Bayesian model, we resort to the variational Bayes methodology. Hence, a computationally efficient variational Bayes algorithm is then presented, where approximating posteriors for all model parameters are derived. 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.

Contents

1 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.

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 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, [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]. Semi-supervised unmixing, [9, 10], which is considered in this paper, 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.

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., [11–14], in semi-supervised unmixing. A number of such semi-supervised unmixing techniques has been recently proposed in [10, 15, 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 technical report, a hierarchical Bayesian approach for semi - supervised hyperspectral unmixing is adopted, which is based on the sparsity hypothesis and the non-negativity property of the abundances. The adopted Bayesian model has been recently presented in [17]. In this 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 non-negativity of the abundances, a truncated non-negative Gaussian distribution is used as a first level prior. The variance parameters 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, [18, 19]. In addition, compared to other related hierarchical models, [14,20,21], 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 non-negativity constrained variant of the adaptive least absolute shrinkage and selection operator (Lasso) criterion of [22], 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, we rely on a VB algorithm to perform statistical inference for the model parameters. Closed form expressions are provided for the updating of the parameters of all posterior approximating distributions. More importantly, based on suitable algebraic manipulations, a fast scheme is derived that allows us to reduce the computational complexity of the VB algorithm by one order of magnitude, [23]. This scheme performs Bayesian inference for all model parameters, and hence, there is no need for parameter cross-validation (as opposed to deterministic methods, e.g. SUnSAL, [24]). To demonstrate the efficiency of the proposed scheme, experimental results on both simulated and real hyperspectral data are provided.

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 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.

2 Problem formulation

In this section, we provide definitions and formulate rigorously the sparse semi-supervised 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 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 $\mathbf{\Phi}$ 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 non-negative, 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 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.

This problem can be solved using either one of the recently proposed compressive sensing techniques, e.g., [11, 13, 14, 20], that focus only on the sparsity issue, or quadratic programming techniques, e.g., [8], that successfully enforce the constraints given in eq. (2), but do not exploit sparsity. In the following, a hierarchical Bayesian model is adopted, that favors sparsity and takes into account the non-negativity constraint of the problem. Then, the variational Bayes framework is used to perform Bayesian inference for the model parameters.

3 Hierarchical Bayesian model

This section describes the recently proposed hierarchical Bayesian model, [17], used to estimate the sparse abundance vector $\mathbf{w}$ from (1), subject to the non-negativity 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,

3.1 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

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

3.2 Parameter prior distributions

The Bayesian formulation requires that both the *sparsity* and *non-negativity* properties of $\mathbf{w}$ should emanate from a suitably selected prior distribution. A widely used prior that favors sparsity, [14, 18, 20, 21, 25], is the zero-mean Laplace probability density function, which, for a single w_i, is defined as

$$\mathcal{L}(w_i|\lambda) = \frac{\lambda}{2}\exp\left[-\lambda|w_i|\right], \tag{4}$$

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

$$\mathcal{L}(\mathbf{w}|\lambda) = \prod_{i=1}^{N}\mathcal{L}(w_i|\lambda) = \left(\frac{\lambda}{2}\right)^N \exp\left[-\lambda\left\|\mathbf{w}\right\|_1\right]. \tag{5}$$

It can be easily shown, [18], that under the Laplace prior, the maximum a posteriori (MAP) estimate of $\mathbf{w}$ is given by

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}}\left\{\frac{\beta}{2}\left\|\mathbf{y}-\mathbf{\Phi}\mathbf{w}\right\|_2^2 + \lambda\left\|\mathbf{w}\right\|_1\right\}, \tag{6}$$

which is the solution of the Lasso criterion of [26]. 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 (3), and hence, the posterior probability density function of $\mathbf{w}$ could not be derived in closed form. As noted in [27], 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.,

$$\frac{\lambda}{2}\exp\left[-\lambda\left|w_i\right|\right] = \int_0^{+\infty}\frac{1}{\sqrt{2\pi s}}\exp\left[-\frac{w_i^2}{2s}\right]\frac{\lambda^2}{2}\exp\left[-\frac{\lambda^2 s}{2}\right]ds, \quad \lambda > 0, \tag{7}$$

In the framework of the problem at hand, eq. (7) 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), [28], has been adopted in [14, 18, 20, 21, 25, 29]. The main advantage of this formulation is that it maintains the conjugacy of the involved parameters.

[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.

In this work, a slightly different Bayesian model is adopted. More specifically, in order to satisfy the non-negativity constraint of the abundance vector $\mathbf{w}$, the proposed hierarchical Bayesian approach uses a *truncated* normal distribution[2] in the non-negative 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 Section 7.1)

$$p(\mathbf{w}|\boldsymbol{\gamma}, \beta) = \mathcal{N}_{\mathbf{R}_+^N}(\mathbf{w}|0, \beta^{-1}\boldsymbol{\Lambda}^{-1}). \tag{8}$$

$\mathbf{R}_+^N$ is the non-negative 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 $\boldsymbol{\Lambda}^{-1} = \text{diag}(\boldsymbol{\gamma})$, where $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \ldots, \gamma_N]^T$. Note that the use of β as a normalization parameter in (8), ensures the unimodality of the posterior distribution of $\mathbf{w}$, [21, 29].

For the second parameter, β, appearing in the likelihood function (3), a Gamma prior distribution is assumed, defined as

$$p(\beta|\kappa, \theta) = \Gamma(\beta|\kappa, \theta) = \frac{\theta^\kappa}{\Gamma(\kappa)}\beta^{\kappa-1}\text{exp}\left[-\theta\beta\right], \tag{9}$$

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 $\text{E}[p(\beta|\kappa, \theta)] = \frac{\kappa}{\theta}$, and $var[p(\beta|\kappa, \theta)] = \frac{\kappa}{\theta^2}$, respectively.

