Bi-objective Nonnegative Matrix Factorization: Linear Versus Kernel-based Models

Fei Zhu, Paul Honeine

Bi-objective Nonnegative Matrix Factorization: Linear Versus Kernel-based Models

Fei Zhu, Paul Honeine, *Member, IEEE*

***Abstract*—Nonnegative matrix factorization (NMF) is a powerful class of feature extraction techniques that has been successfully applied in many fields, in particular in signal and image processing. Current NMF techniques have been limited to a single-objective optimization problem, in either its linear or nonlinear kernel-based formulation. In this paper, we propose to revisit the NMF as a multi-objective problem, in particular a bi-objective one, where the objective functions defined in both input and feature spaces are taken into account. By taking the advantage of the sum-weighted method from the literature of multi-objective optimization, the proposed bi-objective NMF determines a set of nondominated, Pareto optimal, solutions. Moreover, the corresponding Pareto front is approximated and studied. Experimental results on unmixing synthetic and real hyperspectral images confirm the efficiency of the proposed bi-objective NMF compared with the state-of-the-art methods.**

***Index Terms*—Kernel machines, nonnegative matrix factorization, Pareto optimal, hyperspectral image, unmixing problem.**

I. Introduction

NONNEGATIVE MATRIX FACTORIZATION (NMF) has become a versatile technique with plenty of applications [1]. As opposed to other dimensionality reduction approaches, *e.g.,* principal component analysis, vector quantization and linear discriminant analysis, the NMF is based on the additivity of the contributions of the bases to approximate the original data. Such decomposition model often yields a physical interpretation, as illustrated in many real world applications including hyperspectral unmixing [2], face and facial expression recognition [3], gene expression data [4], blind source separation [5], and clustering [6], to name a few.

The NMF approximates a nonnegative input matrix by the product of two low-rank nonnegative ones. As a consequence, it provides a decomposition suitable for many signal processing and data analysis problems, and in particular the hyperspectral unmixing problem. Indeed, a hyperspectral image is a cube that consists of a set of images of the scene under scrutiny, each corresponding to a ground scene from which the light of certain wavelength is reflected. Namely, a reflectance spectrum over a wavelength range is available for each pixel. It is assumed that each spectrum is a mixture of a few "pure" materials, called endmembers. The hyperspectral unmixing problem consists of extracting the endmembers (recorded in the first low-rank matrix), and estimating the abundance of each endmember at every pixel (recorded in the second one). Obviously, the above physical interpretation requires the nonnegativity on both abundances and endmember spectrums.

F. Zhu is with the Institut Charles Delaunay (CNRS), Université de Technologie de Troyes, France. (fei.zhu@utt.fr)

P. Honeine is with the LITIS lab, Université de Rouen, France. (paul.honeine@univ-rouen.fr)

The NMF is a linear model, since each input spectrum is approximated by a linear combination of a set of (bases) spectra. To estimate the decomposition, the objective function for minimization is defined in the so-called *input space* $\mathcal{X}$, where the difference between the input matrix and the product of the estimated ones is usually measured either by the Frobenius norm or by the generalized Kullback-Leibler divergence [1]. These objective functions are often augmented by including different regularization terms, the sparsity constraint [7], the temporal smoothness and spatial decorrelation regularization [8], and the minimum dispersion regularization [9].

Many studies have shown the limits of a linear decomposition, as shown in hyperspectral unmixing [10], [11], [12]. To extent the linear NMF model to the nonlinear scope, several kernel-based NMF have been proposed within the framework of kernel machines [13], [14]. Employing a nonlinear function, the kernel-based formulations map the columns of the data matrix to a so-called *feature space* $\mathcal{H}$, where the existing linear techniques are performed on the transformed data. The kernel trick enables the evaluation of the inner product between any pair of mapped data, without the need to explicit the nonlinear map function, as studied in [13], [15], [16] for kernel-based NMF. A major handicap of these methods resides in having the bases lying in the feature space, making them unavailable explicitly. This is due to the pre-image problem, an obstacle inherited from kernel machines [17]. In [18], [19], these difficulties are circumvented by defining a model in the feature space that can be optimized directly in the input space.

In either its linear conventional formulation or its nonlinear kernel-based formulation, as well as all of their variants, the NMF has been tackling a single-objective optimization problem. In essence, the underlying assumption is that it is known in prior that the linear model dominates the nonlinear one, or vice versa, for the data under study. To obtain such prior information about the given data is not practical in real-world applications. Moreover, it is possible that the combination of the linear and nonlinear models reveals the latent variables closer to the ground truth than each single model considered alone. Independently from the NMF framework, such combination of the linear model with a nonlinear fluctuation was recently studied in [11] and [20] where, in the former, the nonlinearity depends only on the spectral content, while it is defined by a post-nonlinear model in the latter. A multiple-kernel learning approach was studied in [21] and a Bayesian approach in [22]. While all these methods show the relevance of combining linear and nonlinear models, they share a major drawback: they only consist in estimating the abundances, while the endmembers need to be extracted in a pre-processing stage using any conventional linear technique (*e.g.*, N-Findr).

As opposed to such separation in the optimization problems, the NMF provides an elegant framework for estimating jointly the endmembers and the abundances. To the best of our knowledge, there have been no previous studies that combine the linear and nonlinear models within the NMF framework.

In this paper, we study the bi-objective optimization problem that performs the NMF in both input and feature spaces, by combining the linear and kernel-based models. The first objective function to optimize stems from the conventional linear NMF, while the second objective function, defined in the feature space, is derived from the kernel-based NMF model. In case of two conflicting objective functions, there exists a set of nondominated, noninferior or Pareto optimal solutions. In order to acquire the Pareto optimal solutions, we investigate the sum-weighed method from the literature of multi-objective optimization, due to its ease for being integrated to the proposed framework. Moreover, we attempt to approximate the corresponding Pareto front. The multiplicative update rules are derived for the resulting sub-optimization problem when the feature space is induced by the Gaussian kernel. The complexity and the convergence of the algorithm are discussed, as well as the stopping criterion.

The remainder of this paper is organized as follows. We first revisit the conventional and kernel-based NMF. The differences between the input and the feature space optimization are discussed in Section III, with physical interpretation. In Section IV, we present the proposed bi-objective NMF framework. Section V demonstrates the efficiency of the proposed method for unmixing both synthetic and real hyperspectral images. Conclusions and future works are reported in Section VI.

II. A primer on the linear and nonlinear NMF

The conventional NMF approximates a given nonnegative data matrix $\boldsymbol{X} \in \Re^{L\times T}$ with the product of two low-rank nonnegative matrices $\boldsymbol{E} \in \Re^{L\times N}$ and $\boldsymbol{A} \in \Re^{N\times T}$, namely

$$\boldsymbol{X} \approx \boldsymbol{E}\boldsymbol{A}, \tag{1}$$

under the constraints $\boldsymbol{E} \geq 0$ and $\boldsymbol{A} \geq 0$, where the nonnegativity is element-wise [1]. An equivalent vector-wise model is given by considering separately each column of the matrix $\boldsymbol{X}$, namely $\boldsymbol{x}_t$ for $t = 1, \ldots, T$, with

$$\boldsymbol{x}_t \approx \sum_{n=1}^{N} a_{nt}\, \boldsymbol{e}_n, \tag{2}$$

where $\boldsymbol{e}_n$ are the columns of $\boldsymbol{E}$ and a_{nt} the entries of $\boldsymbol{A}$. The space spanned by the vectors $\boldsymbol{x}_t$, as well as the vectors $\boldsymbol{e}_n$, is denoted the input space $\mathcal{X}$. Both matrices $\boldsymbol{E}$ and $\boldsymbol{A}$ are often estimated by minimizing the Frobenius (squared) error norm $\frac{1}{2}\|\boldsymbol{X} - \boldsymbol{E}\boldsymbol{A}\|_F^2$, subject to $\boldsymbol{E} \geq 0$ and $\boldsymbol{A} \geq 0$. In its vector-wise formulation, the objective function to minimize is

$$J_{\mathcal{X}}(\boldsymbol{E}, \boldsymbol{A}) = \frac{1}{2}\sum_{t=1}^{T} \|\boldsymbol{x}_t - \sum_{n=1}^{N} a_{nt}\, \boldsymbol{e}_n\|^2, \tag{3}$$

where the residual error of (2) is measured in the input space $\mathcal{X}$. The optimization is operated with a two-block coordinate descent scheme, by alternating between the elements of $\boldsymbol{E}$ or of $\boldsymbol{A}$, while keeping the elements in the other matrix fixed [1].