3.3 Hyperparameters' priors

Having defined the truncated Gaussian distribution for w_i's, we now focus on the definition of the exponential distributions for γ_i's, in the spirit of eq. (7). 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 [18, 20]. There, the following exponential priors on γ_i are used

$$p(\gamma_i|\lambda) = \Gamma(\gamma_i|1, \frac{\lambda}{2}) = \frac{\lambda}{2}\text{exp}\left[-\frac{\lambda}{2}\gamma_i\right], \; i = 1, 2, \ldots, N, \tag{10}$$

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 (8), the prior distribution of

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

$\mathbf{w}$ would be given as follows

$$\begin{aligned} p(\mathbf{w}|\lambda,\beta) &= \int p(\mathbf{w}|\boldsymbol{\gamma},\beta)p(\boldsymbol{\gamma}|\lambda)\, d\boldsymbol{\gamma} = \prod_{i=1}^{N}\int_0^{\infty} p(w_i|\gamma_i,\beta)p(\gamma_i|\lambda)d\gamma_i \\ &= (\beta\lambda)^{\frac{N}{2}}\exp\left[-\sqrt{\beta\lambda}\sum_{i=1}^{N}|w_i|\right]\mathcal{I}_{\mathbf{R}_+^N}(\mathbf{w}) = \mathcal{L}\left(\mathbf{w}|\sqrt{\beta\lambda}\right)\mathcal{I}_{\mathbf{R}_+^N}(\mathbf{w}), \end{aligned} \tag{11}$$

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 (5), and the Lasso criterion (6). In a similar way, it can be easily shown that under the truncated Laplace prior given in (11), the MAP estimator of $\mathbf{w}$ would be the solution of a non-negativity constrained Lasso criterion. Moreover, from a Lasso point of view, [26], 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 (10), 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

$$p(\gamma_i|\lambda_i) = \Gamma(\gamma_i|1,\frac{\lambda_i}{2}) = \frac{\lambda_i}{2}\exp\left[-\frac{\lambda_i}{2}\gamma_i\right],\ i=1,2,\ldots,N, \tag{12}$$

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

$$p(\boldsymbol{\gamma}|\boldsymbol{\lambda}) = \prod_{i=1}^{N}\left[\frac{\lambda_i}{2}\exp\left[-\frac{\lambda_i}{2}\gamma_i\right]\right] = \left(\frac{1}{2}\right)^N|\boldsymbol{\Psi}|\exp\left[-\frac{1}{2}\sum_{i=1}^{N}\lambda_i\gamma_i\right], \tag{13}$$

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 (8) and (13), constitute a sparsity-promoting non-negative (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

$$\begin{aligned} p(w_i|\lambda_i,\beta) &= \int_0^{\infty} p(w_i|\gamma_i,\beta)p(\gamma_i|\lambda_i)d\gamma_i \\ &= \sqrt{\beta\lambda_i}\,\exp\left[-\sqrt{\beta\lambda_i}\,|w_i|\right]\mathcal{I}_{\mathbf{R}_+^1}(w_i), \end{aligned} \tag{14}$$

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

$$p(\mathbf{w}|\boldsymbol{\lambda},\beta) = \int p(\mathbf{w}|\boldsymbol{\gamma},\beta)p(\boldsymbol{\gamma}|\boldsymbol{\lambda})d\boldsymbol{\gamma} = \prod_{i=1}^{N}\int_0^{\infty} p(w_i|\gamma_i,\beta)p(\gamma_i|\lambda_i)d\gamma_i$$
$$= \beta^{\frac{N}{2}}|\boldsymbol{\Psi}|^{\frac{1}{2}}\prod_{i=1}^{N}\left[\exp\left[-\sqrt{\beta\lambda_i}\,|w_i|\right]\mathcal{I}_{\mathbf{R}_+^1}(w_i)\right]$$
$$= \beta^{\frac{N}{2}}|\boldsymbol{\Psi}|^{\frac{1}{2}}\exp\left[-\sqrt{\beta}\sum_{i=1}^{N}\sqrt{\lambda_i}\,|w_i|\right]\mathcal{I}_{\mathbf{R}_+^N}(\mathbf{w}). \quad (15)$$

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 [22]. Indeed, as it is shown in Section 7.2, the MAP estimator of $\mathbf{w}$ that follows the truncated Laplace prior of (15) coincides with the estimation of $\mathbf{w}$ resulting via the optimization of the non-negativity constrained adaptive Lasso criterion, which is expressed as

$$\tilde{\mathbf{w}} = \arg\min_{\mathbf{w}}\left\{\frac{\beta}{2}\|\mathbf{y}-\boldsymbol{\Phi}\mathbf{w}\|_2^2 + \sum_{i=1}^{N}\alpha_i w_i\right\},\ \text{s.t. } \mathbf{w}\in\mathbf{R}_+^N, \quad (16)$$

for $\alpha_i = \sqrt{\beta\lambda_i}, i = 1\ldots N$. As shown in (16), 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, [22], which is not the case for the original Lasso estimator (6).

It is obvious from (15) 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 $\boldsymbol{\lambda}$,

$$p(\lambda_i|r,\delta) = \Gamma(\lambda_i|r,\delta) = \frac{\delta^r}{\Gamma(r)}{\lambda_i}^{r-1}\exp\left[-\delta\lambda_i\right],\quad i=1,2,\ldots,N \quad (17)$$

where r and δ are hyperparameters, with $r\geq 0$ and $\delta\geq 0$. Both Gamma priors of β, in (9), and λ_i, in (17), are flexible enough to express prior information, by properly tuning their hyperparameters. In this paper, we use a non-informative Jeffrey's prior ($p(x)\propto\frac{1}{x}$) over these parameters, which is obtained from (9) and (17) by setting all hyperparameters κ,θ,r,δ of the Gamma distributions to zero, as in [9, 19, 20]. A schematic representation of the proposed hierarchical Bayesian model in the form of a directed acyclic graph is shown in Fig. 1.

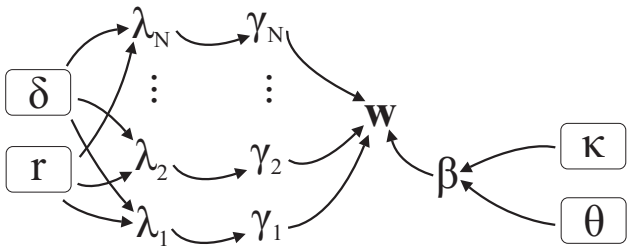


Figure 1: Directed acyclic graph of the proposed Bayesian model. The deterministic model parameters appear in boxes.

4 Bayesian Inference

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

$$p\left(\mathbf{w},\beta,\boldsymbol{\gamma},\boldsymbol{\lambda}|\mathbf{y}\right)=\frac{p\left(\mathbf{y}|\mathbf{w},\beta\right)p\left(\mathbf{w}|\beta,\boldsymbol{\gamma}\right)p\left(\boldsymbol{\gamma}|\boldsymbol{\lambda}\right)p\left(\boldsymbol{\lambda}\right)p\left(\beta\right)}{p(\mathbf{y})}, \tag{18}$$

which is intractable, in the sense that the integral

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

cannot be expressed in closed form. In such cases, the variational Bayes algorithm [30–33] provides an alternative method for overcoming this impediment. Variational Bayesian methods are primarily used (a) to provide an analytical approximation to the posterior probability of the Bayesian model parameters, in order to do statistical inference over these parameters, and (b) to derive a lower bound for the marginal likelihood (sometimes called the "evidence") of the observed data (i.e. the marginal probability of the data given the model). In the following, we derive the exact posterior approximating distributions for all model parameters defined in Section 3.

4.1 Variational Bayes

Assuming posterior independence among model parameters, the joint posterior (18) can be factorized as

$$p\left(\mathbf{w},\beta,\boldsymbol{\gamma},\boldsymbol{\lambda}|\mathbf{y}\right)\approx q\left(\mathbf{w},\beta,\boldsymbol{\gamma},\boldsymbol{\lambda}\right)=q(\mathbf{w})q(\beta)\prod_{i=1}^{N}q(\gamma_i)\prod_{i=1}^{N}q(\lambda_i), \tag{20}$$

and we can derive closed form expressions for all approximate posterior distributions $q(\mathbf{w})$, $q(\boldsymbol{\gamma})$, $q(\boldsymbol{\lambda})$, and $q(\beta)$, by utilizing the Kullback-Leibler (KL) distance minimization criterion, [33]. It is not difficult to verify by simple computations that the posterior $q(\mathbf{w})$ is a non-negatively truncated Gaussian distribution given by

$$q(\mathbf{w}) = \mathcal{N}_{\mathbf{R}_+^N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{21}$$

with

$$\boldsymbol{\mu} = \langle\beta\rangle\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{y}, \text{ and } \boldsymbol{\Sigma} = \langle\beta\rangle^{-1}\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \langle\boldsymbol{\Gamma}^{-1}\rangle\right)^{-1}, \tag{22}$$

where $\langle\cdot\rangle$ denotes expectation of a random variable with respect to its corresponding posterior $q(\cdot)$. The posterior $q(\beta)$ for the precision parameter β is expressed as

$$q(\beta) = \Gamma\left(\frac{M+N}{2} + \kappa, \frac{1}{2}\langle\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}\|^2\rangle + \theta + \frac{1}{2}\langle\mathbf{w}^T\boldsymbol{\Gamma}^{-1}\mathbf{w}\rangle\right). \tag{23}$$

Straightforward computations yield that the approximating posterior pdf of $\gamma_i, i = 1, 2, \ldots, N$ is the following generalized inverse Gaussian distribution

$$q(\gamma_i) = \left(\frac{\langle\lambda_i\rangle}{2\pi}\right)^{\frac{1}{2}} \gamma_i^{-\frac{1}{2}} \exp\left[-\frac{\langle\beta\rangle\langle w_i^2\rangle}{2\gamma_i} - \frac{\langle\lambda_i\rangle}{2}\gamma_i + \sqrt{\langle\beta\rangle\langle\lambda_i\rangle}\langle w_i\rangle\right]. \tag{24}$$

Next, the posterior $q(\lambda_i), i = 1, 2, \ldots, N$ is expressed as

$$q(\lambda_i) = \Gamma\left(\alpha_i|1 + \rho, \frac{\langle\gamma_i\rangle}{2} + \delta\right). \tag{25}$$

It is easy to verify from the resulting posterior distributions that the model parameters are interrelated. This gives rise to an iterative updating procedure, where the distributions' moments are easily evaluated using the follow-

ing results

$$\langle w_i \rangle = \mu_{i,tr}, \tag{26}$$

$$\langle w_i^2 \rangle = \mu_{i,tr}^2 + \sigma_{ii,tr} \tag{27}$$

$$\langle \gamma_i \rangle = \sqrt{\frac{\langle \beta \rangle \langle w_i^2 \rangle}{\langle \lambda_i \rangle}} + \frac{1}{\langle \lambda_i \rangle}, \tag{28}$$

$$\langle \gamma_i^{-1} \rangle = \sqrt{\frac{\langle \lambda_i \rangle}{\langle \beta \rangle \langle w_i^2 \rangle}} \tag{29}$$

$$\langle \lambda_i \rangle = \frac{1+\rho}{\frac{1}{2}\langle \gamma_i \rangle + \delta} \tag{30}$$

$$\langle \|\mathbf{y} - \mathbf{\Phi w}\|^2 \rangle = \|\mathbf{y} - \mathbf{\Phi} \boldsymbol{\mu}_{tr}\|^2 + \text{Trace}\left[\mathbf{\Phi} \mathbf{\Sigma}_{\text{tr}} \mathbf{\Phi}^{\text{T}}\right] \tag{31}$$

$$\langle \mathbf{w}^T \mathbf{\Gamma}^{-1} \mathbf{w} \rangle = \sum_{i=1}^{N} \left[\langle \gamma_i^{-1} \rangle \langle w_i^2 \rangle\right] \tag{32}$$

$$\langle \beta \rangle = \frac{\frac{M+N}{2} + \kappa}{\frac{1}{2}\langle \|\mathbf{y} - \mathbf{\Phi w}\|^2 \rangle + \theta + \frac{1}{2}\langle \mathbf{w}^T \mathbf{\Gamma}^{-1} \mathbf{w} \rangle}, \tag{33}$$

where $\boldsymbol{\mu}_{tr} = [\mu_{1,tr}, \mu_{2,tr}, \ldots, \mu_{N,tr}]^T$ is the mean and $\mathbf{\Sigma}_{tr}$ is the covariance matrix of the truncated Gaussian distribution $q(\mathbf{w})$ in (21). Note that $\boldsymbol{\mu}_{tr}$ will be the estimate of the sparse abundance vector of the pixel $\mathbf{y}$. The proposed VB scheme iterates among the parameters of the approximating posterior distributions $q(\mathbf{w})$, $q(\gamma_i)$, $q(\lambda_i)$, $q(\beta)$, utilizing the required moments in (26)-(33). Convergence is achieved since in each step the KL distance between the true posterior (18) and the approximating distribution (20) is decreased. The most computationally demanding tasks of the proposed VB algorithm involve the computation of $\boldsymbol{\mu}_{tr}$ and $\mathbf{\Sigma}_{tr}$ of the truncated Gaussian distribution (21). To reduce complexity significantly, an efficient scheme is presented next for the computation of $\boldsymbol{\mu}_{tr}$. Moreover, the need to compute $\mathbf{\Sigma}_{tr}$ analytically is alleviated by making the reasonable approximations $\langle \|\mathbf{y} - \mathbf{\Phi w}\|^2 \rangle = \|\mathbf{y} - \mathbf{\Phi} \boldsymbol{\mu}_{tr}\|^2$ and $\langle w_i^2 \rangle = \mu_{i,tr}^2$. The details of the proposed variational Bayes algorithm are also described in [23].