A generalization to the nonlinear form is proposed within the framework offered by kernel machines. In the following, we present the kernel-based NMF recently proposed in [18], [19]. Other formulations can also be investigated such as the ones studied in [13], [16], [15]; the price to pay is that these variants cannot construct the bases in the input space, due to the pre-image problem [19]. See Section III-B for more details.

Consider a nonlinear function $\Phi(\cdot)$ that maps the input space $\mathcal{X}$ to some feature space $\mathcal{H}$. The associated norm is denoted $\|\cdot\|_{\mathcal{H}}$, and the corresponding inner product $\langle \Phi(\boldsymbol{x}_t), \Phi(\boldsymbol{x}_{t'})\rangle_{\mathcal{H}}$, which can be evaluated using the so-called kernel function $\kappa(\boldsymbol{x}_t, \boldsymbol{x}_{t'})$ in kernel machines. Examples of kernels are the Gaussian and the polynomial kernels. By analogy with the model (1)-(2), we consider the matrix factorization $[\Phi(\boldsymbol{x}_1)\ \cdots\ \Phi(\boldsymbol{x}_T)] \approx [\Phi(\boldsymbol{e}_1)\ \cdots\ \Phi(\boldsymbol{e}_N)]\boldsymbol{A}$, namely

$$\Phi(\boldsymbol{x}_t) \approx \sum_{n=1}^{N} a_{nt}\, \Phi(\boldsymbol{e}_n). \tag{4}$$

Under the nonnegativity of all $\boldsymbol{e}_n$ and a_{nt}, the optimization problem consists in minimizing the sum of the residual errors in the feature space $\mathcal{H}$, namely

$$J_{\mathcal{H}}(\boldsymbol{E}, \boldsymbol{A}) = \frac{1}{2}\sum_{t=1}^{T} \Big\|\Phi(\boldsymbol{x}_t) - \sum_{n=1}^{N} a_{nt}\, \Phi(\boldsymbol{e}_n)\Big\|_{\mathcal{H}}^2. \tag{5}$$

By analogy to the linear case, a two-block coordinate descent scheme can be investigated to solve this optimization problem.

III. Nonlinear models for unmixing

In this section, we provide connections between state-of-the-art nonlinear models and the proposed model.

A. On augmenting the linear model with a nonlinearity

Several nonlinear models have been proposed within the hyperspectral unmixing scope, as reviewed in [24], [25]. Often advocated by a physical model, these nonlinear variations mainly consist in a combination of the linear model with an additive nonlinear term, thus of the form

$$\boldsymbol{x}_t = \sum_{n=1}^{N} a_{nt}\, \boldsymbol{e}_n + \psi(\boldsymbol{E}, \boldsymbol{a}_t),$$

where ψ is an $\mathcal{X}$-valued nonlinear function, as detailed next. It is worth noting that the same abundances and endmembers intervene in both the linear and nonlinear terms.

Bilinear models introduce bilinear mixtures of endmembers, such as the generalized bilinear model (GBM) [26] and the post-nonlinear mixing model [27], as well as the GBM-based semi-NMF approach [28]. Several kernel-based models have been proposed to define the nonlinearity term ψ in some feature space. In [11], the nonlinearity depends exclusively on the endmembers, namely $\psi(\boldsymbol{E})$. In [21], the above additive fluctuation is relaxed by considering a convex combination with multiple kernel learning. More recently, the abundances are incorporated in the nonlinear model, with a post-nonlinear model $\psi(\boldsymbol{E}\boldsymbol{a}_t)$ in [20] and a Bayesian approach is used in [22]. Another model is proposed in [12] in the context of supervised learning. All these methods consider that the

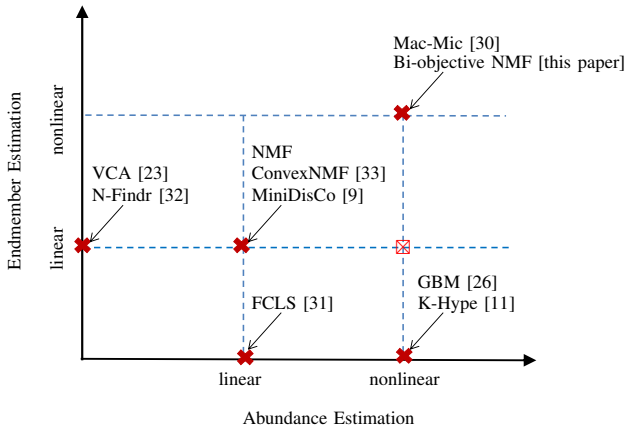

Fig. 1: Schema illustrating linear versus nonlinear models, and single versus joint estimation. Marker ⊠ shows the combinations between VCA/N-Findr and GBM/K-Hype, for instance.

endmembers $\boldsymbol{e}_n$ were already estimated using some linear technique such as N-Finder and VCA [23]; only the abundances are estimated with nonlinear models. See Fig. 1 for a schematic illustration of these differences with respect to our work that is described next (see Section III-C for connections to the Mac-Mic [29], [30]).

B. From kernelized NMF to the proposed approach

The NMF allows to estimate simultaneously the endmembers and the abundances. It has been applied either in its linear model, *i.e.,* in the input space, or in a kernel-based formulation, *i.e.,* in the feature space. In the former as studied for instance in [1], [33], [9], each sample $\boldsymbol{x}_t$ is approximated with a linear combination of basis elements $\boldsymbol{e}_n$, by minimizing the distance in the input space between each $\boldsymbol{x}_t$ and $\widehat{\boldsymbol{x}}_t = \sum_{n=1}^N a_{nt}\, \boldsymbol{e}_n$. In the latter as conducted in [13], [15], [16], [34], the basis elements $\boldsymbol{e}_n^\Phi$ belong to some kernel-induced feature space where the optimization occurs, by minimizing the distance between $\Phi(\boldsymbol{x}_t)$ and $\sum_{n=1}^N a_{nt}\, \boldsymbol{e}_n^\Phi$.

To the best of our knowledge, there has not been any attempt to examine simultaneously linear and nonlinear NMF. This is mainly due to the fact that, while one may assume that the abundances a_{nt} are the same in both representations, this is not the case of the endmembers. The linear endmembers are $\boldsymbol{e}_n \in \mathcal{X}$ while the nonlinear ones are $\boldsymbol{e}_n^\Phi \in \mathcal{H}$. It is not obvious to connect the former to the latter. Indeed, one needs to estimate $\boldsymbol{e}'_n \in \mathcal{X}$ whose $\Phi(\boldsymbol{e}'_n)$ is as close as possible to $\boldsymbol{e}_n^\Phi$. This is the curse of the pre-image problem, an ill-posed problem inherited from kernel machines [17]. Moreover, the simultaneous optimization of the linear and nonlinear NMF yields two different sets of endmembers, $\boldsymbol{e}_n$ and $\boldsymbol{e}'_n$, without any connection between them and difficult interpretation. For all these reasons, MercerNMF and KconvexNMF are not shown in Fig. 1; while the underlying models are nonlinear, the endmembers cannot be estimated.