4.2 Fast computation of the abundance vector estimate $\boldsymbol{\mu}_{tr}$

In [17], an iterative scheme has been proposed to compute the expectation of a multivariate Gaussian distribution truncated in the non-negative orthant of $\mathcal{R}^N$. In this paper, we propose a more computationally efficient implementation of this scheme, based on suitable algebraic manipulations. The scheme

proposed in [17] iterates among the means of the one-dimensional conditional distributions of the i-th element of $\mathbf{w}$ conditioned on the remaining elements $\boldsymbol{\mu}_{\neg i,tr} = [\mu_{1,tr}, \ldots, \mu_{i-1,tr}, \mu_{i+1,tr}, \ldots, \mu_{N,tr}]^T$. These conditional distributions are expressed as, [17],

$$w_i|\boldsymbol{\mu}_{\neg i,tr} \sim \mathcal{N}_{\mathbf{R}_+^1}\left(w_i|\mu_i^*, \sigma_{ii}^{*2}\right) \tag{34}$$

with

$$\mu_i^* = \mu_i + \boldsymbol{\sigma}_{\neg i}^T \boldsymbol{\Sigma}_{\neg i \neg i}^{-1}\left(\boldsymbol{\mu}_{\neg i,tr} - \boldsymbol{\mu}_{\neg i}\right) \tag{35}$$

$$\sigma_{ii}^{*2} = \sigma_{ii} - \boldsymbol{\sigma}_{\neg i}^T \boldsymbol{\Sigma}_{\neg i \neg i}^{-1} \boldsymbol{\sigma}_{\neg i}, \tag{36}$$

where μ_i and σ_{ii} represent the i-th and ii-th 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 i-th row and the i-th 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, and $\boldsymbol{\mu}_{\neg i}$ is the vector resulting from $\boldsymbol{\mu}$ after removing its i-th element μ_i. The j-th iteration of the proposed scheme can be expressed as

$$\begin{aligned}
&1.\ \mu_{1,tr}^{(j)} = \mathrm{E}[p(w_1|\mu_{2,tr}^{(j-1)}, \mu_{3,tr}^{(j-1)}, \ldots, \mu_{N,tr}^{(j-1)})] \\
&2.\ \mu_{2,tr}^{(j)} = \mathrm{E}[p(w_2|\mu_{1,tr}^{(j)}, \mu_{3,tr}^{(j-1)}, \ldots, \mu_{N,tr}^{(j-1)})] \\
&\qquad\vdots \\
&N.\ \mu_{N,tr}^{(j)} = \mathrm{E}[p(w_N|\mu_{1,tr}^{(j)}, \mu_{2,tr}^{(j)}, \ldots, \mu_{N-1,tr}^{(j)})].
\end{aligned} \tag{37}$$

Note that in the one-dimensional case, the expectation of a random variable $x \sim \mathcal{N}_{\mathbf{R}_+^1}(x|\mu^*, \sigma^{*2})$, such as those in (37), can be computed as, [17],

$$\mathrm{E}[x] = \mu^* + \frac{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{\mu^{*2}}{\sigma^{*2}}\right)}{1 - \frac{1}{2}\mathrm{erfc}\left(\frac{\mu^*}{\sqrt{2}\sigma^*}\right)}\sigma^*, \tag{38}$$

with $\mathrm{erfc}(\cdot)$ being the complementary error function. It has been experimentally verified that this scheme converges after a few iterations, [17].

In the sequel we show that it is possible to drop the dependence on $\boldsymbol{\mu}$ in (35) and sidestep the complex operations of matrix inversions, i.e., the computation of $\boldsymbol{\Sigma}_{\neg i \neg i}^{-1} \forall i$, which have complexity $\mathcal{O}(N(N-1)^3)$. To this end, straightforward computations for μ_i^* in (35) yield that

$$\begin{aligned}
\mu_i^* &= \mu_i + \boldsymbol{\sigma}_{\neg i}^T \boldsymbol{\Sigma}_{\neg i \neg i}^{-1}(\tilde{\mathbf{w}}_{\neg i} - \boldsymbol{\mu}_{\neg i}) \\
&= \boldsymbol{\sigma}_{\neg i}^T \boldsymbol{\Sigma}_{\neg i \neg i}^{-1} \boldsymbol{\mu}_{\neg i,tr} + \begin{bmatrix} -\boldsymbol{\sigma}_{\neg i}^T \boldsymbol{\Sigma}_{\neg i \neg i}^{-1} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}_{\neg i} \\ \mu_i \end{bmatrix}.
\end{aligned} \tag{39}$$

Setting $\mathbf{z} = \mathbf{\Phi}^T\mathbf{y}$, (22) becomes $\boldsymbol{\mu} = \langle\beta\rangle\mathbf{\Sigma}\mathbf{z}$ and we get

$$\left[\begin{array}{c} \boldsymbol{\mu}_{\neg i} \\ \mu_i \end{array}\right] = \mathbf{T}_i\boldsymbol{\mu} = \langle\beta\rangle\mathbf{T}_i\mathbf{\Sigma}\mathbf{z} = \langle\beta\rangle\mathbf{T}_i\mathbf{\Sigma}\mathbf{T}_i^T\mathbf{T}_i\mathbf{z} = \langle\beta\rangle\mathbf{\Sigma}_i(\mathbf{T}_i\mathbf{z}) \tag{40}$$

where $\mathbf{T}_i$ is an appropriate permutation matrix and $\mathbf{\Sigma}_i$ is obtained from $\mathbf{\Sigma}$ by moving its i-th column and row to the end of the matrix,

$$\mathbf{\Sigma}_i = \left[\begin{array}{cc} \mathbf{\Sigma}_{\neg i\neg i} & \boldsymbol{\sigma}_{\neg i} \\ \boldsymbol{\sigma}_{\neg i}^T & \sigma_{ii} \end{array}\right]. \tag{41}$$