In [18], [19], we have defined a novel nonlinear model in the feature space with $\widehat{\Psi}_t = \sum_{n=1}^N a_{nt}\, \Phi(\boldsymbol{e}_n)$; as a consequence, the endmembers $\boldsymbol{e}_n$ are estimated directly in $\mathcal{X}$. In this paper,

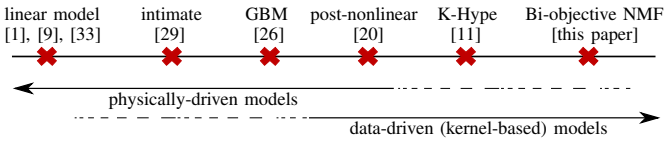

Fig. 2: Schematic illustration of the physical interpretation confronted to data-driven nonlinearity in unmixing models.

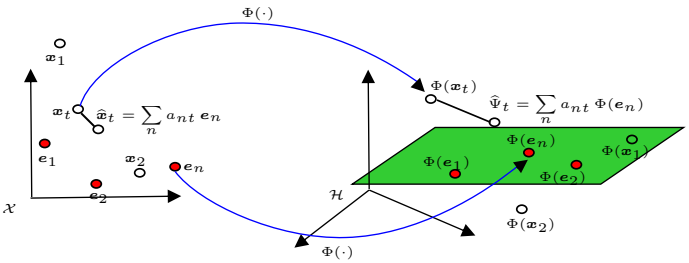

Fig. 3: In the linear NMF, each sample $\boldsymbol{x}_t$ is approximated by $\widehat{\boldsymbol{x}}_t$ in the input space $\mathcal{X}$, while in the kernel-based NMF, the mapped sample $\Phi(\boldsymbol{x}_t)$ is approximated by $\widehat{\Psi}_t$ in the feature space $\mathcal{H}$. The proposed bi-objective NMF solves simultaneously the two optimization problems.

as illustrated in Fig. 3, we combine the estimation of this model with the linear one. To this end, we minimize simultaneously $J_\mathcal{X}$ and $J_\mathcal{H}$, namely the distance in the input space between each $\boldsymbol{x}_t$ and $\widehat{\boldsymbol{x}}_t = \sum_{n=1}^N a_{nt}\, \boldsymbol{e}_n$, and the distance in $\mathcal{H}$ between $\Phi(\boldsymbol{x}_t)$ and $\widehat{\Psi}_t = \sum_{n=1}^N a_{nt}\, \Phi(\boldsymbol{e}_n)$. The resulting problem is the bi-objective NMF. We shall take advantage of the sum-weighted method to tackle this problem as a sequence of single-objective optimization problems, each corresponding to a fusion of the linear and nonlinear optimization problems, at different levels characterized by a parameter α, namely

$$\min_{\boldsymbol{E},\boldsymbol{A}} \alpha J_\mathcal{X}(\boldsymbol{E},\boldsymbol{A}) + (1-\alpha) J_\mathcal{H}(\boldsymbol{E},\boldsymbol{A}). \tag{6}$$

C. Remarks on the physical interpretation

We study the interpretation of the proposed bi-objective NMF by connecting it to several state-of-the-art models.

The nonlinear model $\Phi(\boldsymbol{x}_t) \approx \sum_{n=1}^N a_{nt} \Phi(\boldsymbol{e}_n)$ in (4) is closely related to the microscopic mixture of the Hapke model [29], [35], [30]. The latter uses the widely known bidirectional reflectance distribution function for microscopic mixtures, which describes the relationship of observed reflectance to the albedo of materials within the scene under scrutiny [29]. Indeed, the single-scattering albedo (SSA), for a wavelength λ, is defined as $w_\lambda = \sum_{n=1}^N f_n w_{n\lambda}$, where $w_{n\lambda}$ are the material albedos and f_n the corresponding fractional proportions. It is easy to see that this microscopic mixing model as a linear model in the albedo domain, while it is nonlinear in the reflectance domain. Indeed, the model in the latter takes the form $\boldsymbol{x}_t \approx R(\sum_{n=1}^N f_{nt} \boldsymbol{w}_n)$, where R is the nonlinear Hapke's reflectance function and $\boldsymbol{w}_n$ is the vector of SSA at all wavelengths. By mapping the reflectance data to the albedo-domain, the unknown microscopic proportions are estimated using the model $R^{-1}(\boldsymbol{x}_t) \approx \sum_{n=1}^N f_{nt} R^{-1}(\boldsymbol{e}_n)$,

where we have used $\boldsymbol{w}_n = R^{-1}(\boldsymbol{e}_n)$ as recommended in [30]. The nonlinear model in (4) has the same structure, where the difference lies in a nonlinearity R^{-1} characterized by a nonlinear kernel.

Machine learning with kernel-based models allow to alleviate missing physical interpretation of the underlying nonlinearity, as have been largely investigated in the literature [21], [11], [20], [12]. Fig. 2 attempts to categorize unmixing models/techniques in terms of both their "level" of physical interpretation and their data-driven modeling to describe nonlinear relations. Consider for instance the post-nonlinear model of the form $\psi(\boldsymbol{E}\boldsymbol{a}_t)$; while it has a physical interpretation as stipulated in [20], the nonlinear function $\psi(\cdot)$ is estimated from data with kernel-based methods, thus without any physical interpretation. It is worth noting that linear and quadratic models can be viewed as special cases of kernel-based models. An analysis in depth of parametric, semi-parametric, and nonparametric modeling is beyond the scope of this paper.

As opposed to augmenting the linear model with a nonlinearity (see Section III-A), the proposed model is related to the Mac-Mic presented in [30] (see Fig. 1). Indeed, the latter confronts two models for each pixel, the linear model, called macroscopic, and the aforementioned microscopic one. The proposed bi-objective NMF can also be viewed as confronting two models, a "regularized" linear model and a "regularized" nonlinear one. One way to understand this property is through two complementary viewpoints of the bi-objective optimization problem (6). In the first one, the investigated model is $\boldsymbol{x}_t \approx \sum_{n=1}^{N} a_{nt}\, \boldsymbol{e}_n$ (results from minimizing $J_{\mathcal{X}}$), while the minimization of $J_{\mathcal{H}}$ operates as a regularization. In the second viewpoint, one can likewise say that the underlying model is the nonlinear model $\Phi(\boldsymbol{x}_t) \approx \sum_{n=1}^{N} a_{nt}\, \Phi(\boldsymbol{e}_n)$, while the minimization of $J_{\mathcal{X}}$ operates as a regularization by emphasizing that the nonlinear model should not be very "distinct" from the linear one.