By substituting (41) in (40), and then in (39), it easily follows that

$$\mu_i^* = \boldsymbol{\sigma}_{\neg i}^T\mathbf{\Sigma}_{\neg i\neg i}^{-1}\boldsymbol{\mu}_{\neg i,tr} + \langle\beta\rangle\sigma_{ii}^{*2}z_i \tag{42}$$

Let us denote with $\mathbf{v}_{\neg i}^T$ the i-th row of $\mathbf{V} = \langle\beta\rangle^{-1}\mathbf{\Sigma}^{-1}$ excluding its i-th element v_{ii}. From (41) and the partitioned covariance matrix inversion formula, [34], we get

$$\mathbf{v}_{\neg i}^T = -\frac{\langle\beta\rangle^{-1}}{\sigma_{ii}^*}\boldsymbol{\sigma}_{\neg i}^T\mathbf{\Sigma}_{\neg i\neg i}^{-1}, \tag{43}$$

and

$$\sigma_{ii}^{*2} = \frac{\langle\beta\rangle^{-1}}{v_{ii}}. \tag{44}$$

Using (43), (42) becomes

$$\mu_i^* = \frac{1}{v_{ii}}(z_i - \mathbf{v}_{\neg i}^T\boldsymbol{\mu}_{\neg i,tr}), \tag{45}$$

that is, each μ_i^* is efficiently computed with N operations.

The proposed algorithm is summarized in Table 1. Note that matrix inversions have been completely eliminated and the required computational complexity of the algorithm is $\mathcal{O}(N^2)$ per iteration t, which is one order of magnitude less than the original BI-ICE algorithm, [17]. In addition, both algorithms converge very fast, exhibit similar estimation performance, and produce sparse estimates without the need of tuning or cross-validating any parameters.

Input $\mathbf{y}, \mathbf{\Phi}$
Initialize $\beta, \boldsymbol{\gamma}, \boldsymbol{\lambda}$
Compute $\mathbf{A} = \mathbf{\Phi}^T\mathbf{\Phi}$, and $\mathbf{z} = \mathbf{\Phi}^T\mathbf{y}$
for $t = 1, 2, \ldots$
- compute $\mathbf{V}(t) = \mathbf{A} + \mathbf{\Gamma}^{-1}(t)$
for $i = 1, 2, \ldots, N$
- extract $\mathbf{v}_{\neg i}(t)$ and $v_{ii}(t)$ from $\mathbf{V}(t)$
- compute $\sigma_{ii}^{*2}(t)$ from (43) and $\mu_i^*(t)$ from (45)
- compute $\mu_{i,tr}(t)$ from (38)
end for
- compute $\beta(t)$ from (33)
- compute $\boldsymbol{\gamma}(t)$ from (28)
- compute $\boldsymbol{\lambda}(t)$ from (30)
end for

Table 1: The proposed fast variational Bayes algorithm.

4.3 Embedding the sum-to-one constraint

The sparsity-promoting hierarchical Bayesian model presented in the previous sections takes into consideration the non-negativity of the abundance vector $\mathbf{w}$. However, the abundances' sum-to-one constraint has not yet been considered. As noted in [24], 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 $\mathbf{\Phi}$ having non-negative entries already admits a generalized sum-to-one constraint, [35]. 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], [36, p. 586], to impose the sum-to-one constraint deterministically,

by augmenting the initial LMM of (1) with an extra equation as follows,

$$\begin{bmatrix} \mathbf{y} \\ \alpha \end{bmatrix} = \begin{bmatrix} \mathbf{\Phi} \\ \alpha \mathbf{1}^T \end{bmatrix} \mathbf{w} + \begin{bmatrix} \mathbf{n} \\ 0 \end{bmatrix} \tag{46}$$

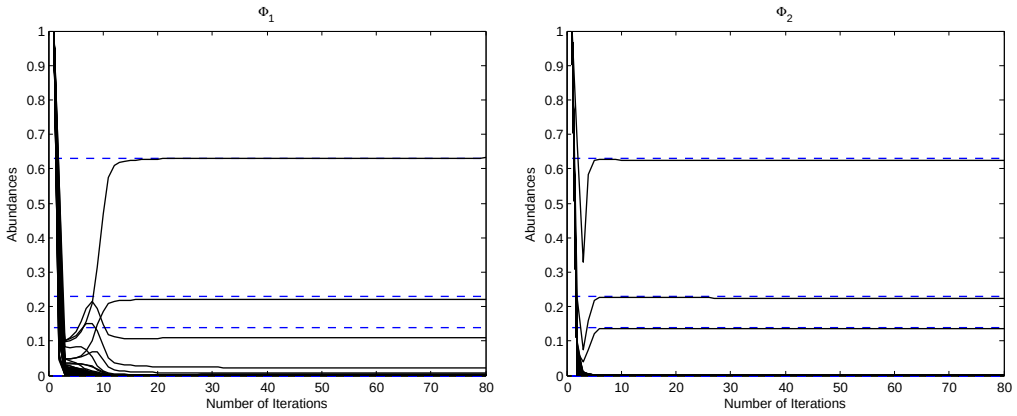
where α is a scalar parameter, which controls the effect of the sum-to-one constraint on the estimation of $\mathbf{w}$. Specifically, the larger the value of α is, the closer the sum of the estimated w_i's will be to one. It should be noticed that the augmentation of the LMM as in (46) does not affect the proposed hierarchical Bayesian model and the subsequent analysis.

5 Experimental Results

5.1 Simulation Results on Synthetic Data

This section illustrates the effectiveness of the proposed variational Bayes algorithm, by a series of experiments related to the unmixing of a synthetic hyperspectral image. Following the experimental settings of [24], where a thorough comparison of several sparse semi-supervised unmixing algorithms is presented, we consider two spectral data sets for the simulated hyperspectral scene: (a) $\mathbf{\Phi}_1 \in \mathcal{R}^{453\times 220}$, which is a matrix containing the spectral signatures of 220 endmembers selected from the USGS spectral library, [37], and (b) $\mathbf{\Phi}_2 \in \mathcal{R}^{453\times 220}$, 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 $\mathbf{\Phi}_1$ are highly correlated. The condition number and the mutual coherence, [24], of $\mathbf{\Phi}_1$ are 36.182×10^6 and 0.999933 respectively, whereas, for 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 non-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. 2 are the average of 50 noise realizations. We observe that less than 15 iterations are sufficient for the variational Bayes algorithm to converge to



(a) a (b) b

Figure 2: Estimation of the entries of the sparse vector $\mathbf{w}$, as the variational Bayes algorithm 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.