IV. Bi-objective optimization for NMF

A. Problem formulation

We propose to minimize simultaneously the objective functions $J_{\mathcal{X}}(\boldsymbol{E}, \boldsymbol{A})$ and $J_{\mathcal{H}}(\boldsymbol{E}, \boldsymbol{A})$, namely in both input and feature spaces as shown in Fig. 3. Such problem is in a sense an ill-defined one. Indeed, it is not possible in general to find a solution that is optimal for both objective functions. As opposed to single-objective optimization problems where the main focus would be on the decision solution space, namely the space of all entries $(\boldsymbol{E}, \boldsymbol{A})$ (of dimension $LN + NT$), the bi-objective optimization problem brings the focus on the *objective space*, namely the space of the *objective vector* $[J_{\mathcal{X}}(\boldsymbol{E}, \boldsymbol{A}) \quad J_{\mathcal{H}}(\boldsymbol{E}, \boldsymbol{A})]$. To study and solve this optimization problem, we revisit in our context the following definitions from the literature of multi-objective optimization:

- **Pareto dominance**: The solution $(\boldsymbol{E}_1, \boldsymbol{A}_1)$ is said to dominate $(\boldsymbol{E}_2, \boldsymbol{A}_2)$ if and only if $J_{\mathcal{X}}(\boldsymbol{E}_1, \boldsymbol{A}_1) \leq J_{\mathcal{X}}(\boldsymbol{E}_2, \boldsymbol{A}_2)$ and $J_{\mathcal{H}}(\boldsymbol{E}_1, \boldsymbol{A}_1) \leq J_{\mathcal{H}}(\boldsymbol{E}_2, \boldsymbol{A}_2)$, where at least one inequality is strict.
- **Pareto optimal**: A solution is a global (respectively local) Pareto optimal if and only if it is not dominated by any other solution (respectively in its neighborhood). That is, the objective vector $[J_{\mathcal{X}}(\boldsymbol{E}^*, \boldsymbol{A}^*) \quad J_{\mathcal{H}}(\boldsymbol{E}^*, \boldsymbol{A}^*)]$ corresponding to a Pareto optimal $(\boldsymbol{E}^*, \boldsymbol{A}^*)$ cannot be improved in any space (input or feature space) without any degradation in the other space.
- **Pareto front**: The set of the objective vectors corresponding to the Pareto optimal solutions forms the Pareto front in the objective space.

Various multi-objective optimization techniques have been successfully proposed *e.g.,* evolutionary algorithms, sum-weighted method, ε-constraint method, normal boundary intersection method, to name a few. See [36], [37] for a survey. Among the existing methods, the sum-weighted or scalarization method has been always the most popular one, since it is straightforward and easy to implement [38], [39]. It converts a multi-objective problem into a single-objective problem by combining the multiple objectives. Under some conditions, the resulting objective vector belongs to the convex part of multi-objective problem's Pareto front. Thus, by changing appropriately the weights among the objectives, the Pareto front of the original problem is approximated. The main drawback of this method is that the nonconvex part of the Pareto front is often unattainable [38]. Nevertheless, it is the most practical one, in view of the complexity of the NMF problem, which is nonconvex, ill-posed and NP-hard [40].

B. Bi-objective optimization with the sum-weighted method

Following the formulation introduced in the previous section, we study the minimization of the bi-objective function $[J_{\mathcal{X}}(\boldsymbol{E}, \boldsymbol{A}) \quad J_{\mathcal{H}}(\boldsymbol{E}, \boldsymbol{A})]$, under the nonnegativity of the matrices $\boldsymbol{E}$ and $\boldsymbol{A}$. The decision solution, of size $LN + NT$, corresponds to the entries in the unknown matrices $\boldsymbol{E}$ and $\boldsymbol{A}$. We transform this bi-objective optimization problem into an aggregated objective function (*i.e.*, sum-weighted objective function, also called scalarization value) which is a convex combination of the two original objective functions, namely

$$\begin{aligned} \min_{\boldsymbol{E},\boldsymbol{A}} \quad & \alpha J_{\mathcal{X}}(\boldsymbol{E}, \boldsymbol{A}) + (1-\alpha) J_{\mathcal{H}}(\boldsymbol{E}, \boldsymbol{A}) \\ \text{subject to} \quad & \boldsymbol{E} \geq 0 \text{ and } \boldsymbol{A} \geq 0 \end{aligned} \tag{7}$$

where the weight $\alpha \in [0, 1]$ controls the relative importance between objectives $J_{\mathcal{X}}$ and $J_{\mathcal{H}}$. For a fixed value of α, this problem is called the sub-optimization problem. Its solution is a Pareto optimal for the original bi-objective problem, as proven in [38] for the general case. By solving the sub-optimization problem with a spread of values of α, we obtain an approximation of the Pareto front. It is obvious that the single-objective conventional NMF in (3) is given by $\alpha = 1$, while $\alpha = 0$ leads to the kernel variant in (5).

Similar to the NMF, which is ill-posed, nonconvex and NP-hard [40], the optimization problem (7) is difficult to solve. It has no closed-form solution, a drawback inherited from most nonnegative constrained optimization problems. Moreover, the objective function is nonlinear, making the optimization problem more difficult. As in NMF algorithms, the global optimal solution cannot be guaranteed, thus the term Pareto optimal referred in the following is in the local sense.

TABLE I: Some common kernels and their gradients w.r.t. $\boldsymbol{e}_n$

Kernel	$\kappa(\boldsymbol{e}_n, \boldsymbol{z})$	$\nabla_{\boldsymbol{e}_n}\kappa(\boldsymbol{e}_n, \boldsymbol{z})$
Gaussian	$\exp(\frac{-1}{2\sigma^2}\|\boldsymbol{e}_n - \boldsymbol{z}\|^2)$	$-\frac{1}{\sigma^2}\kappa(\boldsymbol{e}_n, \boldsymbol{z})(\boldsymbol{e}_n - \boldsymbol{z})$
Polynomial	$(\boldsymbol{z}^\top \boldsymbol{e}_n + \mathrm{c})^d$	$d\,(\boldsymbol{z}^\top \boldsymbol{e}_n + \mathrm{c})^{(d-1)}\boldsymbol{z}$
Exponential	$\exp(\frac{-1}{2\sigma^2}\|\boldsymbol{e}_n - \boldsymbol{z}\|)$	$-\frac{1}{2\sigma^2}\kappa(\boldsymbol{e}_n, \boldsymbol{z})\mathrm{sgn}(\boldsymbol{e}_n - \boldsymbol{z})$
Sigmoid	$\tanh(\gamma \boldsymbol{z}^\top \boldsymbol{e}_n + \mathrm{c})$	$\gamma \mathrm{sech}^2(\gamma \boldsymbol{z}^\top \boldsymbol{e}_n + \mathrm{c})\boldsymbol{z}$

Substituting the expressions given in (3) and (5) for $J_{\mathcal{X}}$ and $J_{\mathcal{H}}$, the aggregated objective function becomes

$$\frac{\alpha}{2}\sum_{t=1}^{T}\Big\|\boldsymbol{x}_t - \sum_{n=1}^{N} a_{nt}\,\boldsymbol{e}_n\Big\|^2 + \frac{1-\alpha}{2}\sum_{t=1}^{T}\Big\|\Phi(\boldsymbol{x}_t) - \sum_{n=1}^{N} a_{nt}\,\Phi(\boldsymbol{e}_n)\Big\|_{\mathcal{H}}^2.$$

This objective function becomes, after removing the constant terms that are independent of a_{nt} and $\boldsymbol{e}_n$,

$$\begin{aligned} J = \alpha\sum_{t=1}^{T}\Big(-\sum_{n=1}^{N} a_{nt}\boldsymbol{e}_n^\top \boldsymbol{x}_t + \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_{nt}a_{mt}\boldsymbol{e}_n^\top \boldsymbol{e}_m\Big) \\ + (1-\alpha)\sum_{t=1}^{T}\Big(-\sum_{n=1}^{N} a_{nt}\kappa(\boldsymbol{e}_n, \boldsymbol{x}_t) + \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_{nt}a_{mt}\kappa(\boldsymbol{e}_n, \boldsymbol{e}_m)\Big). \end{aligned} \tag{8}$$

In the following, we derive iterative techniques to minimize it by alternating over the matrices $\boldsymbol{E}$ or $\boldsymbol{A}$, while keeping the other matrix fixed. The derivative of (8) with respect to a_{nt} is

$$\begin{aligned} \nabla_{a_{nt}} J = \alpha\Big(-\boldsymbol{e}_n^\top \boldsymbol{x}_t + \sum_{m=1}^{N} a_{mt}\,\boldsymbol{e}_n^\top \boldsymbol{e}_m\Big) \\ + (1-\alpha)\Big(-\kappa(\boldsymbol{e}_n, \boldsymbol{x}_t) + \sum_{m=1}^{N} a_{mt}\,\kappa(\boldsymbol{e}_n, \boldsymbol{e}_m)\Big). \end{aligned} \tag{9}$$