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.

contrast,

5.2 Simulation Results on Real Data

This section describes the application of the variational Bayes algorithm to real hyperspectral image data sets. The first data set was collected by the airborne visible/infrared imaging spectrometer (AVIRIS) flight over the Cuprite mining site, Nevada, in 1997, [38]. 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 [6,39–41]. 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. 3.

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

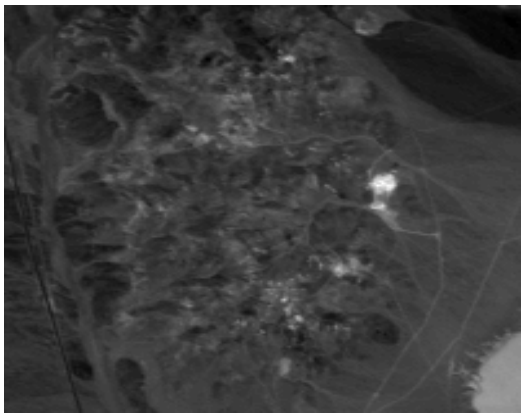
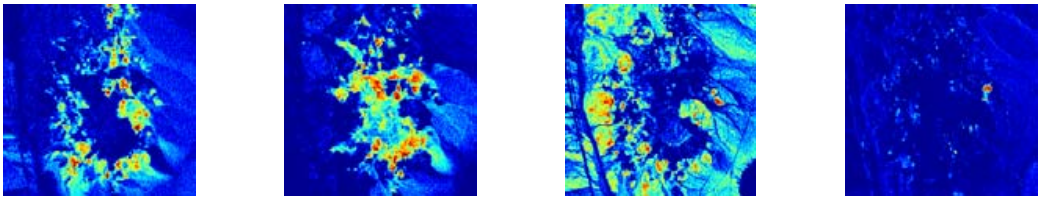


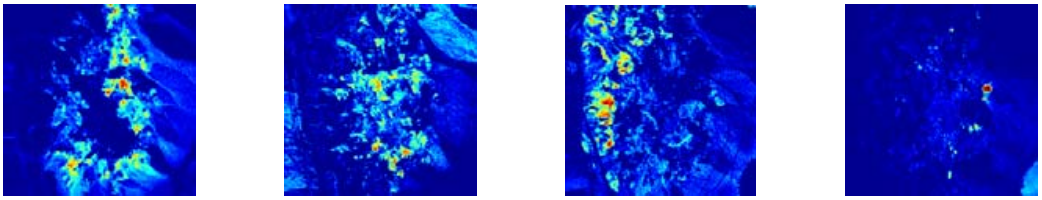
Figure 3: Band 150 of a subimage of the Cuprite Aviris hyperspectral data set.

the variational Bayes 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 variational Bayes algorithms, are shown in Fig. 4a, Fig. 4b, and Fig. 4c, 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 the variational Bayes algorithm share a high degree of similarity and are in full agreement with previous results concerning the selected abundances and reported in [6, 41], as well as with the conclusions derived in Section 5.1.

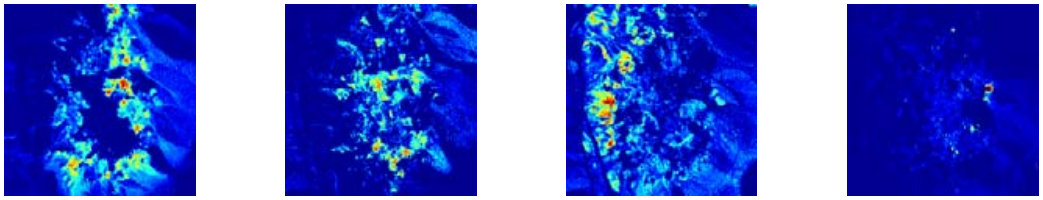
Next, we test the proposed VB algorithm on the calibrated OMEGA cube of the Syrtis Major area used in [42]. The endmember matrix $\mathbf{\Phi}$ contains the spectral signatures of 32 mineral, previously detailed in [42]. Syrtis Major is a Hesperian volcanic complex composed essentially of basalts. Mafic minerals, olivine and both low-calcium (LCP) and rich-calcium (HCP) pyroxenes [43], as well as phyllosilicates [44] have been identified in the area. Moreover, there is a significant presence of hydrated minerals [45], feldspar [44] and iron-bearing minerals such as iron oxides [46]. In the study area, Mustard et. al. in [43] have already identified the presence of three specific mafic



(a) LS algorithm



(b) QP algorithm



(c) Variational Bayes algorithm

Figure 4: Estimated abundance values of four endmembers using: (a) the LS algorithm, (b) the QP algorithm, (c) the variational Bayes algorithm

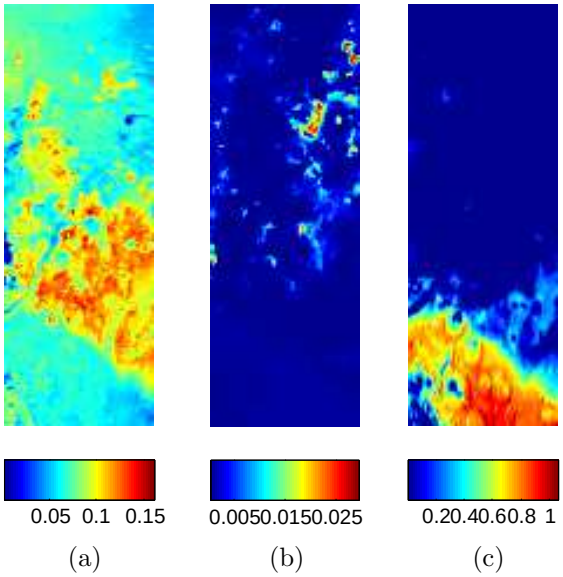


Figure 5: Abundance maps of (a) hypersthene (b) diopside and (c) magnetite in Syrtis Major obtained using the proposed VB algorithm.

minerals (i) hypersthene, (ii) diopside and (iii) fayalite.

Abundance maps obtained by applying the proposed algorithm reveal the presence of three areas with distinct characteristics in the image. In the middle part of the image, LCP pyroxenes prevail, as shown in Fig. 5a. The results shown in Fig. 5a are in accordance to the corresponding maps in [42]. The upper part of the image presents low reflectance and low abundance values and no spatially predominant mineral. HCP diopside is observed only in this area but in localized outcrops, as shown in Fig. 5b. In addition, the lower zone of the image is characterized by strong presence of iron oxides, such as magnetite (Fig. 5c) and hematite, accompanied by clay minerals such as nontronite. Finally, olivines are detected in few pixels with low abundance values in the middle left part of the image, while phyllosilicates such as muscovite are detected in the whole image, although having low abundances. The latter results are not shown here due to space limitations.

Furthermore, the mean number of abundances of value higher than 0.1 is 1.74, i.e., in the mean, approximately two endemembers are present in each pixel, which justifies the use of a sparsity-promoting unmixing scheme. The sum of abundances per pixel in the upper half of the image varies around 0.35 while in the bottom half the same sum exceeds 1.5. This is possible since the sum-to-one constraint, i.e., $\sum_{i=1}^{N} w_i = 1$, is not imposed in the

proposed model and is an indication that the endmembers library used in the unmixing process may be insufficient to effectively describe the exact mineral composition of the scene, as also noted in [42].

6 Conclusion

A variational Bayesian method for sparse semi-supervised hyperspectral unmixing has been presented in this report. 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 variational Bayes 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. Extensions of the proposed algorithm that also take into account the spatial information of the hyperspectral scene are currently under investigation.

7 Appendices

7.1 Derivation of the truncated Gaussian prior distribution of w

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

$$\begin{aligned} p(\mathbf{w}|\boldsymbol{\gamma},\beta) &= \prod_{i=1}^{N}\left[\mathcal{N}_{\mathbf{R}_+^1}(w_i|0,\frac{\gamma_i}{\beta})\right] = \prod_{i=1}^{N}\left[2\,(2\pi)^{-\frac{1}{2}}\,\beta^{\frac{1}{2}}\gamma_i^{-\frac{1}{2}}\exp\left[-\frac{\beta}{2}\frac{w_i^2}{\gamma_i}\right]\mathcal{I}_{\mathbf{R}_+^1}(w_i)\right] \\ &= 2^N\,(2\pi)^{-\frac{N}{2}}\,\beta^{\frac{N}{2}}\,|\boldsymbol{\Lambda}|^{\frac{1}{2}}\exp\left[-\frac{\beta}{2}\mathbf{w}^T\boldsymbol{\Lambda}\mathbf{w}\right]\mathcal{I}_{\mathbf{R}_+^N}(\mathbf{w}) = \mathcal{N}_{\mathbf{R}_+^N}(\mathbf{w}|0,\beta^{-1}\boldsymbol{\Lambda}^{-1}), \end{aligned} \tag{47}$$

where $\mathbf{R}_+^1$ is the set of non-negative real numbers and $\mathbf{R}_+^N$ is the non-negative 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 $\boldsymbol{\Lambda}^{-1} = \mathrm{diag}(\boldsymbol{\gamma})$.

7.2 The Non-negativity Constrained Bayesian Adaptive Lasso

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

$$\mathbf{w}_{\text{MAP}} = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{y}). \tag{48}$$

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

$$\mathbf{w}_{\text{MAP}} = \arg\max_{\mathbf{w}} p(\mathbf{y}|\mathbf{w},\beta)p(\mathbf{w}|\boldsymbol{\lambda},\beta) = \arg\min_{\mathbf{w}} \left\{-\log\left[p(\mathbf{y}|\mathbf{w},\beta)p(\mathbf{w}|\boldsymbol{\lambda},\beta)\right]\right\}. \tag{49}$$

Then, substituting in (49) the likelihood function from (3) and the truncated Laplace prior from (15), the MAP estimator can be expressed as

$$\begin{aligned}\mathbf{w}_{\text{MAP}} &= \arg\min_{\mathbf{w}} \Bigg\{-\log\Bigg[(2\pi)^{-\frac{M}{2}}\,\beta^{\frac{M}{2}} \qquad (50)\\ &\quad \exp\left[-\frac{\beta}{2}\|\mathbf{y}-\boldsymbol{\Phi}\mathbf{w}\|_2^2\right]\beta^{\frac{N}{2}}\,|\boldsymbol{\Psi}|^{\frac{1}{2}}\exp\left[-\sqrt{\beta}\sum_{i=1}^{N}\sqrt{\lambda_i}\,|w_i|\right]\mathcal{I}_{\mathbf{R}_+^N}(\mathbf{w})\Bigg]\Bigg\}\\ &= \arg\min_{\mathbf{w}}\left[\frac{\beta}{2}\|\mathbf{y}-\boldsymbol{\Phi}\mathbf{w}\|_2^2 + \sum_{i=1}^{N}\sqrt{\beta\lambda_i}\,|w_i| - \log\left(\mathcal{I}_{\mathbf{R}_+^N}(\mathbf{w})\right)\right]. \qquad (51)\end{aligned}$$

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 (15), is equivalent to solving the adaptive Lasso criterion of (16), for $\alpha_i = \sqrt{\beta\lambda_i}, i = 1,\ldots,N$, subject to $\mathbf{w}$ being non-negative, i.e., $\mathbf{w} \in \mathbf{R}_+^N$.