The gradient of (8) with respect to $\boldsymbol{e}_n$ is

$$\begin{aligned} \nabla_{\boldsymbol{e}_n} J = \alpha\sum_{t=1}^{T} a_{nt}\Big(-\boldsymbol{x}_t + \sum_{m=1}^{N} a_{mt}\boldsymbol{e}_m\Big) \\ + (1-\alpha)\sum_{t=1}^{T} a_{nt}\Big(-\nabla_{\boldsymbol{e}_n}\kappa(\boldsymbol{e}_n, \boldsymbol{x}_t) + \sum_{m=1}^{N} a_{mt}\nabla_{\boldsymbol{e}_n}\kappa(\boldsymbol{e}_n, \boldsymbol{e}_m)\Big). \end{aligned} \tag{10}$$

Here, $\nabla_{\boldsymbol{e}_n}\kappa(\boldsymbol{e}_n, \cdot)$ represents the gradient of the kernel with respect to its first argument $\boldsymbol{e}_n$, and can be determined for most valid kernels, as shown in TABLE I. Without loss of generality, we restrict the presentation to the Gaussian kernel for the objective function $J_{\mathcal{H}}$. In this case, expression (10) becomes (11) (given on the top of the next page).

1) Optimization over $\boldsymbol{E}$ using projected gradient (PG): We apply the projected gradient method (PG) [41], [9] to address the bound optimization problem

$$\begin{aligned} &\min_{\boldsymbol{E}} \; J(\boldsymbol{E}) \\ &\text{subject to } \boldsymbol{L}_{ln} \le \boldsymbol{E}_{ln} \le \boldsymbol{U}_{ln} \text{ for } l = 1, \ldots, L \text{ and } n = 1, \ldots, N, \end{aligned}$$

where the function $J(\boldsymbol{E}) : \Re^{L\times N} \longrightarrow \Re$ is continuously differential, and $\boldsymbol{L}$ and $\boldsymbol{U}$ are lower and upper bound matrices. At iteration k, the PG update takes the form

$$\boldsymbol{E}^{k+1} = \mathrm{P}[\boldsymbol{E}^k - \eta_k \nabla_{\boldsymbol{E}} J(\boldsymbol{E}^k)],$$

where η_k is the stepsize and $\mathrm{P}[\cdot]$ is the projection operator that maps the elements of $\boldsymbol{E}$ back to the feasible bounded region.

Algorithm 1 The k-th iteration of the PG, following [9]

Input: $0 < \rho < 1$
1: $\eta_k \leftarrow \eta_{k-1}$, $p = 1$
2: **if** η_k satisfies (12), **then**
3: **while** η_k/ρ^p satisfies (12) **do**
4: $\eta_k \leftarrow \eta_k/\rho^p$, $p \leftarrow p + 1$
5: **end while**
6: **else**
7: **while** η_k does not satisfy (12) **do**
8: $\eta_k \leftarrow \eta_k\,\rho^p$, $p \leftarrow p + 1$
9: **end while**
10: **end if**
11: update $\boldsymbol{E}^{k+1} = \mathrm{P}[\boldsymbol{E}^k - \eta_k \nabla_{\boldsymbol{E}} J(\boldsymbol{E}^k)]$.

Algorithm 2 The proposed bi-objective NMF, for a fixed $\alpha_m \in \{\alpha_1, \alpha_2, ..., \alpha_M\}$

Input: $k = 0$, warm start by $\boldsymbol{E}_m^0 = \boldsymbol{E}_{m-1}$ and $\boldsymbol{A}_m^0 = \boldsymbol{A}_{m-1}$
1: **repeat**
2: update $\boldsymbol{E}^{k+1}$ with Algorithm 1
3: update $\boldsymbol{A}^{k+1}$ with (14)
4: $k = k + 1$
5: **until** stopping criterion
Output: $\boldsymbol{E}_m$ and $\boldsymbol{A}_m$

To estimate η_k, we investigate the backtracking-Armijo line search, proved effective for NMF [41], [9]. Let $\nabla_{\boldsymbol{E}} J = [\nabla_{\boldsymbol{e}_1} J \;\; \nabla_{\boldsymbol{e}_2} J \;\; \cdots \;\; \nabla_{\boldsymbol{e}_N} J]$. At each iteration, if the condition

$$J(\boldsymbol{E}^k) - J(\boldsymbol{E}^{k+1}) \le \gamma\,\eta_k\,\mathrm{vec}(\nabla_{\boldsymbol{E}} J)^\top \mathrm{vec}(\boldsymbol{E}^{\mathrm{k}} - \boldsymbol{E}^{\mathrm{k}}) \tag{12}$$

is satisfied, a sufficient decrease of objective function is achieved. Here $\mathrm{vec}(\cdot)$ reshapes the matrix into a vector, and γ characterizes the decrease level and is often set to 1%. As given in Algorithm 1, this modified PG accelerates the stepsize search by eliminating the upper bound required in [41].

2) Optimization over $\boldsymbol{A}$ using multiplicative update (MU): The PG update rule for $\boldsymbol{A}$ can be derived in the same way as for $\boldsymbol{E}$. However, the stepsize estimation in PG rule is very time consuming. To alleviate this problem, we develop the multiplicative update (MU) for $\boldsymbol{A}$. Initially proposed in [42], the MU has been largely investigated for NMF [1]. Thanks to the convexity of the subproblem $J(\boldsymbol{A})$, the MU for $\boldsymbol{A}$ yields a monotone decrease in the objective function. Denote the matrix Λ_k the stepsize matrix at iteration k, where $(\Lambda_k)_{nt} = \lambda_{k,nt}$. The PG update rule in terms of $\boldsymbol{A}$ is

$$\boldsymbol{A}^{k+1} = \mathrm{P}[\boldsymbol{A}^k - \Lambda_k \nabla_{\boldsymbol{A}} J(\boldsymbol{A}^k)]. \tag{13}$$

Here, the stepsize balances the rate of convergence with the accuracy of optimization, and can be set differently depending on n and t. We choose the stepsize parameter in (13) as

$$\lambda_{k,nt} = \frac{a_{nt}^k}{\alpha\sum_m a_{mt}^k \boldsymbol{e}_n^\top \boldsymbol{e}_m + (1-\alpha)\sum_m a_{mt}^k\,\kappa(\boldsymbol{e}_n, \boldsymbol{e}_m)},$$

which yields

$$a_{nt}^{k+1} = a_{nt}^k\,\frac{\alpha\,\boldsymbol{e}_n^\top \boldsymbol{x}_t + (1-\alpha)\,\kappa(\boldsymbol{e}_n, \boldsymbol{x}_t)}{\alpha\sum_m a_{mt}^k \boldsymbol{e}_n^\top \boldsymbol{e}_m + (1-\alpha)\sum_m a_{mt}^k\,\kappa(\boldsymbol{e}_n, \boldsymbol{e}_m)}. \tag{14}$$

It is noteworthy that the multiplicative update rule for $\boldsymbol{e}_n$ can be elaborated in the same way, by using the so-called split gradient method. However, since the sub-optimization

$$\nabla_{\boldsymbol{e}_n} J = \alpha \sum_{t=1}^{T} a_{nt}\Big(-\boldsymbol{x}_t + \sum_{m=1}^{N} a_{mt}\boldsymbol{e}_m\Big) + \frac{1-\alpha}{\sigma^2}\sum_{t=1}^{T} a_{nt}\Big(\kappa(\boldsymbol{e}_n,\boldsymbol{x}_t)(\boldsymbol{e}_n-\boldsymbol{x}_t) - \sum_{m=1}^{N} a_{mt}\kappa(\boldsymbol{e}_n,\boldsymbol{e}_m)(\boldsymbol{e}_n-\boldsymbol{e}_m)\Big). \tag{11}$$

on $\boldsymbol{e}_n$ is possibly nonconvex[1], the monotone property is not guaranteed with an arbitrary kernel. That is, for a given weight α, although the aggregated objective function J globally decreases, the overshoot of stepsize in updating $\boldsymbol{E}$ may occur during iterations. This discussion is summarized in TABLE II.