7.3 The approximating posterior distribution $q(\gamma_i|\mathbf{y}, w_i, \lambda_i, \beta)$ and its mean

Using (8) and (12) the posterior conditional distribution $p(\gamma_i|\mathbf{y}, w_i, \lambda_i, \beta)$ for $w_i \geq 0$ can be computed as

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

$$= \frac{\gamma_i^{-\frac{1}{2}} \exp\left[-\frac{\beta}{2}\frac{w_i^2}{\gamma_i} - \frac{\lambda_i}{2}\gamma_i\right]}{\int_0^\infty \gamma_i^{-\frac{1}{2}} \exp\left[-\frac{\beta}{2}\frac{w_i^2}{\gamma_i} - \frac{\lambda_i}{2}\gamma_i\right] d\gamma_i} = \frac{\gamma_i^{-\frac{1}{2}} \exp\left[-\frac{\beta}{2}\frac{w_i^2}{\gamma_i} - \frac{\lambda_i}{2}\gamma_i\right]}{\sqrt{\frac{2\pi}{\lambda_i}} \exp\left[-\sqrt{\beta\lambda_i w_i^2}\right]}$$

$$= \left(\frac{\lambda_i}{2\pi}\right)^{\frac{1}{2}} \gamma_i^{-\frac{1}{2}} \exp\left[-\frac{\beta w_i^2}{2\gamma_i} - \frac{\lambda_i}{2}\gamma_i + \sqrt{\beta\lambda_i}\,|w_i|\right], \tag{52}$$

where we used [47, equation 3.471.15] to compute the integral. The mean of (52) is computed as

$$\begin{aligned}
\mathrm{E}\left[q(\gamma_i|\mathbf{y}, w_i, \lambda_i, \beta)\right] &= \int_0^\infty \gamma_i \; p(\gamma_i|\mathbf{y}, w_i, \lambda_i, \beta) d\gamma_i \\
&= \int_0^\infty \left(\frac{\lambda_i}{2\pi}\right)^{\frac{1}{2}} \gamma_i^{\frac{1}{2}} \exp\left[-\frac{\beta w_i^2}{2\gamma_i} - \frac{\lambda_i}{2}\gamma_i + \sqrt{\beta\lambda_i}\,|w_i|\right] d\gamma_i \\
&= \left(\frac{\lambda_i}{2\pi}\right)^{\frac{1}{2}} \exp\left[\sqrt{\beta\lambda_i}\,|w_i|\right] \int_0^\infty \gamma_i^{\frac{1}{2}} \exp\left[-\frac{\beta w_i^2}{2\gamma_i} - \frac{\lambda_i}{2}\gamma_i\right] d\gamma_i \\
&= \left(\frac{2\lambda_i}{\pi}\right)^{\frac{1}{2}} \left(\frac{\beta w_i^2}{\lambda_i}\right)^{\frac{3}{4}} \exp\left[\sqrt{\beta\lambda_i}\,|w_i|\right] K_{3/2}\left(\sqrt{\beta\lambda_i}\,|w_i|\right),
\end{aligned} \tag{53}$$

where we used [47, equation 3.471.9] for the integral computation. Note that this does not affect the variational Bayes algorithm, since w_i's are guaranteed to be non-negative (the fact $w_i < 0$ is impossible by the formulation of the problem).

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 Annu. JPL Airborne Geosci. Workshop*, vol. 1, Washington, DC, 1993, pp. 11–14.

[5] 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, Jul. 1999, pp. 266 –275.

[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 Sensing*, 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 Journal on Optimization*, vol. 6, pp. 1040–1058, 1996.

[9] N. Dobigeon, J.-Y. Tourneret, and C.-I. Chang, "Semi-supervised 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, "Semi-supervised hyperspectral unmixing via the weighted Lasso," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'10)*, Dallas, Texas, Mar. 2010.

[11] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, 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," Department of Statistics, Stanford University, Tech. Rep., 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 *Geoscience and Remote Sensing Symposium,*

2009 IEEE International, IGARSS 2009, vol. 4, Cape Town, Jul. 2009, pp. 85 –88.

[16] J. Bioucas-Dias and 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, Jun. 2010.

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

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

[19] 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.

[20] 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.

[21] T. Park and C. George, "The Bayesian Lasso," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681–686, Jun. 2008.

[22] H. Zou, "The adaptive Lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, pp. 1418–1429, Dec. 2006.

[23] A. Rontogiannis, K. Themelis, O. Sykioti, and K. Koutroumbas, "A fast variational Bayes algorithm for sparse semi-supervised unmixing of OMEGA/Mars Express data," in *5th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, Florida, Jun 2013.

[24] M.-D. Iordache, J. Bioucas-Dias, and A. Plaza, "Sparse unmixing of hyperspectral data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 49, no. 6, pp. 2014 –2039, June 2011.

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

[26] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society*, vol. 58, no. 1, pp. 267–288, 1996.

[27] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society, Series B*, vol. 36, no. 1, pp. 99–102, 1974.

[28] 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.

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

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

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

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

[33] D. Tzikas, A. Likas, and N. Galatsanos, "The variational approximation for Bayesian inference," *Signal Processing Magazine, IEEE*, vol. 25, no. 6, pp. 131 –146, Nov. 2008.

[34] L. L. Scharf, *Statistical Signal Processing.* Prentice Hall, 1991.

[35] 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.

[36] G. H. Golub and C. F. Van Loan, *Matrix Computations (Johns Hopkins Studies in Mathematical Sciences)(3rd Edition).* The Johns Hopkins University Press, 1996.

[37] 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, http://speclab.cr.usgs.gov/spectral.lib06/ds231/datatable.html.

[38] AVIRIS free standard data products. [Online]. Available: http://aviris.jpl.nasa.gov/html/aviris.freedata.html

[39] 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. 5–15–44, Dec. 1993.

[40] 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.

[41] 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.

[42] F. Schmidt, S. Bourguignon, S. Le Mouelic, N. Dobigeon, C. Theys, and E. Treguier, "Accuracy and performance of linear unmixing techniques for detecting minerals on omega/mars express," in *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2011 3rd Workshop on*, June 2011, pp. 1 –4.

[43] J. Mustard, F. Poulet, A. Gendrin, J.-P. Bibring, Y. Langevin, B. Gondet, N. Mangold, B. G., and F. Altierri, "Olivine and pyroxene diversity in the crust of Mars," *Science*, vol. 307, pp. 1594–1597, 2005.

[44] F. Poulet, J.-P. Bibring, J. Mustard, A. Gendrin, N. Mangold, Y. Langevin, R. Arvidson, B. Gondet, and C. Gomez, "Phyllosilicates on Mars and implications for early martian climate," *Nature*, vol. 438, no. 7068, pp. 623–627, 2005.

[45] J. Bibring *et al.*, "Mars surface diversity as revealed by the OMEGA/Mars express observations," *Science*, vol. 307, pp. 1576–1581, 2005.

[46] J. Mustard, S. Erard, J. Bibring, J. Head, S. Hurtrez, Y. Langevin, C. Pieters, and C. Sotin, "The surface of Syrtis Major: Composition of the volcanic substrate and mixing with altered dust and soil," *Journal of Geophysical Research*, vol. 98, no. E2, pp. 3387–3400, 1993.

[47] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*. New York: Academic, 1980.