TABLE II: The convexity and the corresponding optimization methods for the subproblem

	Convexity	PG	MU
$\min_{\boldsymbol{E}} J(\boldsymbol{E})$		✓	
$\min_{\boldsymbol{A}} J(\boldsymbol{A})$	✓	✓	✓

C. On the complexity, convergence and stopping criterion

The complexity of the PG for $\boldsymbol{E}$ is $\mathcal{O}(pTLN^2)$, where p is the average number of checking condition (12). See [41] for details on the complexity of PG. The complexity of the MU for $\boldsymbol{A}$ is $\mathcal{O}(TLN^2)$. Thus, the total complexity of Algorithm 2 is $\mathcal{O}(k(p+1)TLN^2)$ after k iterations. This complexity holds using any commonly-used kernel listed in TABLE I, with essentially the same complexity $\mathcal{O}(L)$ for each kernel.

Similar to the PG and MU rules initially presented for the linear NMF, the proposed algorithm is a stationary point method. See also the discussions on the convergence of the conventional NMF in [44], [43]. We use the two-fold stopping criterion, that is, either a stationary point is attained, or the preset maximum number of iterations is reached. To be more specific, the algorithm stops when either the condition $\|J(\boldsymbol{E}^{k+1},\boldsymbol{A}^{k+1}) - J(\boldsymbol{E}^{k},\boldsymbol{A}^{k}))\| < \varepsilon$ is satisfied, or $k = k_{\max}$, *e.g.,* $k_{\max} = 2000$. The threshold of the error difference between successive iterations is set to $\varepsilon = 10^{-4}$.

D. Posteriori analyse of the approximated Pareto front

It is worth noting that we apply the sum-weighted method as a *posteriori* method, where different Pareto optimal solutions are generated, and the decision maker (DM) makes the final compromise among optimal solutions. Alternatively, in a *priori* method, the DM specifies the weight α in advance to generate a solution. See [37] for more details.

All the points on the approximated Pareto front are optimal in some sense. To choose the α suitable to the studied data, we employ the so-called level diagrams approach proposed in [45]. This posteriori method classifies the points on the Pareto front according to their proximities to the ideal point, defined by $J^{**} = [\min J_{\mathcal{X}} \quad \min J_{\mathcal{H}}]$ in our case, where $\min J_{\mathcal{X}}$ and $\min J_{\mathcal{H}}$ denote respectively the minimum values of the two objectives obtained on the Pareto front. For this purpose, each point $J = [J_{\mathcal{X}} \quad J_{\mathcal{H}}]$ is first normalized to $\overline{J} = [\overline{J_{\mathcal{X}}} \quad \overline{J_{\mathcal{H}}}]$ using the maximum and minimum values achieved, that is

$$\overline{J_{\mathcal{X}}} = \frac{J_{\mathcal{X}} - \min J_{\mathcal{X}}}{\max J_{\mathcal{X}} - \min J_{\mathcal{X}}}, \quad \text{and} \quad \overline{J_{\mathcal{H}}} = \frac{J_{\mathcal{H}} - \min J_{\mathcal{H}}}{\max J_{\mathcal{H}} - \min J_{\mathcal{H}}}.$$

The distance to the ideal point is then evaluated with a particular ℓ_p-norm, *e.g.*, ℓ_1-norm $\|\overline{J}\|_1 = \overline{J_{\mathcal{X}}} + \overline{J_{\mathcal{H}}}$ ℓ_2-norm $\|\overline{J}\|_2 = \big(\overline{J_{\mathcal{X}}}^2 + \overline{J_{\mathcal{H}}}^2\big)^{\frac{1}{2}}$, ℓ_∞-norm $\|\overline{J}\|_\infty = \max(\overline{J_{\mathcal{X}}}, \overline{J_{\mathcal{H}}})$ and $\ell_{-\infty}$-norm $\|\overline{J}\|_{-\infty} = \min(\overline{J_{\mathcal{X}}}, \overline{J_{\mathcal{H}}})$. It is clear that the points with small norms locate nearly to the ideal point, therefore the DM can choose a solution among them.

V. Experiments

In this section, the performance of the proposed algorithm for bi-objective NMF is demonstrated on the unmixing of synthetic and real hyperspectral images. The unmixing performance is evaluated by two criteria, the averaged spectral angle distance between endmembers (SAD) and the root mean square error on the abundances (RMSE), defined as

$$\mathrm{SAD} = \frac{1}{N}\sum_{n=1}^{N} \arccos \frac{\boldsymbol{e}_n^\top \widehat{\boldsymbol{e}}_n}{\|\boldsymbol{e}_n\|\|\widehat{\boldsymbol{e}}_n\|} \qquad \mathrm{RMSE} = \sqrt{\frac{1}{NT}\sum_{t=1}^{T}\|\boldsymbol{a}_t - \widehat{\boldsymbol{a}}_t\|^2}.$$

A. State-of-the-art unmixing methods

The unmixing problem comprises the estimation of endmembers and the corresponding abundance maps. Some existing techniques either extract the endmembers (such as VCA) or estimate the abundances (such as FCLS)[2]; other methods enable the simultaneous estimations, *e.g.,* NMF and its variants. We briefly present state-of-the-art unmixing algorithms.

The most-known endmember extraction techniques include the vertex component analysis (VCA) [23], the N-Findr [32] and the NMF based ones [47]. For fair comparison, the linear NMF is applied for endmember extraction, jointly with three abundance estimation techniques. The fully constrained least squares algorithm (FCLS) [31] investigates the linear mixture model to estimate the abundances with the nonnegativity and sum-to-one constraints. K-Hype uses a linear-mixture with an additive nonlinear-fluctuation for abundance estimation, where the nonlinear term is described as a kernel-based model [11]. In [28], a generalized bilinear model is formulated with parameters optimized using the semi-nonnegative matrix factorization (GBM-sNMF). We also consider the nonlinear macroscopic/microscopic mixture model (Mac-Mic) [30].

We further consider NMF-based techniques that estimate jointly the endmembers and abundances. The minimum dispersion constrained NMF (MiniDisCo) [9] includes the dispersion regularization to the conventional NMF, by integrating the sum-to-one constraint for each pixel's abundance fractions and the minimization of variance within each endmember. The problem is solved by exploiting an alternate projected gradient

[1]In conventional NMF, the subproblem of estimating each matrix separately is convex. From this, the monotone decreasing property of the MU was proved by constructing an auxiliary function as an upper bound [1], [43]. In our work, the proposed framework involves a nonconvex optimization problem on $\boldsymbol{e}_n$, since the Hessian matrix is no longer guaranteed to be positive semidefinite.

[2]See [46] for estimating abundances with endmember extraction techniques.

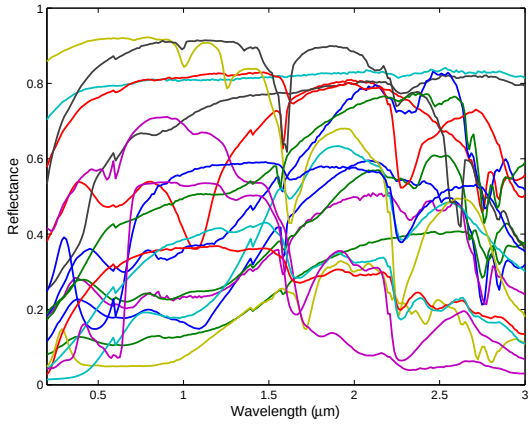

Fig. 4: The USGS spectra used for synthetic data generation.

scheme. In the convex nonnegative matrix factorization (ConvexNMF) [33], the endmember matrix is restricted to the span of the input data. The kernel convex-NMF (KconvexNMF) is essentially a kernelized variants of the ConvexNMF [16]. Nonlinear NMF based on constructing Mercer kernels (MercerNMF) [34] uses a self-constructed kernel that preserves the nonnegativity of the embedded bases and their coefficients; the embedded data being finally factorized with the classical NMF.

TABLE III: Unmixing performance on synthetic data ($\times 10^{-2}$)

		$N=3$				$N=6$			
		SNR = 30dB		SNR = 15dB		SNR = 30dB		SNR = 15dB	
		SAD	RMSE	SAD	RMSE	SAD	RMSE	SAD	RMSE
	FCLS	-	32.48	-	31.99	-	30.01	-	28.17
	GBM-sNMF	-	28.91	-	27.48	-	27.79	-	26.49
	K-Hype	-	8.40	-	10.63	-	12.31	-	11.11
	MiniDisCo	8.20	10.49	11.60	12.24	14.53	②**7.66**	17.93	②**7.99**
	ConvexNMF	14.19	21.43	13.91	21.96	19.06	12.15	20.00	12.86
	KconvexNMF	-	14.40	-	16.36	-	12.45	-	12.40
	MercerNMF	-	16.02	-	15.94	-	①**7.60**	-	①**7.54**
	Mac-Mic	9.93	12.72	13.34	12.34	14.48	13.01	19.05	9.04
Bi-objective NMF (this paper) Pareto Optimal	$\alpha=1$	8.29	25.26	11.14	24.18	15.88	37.01	24.08	36.44
	$\alpha=0.9$	①**4.80**	①**4.67**	①**6.22**	③6.83	12.58	③8.83	21.97	③8.44
	$\alpha=0.8$	②**5.34**	②**4.86**	②**6.40**	①**6.37**	②**11.78**	8.93	18.83	8.85
	$\alpha=0.7$	③6.19	③6.13	③6.95	②**6.76**	①**11.77**	8.85	17.36	9.10
	$\alpha=0.6$	7.10	7.81	7.49	7.62	③11.95	9.28	16.56	9.14
	$\alpha=0.5$	7.85	9.06	7.95	8.40	12.27	9.93	16.11	9.80
	$\alpha=0.4$	8.48	9.80	8.45	8.90	12.70	10.80	③15.46	10.45
	$\alpha=0.3$	9.16	10.59	8.90	9.36	13.10	11.72	②**15.19**	10.89
	$\alpha=0.2$	9.92	11.74	9.51	9.82	13.67	12.27	①**15.16**	11.14
	$\alpha=0.1$	10.95	13.00	10.34	10.60	14.42	12.93	15.57	11.08
	$\alpha=0$	12.32	17.55	12.54	15.54	15.22	13.83	16.42	12.32

TABLE IV: Performance on the Urban image ($\times 10^{-2}$)

		Spectral Angle Distance				
		SAD	Asphalt	Grass	Tree	Roof
	MiniDisCo	③30.23	25.91	①**25.62**	13.86	55.51
	ConvexNMF	34.83	48.15	47.87	14.29	35.01
	Mac-Mic	33.53	③10.78	②**43.65**	53.00	26.68
this paper	$\alpha=1$ ($\ell_{-\infty}$-norm)	40.84	87.29	60.03	①**7.92**	③8.14
Pareto opt.	$\alpha=0.48$ (ℓ_2-norm)	31.28	66.74	46.11	②**8.30**	①**3.95**
Pareto opt.	$\alpha=0.40$ (ℓ_∞-norm)	30.45	64.18	③44.95	③8.37	②**4.30**
Pareto opt.	$\alpha=0.04$ ($\ell_{-\infty}$-norm)	②**29.79**	①**8.34**	70.54	9.77	30.46
	$\alpha=0$	①**28.55**	②**8.93**	62.31	10.10	32.84

B. Simulation with synthetic data

The performance of the proposed method is firstly studied on a series of synthetic images, each of size 20×20 pixels. The generalized bilinear model (GBM) [26], is considered with

$$\boldsymbol{x}_t = \sum_{n=1}^{N} a_{nt}\,\boldsymbol{e}_n + \sum_{n=1}^{N-1}\sum_{m=n+1}^{N} \gamma_{nm}\, a_{nt}\, a_{mt}\,(\boldsymbol{e}_n \otimes \boldsymbol{e}_m) + \boldsymbol{n},$$

where $\gamma_{nm} \in [0,1]$ and $\boldsymbol{n} \in \Re^{L\times 1}$ is the additive noise. The data are generated as follows. First, $N=3$ or $N=6$ endmembers are randomly selected from the candidate spectra set. This set is composed by 19 spectra drawn from the United States Geological Survey (USGS) digital spectral library [48], as given in Fig. 4. Second, the abundance vectors are uniformly generated using a Dirichlet distribution on the simplex defined by the nonnegativity and the sum-to-one constraints [48]. Last, the data is corrupted with a Gaussian noise at two levels, with the signal-to-noise ratio of 30 dB and 15 dB.

Experiments are conducted employing the weight set $\alpha \in \{0, 0.1, \ldots, 0.9, 1\}$, which implies the model varying gradually from the nonlinear Gaussian NMF ($\alpha = 0$) to the conventional linear NMF ($\alpha = 1$). For each α from the weight set, the Algorithm 2 is applied. The maximum iteration number is set to $k_{\max} = 2000$ in all the comparing methods. The bandwidth parameter in the Gaussian kernel is roughly set as $\sigma = 3.0$ for all the experiments. By performing ten Monte-Carlo simulations, the average values in terms of SAD and RMSE are compared with the aforementioned unmixing approaches, as given in TABLE III.

We observe the following. For all the considered numbers of endmembers and noise levels, the proposed bi-objective NMF with the Pareto optimal outperforms not only state-of-the-art methods but also the linear ($\alpha = 1$) and Gaussian ($\alpha = 0$) NMF in terms of endmember estimation. Given relatively small number of endmembers with $N = 3$, the proposed method also yields the smallest root mean square error on the abundances regardless of the noise level. For $N = 6$, it provides comparable results to MercerNMF and MinDisCo, being slightly worse in terms of RMSE and slightly better in terms of SAD.

C. Experiments with Urban image

As depicted in Fig. 5, the real hyperspectral image studied is from the Urban image, acquired by the HYDICE sensor. The top left part with 150×150 pixels is taken from the original 307×307 pixels' image. The raw data consists of 210 channels covering the bandwidth from $0.4\mu m$ to $2.5\mu m$. As recommended in [49], only $L = 162$ bands of high-SNR are of interest. According to the ground truth provided in [49], [50], the studied area is mainly composed of four endmembers shown in Fig. 6: asphalt, grass, tree and roof. In experiments, the weight set is chosen as $\alpha \in \{0, 0.04, \ldots, 0.96, 1\}$, and the maximum iteration number is set to $k_{\max} = 300$. Starting from $\alpha_1 = 0$, the matrix $\boldsymbol{E}_1$ is initialized by conducting NMF on 1000 randomly chosen samples, while the elements in $\boldsymbol{A}_1$ are generated using a $[0, 1]$ uniform distribution. The bandwidth in the Gaussian kernel is selected as $\sigma = 4.2$, after preliminary analysis using the single-objective Gaussian NMF with the candidate set $\{0.2, 0.3, \ldots, 9.9, 10, 15, 20, \ldots, 50\}$.

The unmixing performance is shown in TABLE IV, with several ℓ_p-norms as described in Section IV-D. Methods that

Fig. 5: The scene from the Urban image

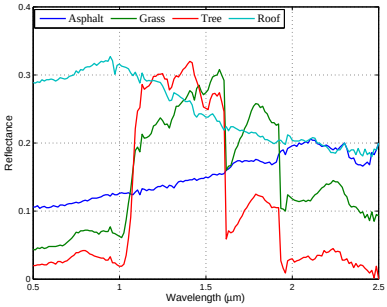

Fig. 6: The four ground truth endmembers in the Urban image.

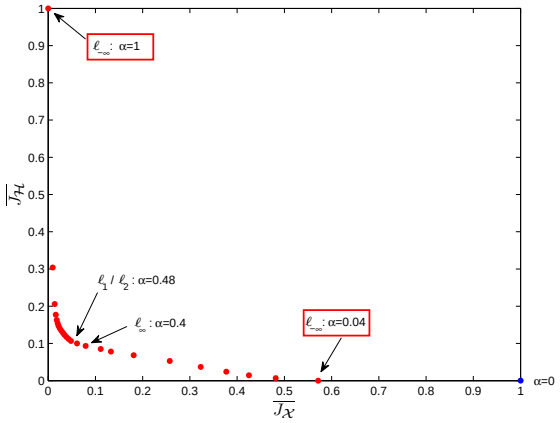

Fig. 7: Illustration of the approximated Pareto front in the objective space for the Urban image. The (normalized) objective vectors of the 25 non-dominated solutions, marked in red, approximate a part of the Pareto front; the dominated solutions are marked in blue.

do not extract endmembers are not included in this table, such as FCLS, sNMF, K-Hype, MercerNMF and KconvexNMF. Compared with the state-of-the-art methods, three endmembers out of four, *i.e.*, Asphalt, Tree and Roof, are better estimated by Pareto optima. The estimated abundance maps corresponding to the four endmembers are shown in Fig. 9.

We compare in TABLE V the computational time of the proposed method with the aforementioned unmixing algorithms that jointly estimate the endmembers and abundances. Nonlinear methods, and in particular kernel-based ones, are time-consuming in general. Regarding the proposed bi-objective NMF, its computational complexity is lower than the one of MercerNMF, for a fixed value of α. When considering a spread of values of α, the sub-optimization problems can be addressed in parallel.

TABLE V: Estimated computational time (in seconds)

nonlinear	MiniDisCo	220
	ConvexNMF	996
	KconvexNMF	2622
	MercerNMF	20332
	Mac-Mic	4244
	Bi-Objective NMF, average per α	5420

D. Approximating the Pareto front

Inherited from nonlinear multi-objective optimization problems, the determination of the whole Pareto front is intractable and the target becomes to approximate the Pareto front by a set of discrete points, as stated in [36]. To this end, we operate as follows: For each value of α, we obtain a solution (endmember and abundance matrices) from the proposed algorithm; by evaluating the objective functions $J_{\mathcal{X}}$ and $J_{\mathcal{H}}$ at this solution, we get a single point in the objective space, as shown in Fig. 7. The evolution of these objectives functions and the aggregated objective function J, evaluated at the solution obtained for each α, are shown in Fig. 8. We observe the following:

1) Regarding the sum-weighted approach, the minimizer of the sub-optimization problem is proven to be a Pareto optimal for the original multi-objective problem, *i.e.*, the corresponding objective vector belongs to the Pareto front in the objective space [38]. For the Urban image, we obtain 25 (out of 26) dominated solutions. The solution for $\alpha = 0$ is dominated by the solutions on the approximated Pareto front, with respect to both objectives. Such phenomenon is not surprising. Indeed, *there exist multiple Pareto optimal solutions in a problem only if the objectives are conflicting to each other*, as demonstrated in [51][3]. As shown in Fig. 7 and Fig. 8, the obtained solutions are Pareto optimal within the objectives-conflicting interval $\alpha \in [0.04, 1]$.
2) A uniform distribution of the values of α from $[0, 1]$ does not lead to a uniform spread of the solutions on the approximated Pareto front. Moreover, the nonconvex part of the Pareto front cannot be attained using any weight. These are two major drawbacks of the sum-weighted method, as stated in [38] and illustrated in Fig. 7.

Nevertheless, the obtained approximation of Pareto front is of high value. On one hand, it provides a set of non-dominated solutions for the DM. On the other hand, an insight of the tradeoff between objectives $J_{\mathcal{X}}$ and $J_{\mathcal{H}}$ reveals the underlying linearity/nonlinearity of the data under study.

VI. Conclusion

This paper presented a bi-objective nonnegative matrix factorization by exploiting the kernel machines, where the decomposition was performed simultaneously in the input and the feature spaces. The multiplicative update rules were

[3]For example, the Pareto optimal solutions for the well-known Schaffer's function, defined by $J(x) = [x^2 \quad (x-2)^2]$, are found only within the interval $[0, 2]$, where a tradeoff between two objectives exists. See [52].

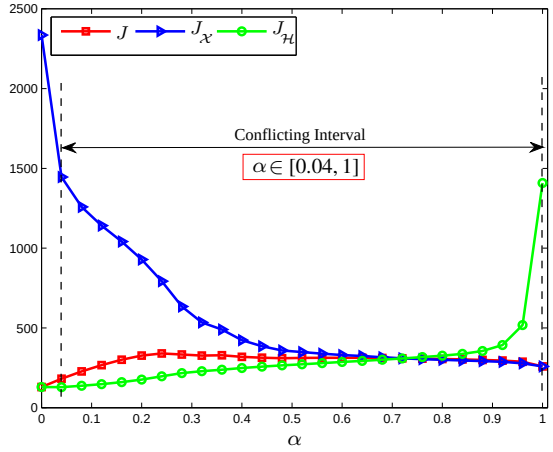

Fig. 8: Visualization of the tradeoff between the two objectives $J_{\mathcal{X}}$ and $J_{\mathcal{H}}$, and the change of the aggregated objective function J, along with the increment of α for the Urban image.

derived. The performance of the method was demonstrated for unmixing synthetic and real hyperspectral images. The approximation of the Pareto front was analyzed. Future work include a more efficient way to determine the good value of α. In addition, we will incorporate physical-based unmixing models, namely the bilinear ones and the macroscopic-microscopic models, by defining appropriately the kernel in the proposed framework. Considering simultaneously several kernels, and consequently several feature spaces, is also under investigation.

ACKNOWLEDGMENT

This work was supported by the French ANR, grant HYPANEMA: ANR-12BS03-0033.

REFERENCES

[1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[2] S. Jia and Y. Qian, "Constrained nonnegative matrix factorization for hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 1, pp. 161–173, Jan. 2009.

[3] R. Zhi, M. Flierl, Q. Ruan, and W. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 41, no. 1, pp. 38–52, Feb. 2011.

[4] P. M. Kim and B. Tidor, "Subsystem identification through dimensionality reduction of large-scale gene expression data." *Genome research*, vol. 13, no. 7, pp. 1706–1718, 2003.

[5] A. Cichocki, R. Zdunek, and S.-I. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5. IEEE, 2006, pp. V–V.

[6] C. Ding, X. He, and H. D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," in *Proc. SIAM Data Mining Conf*, 2005, pp. 606–610.

[7] P. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[8] Z. Chen and A. Cichocki, "Nonnegative matrix factorization with temporal smoothness and/or spatial decorrelation constraints," in *Laboratory for Advanced Brain Signal Processing, RIKEN, Tech. Rep*, 2005.

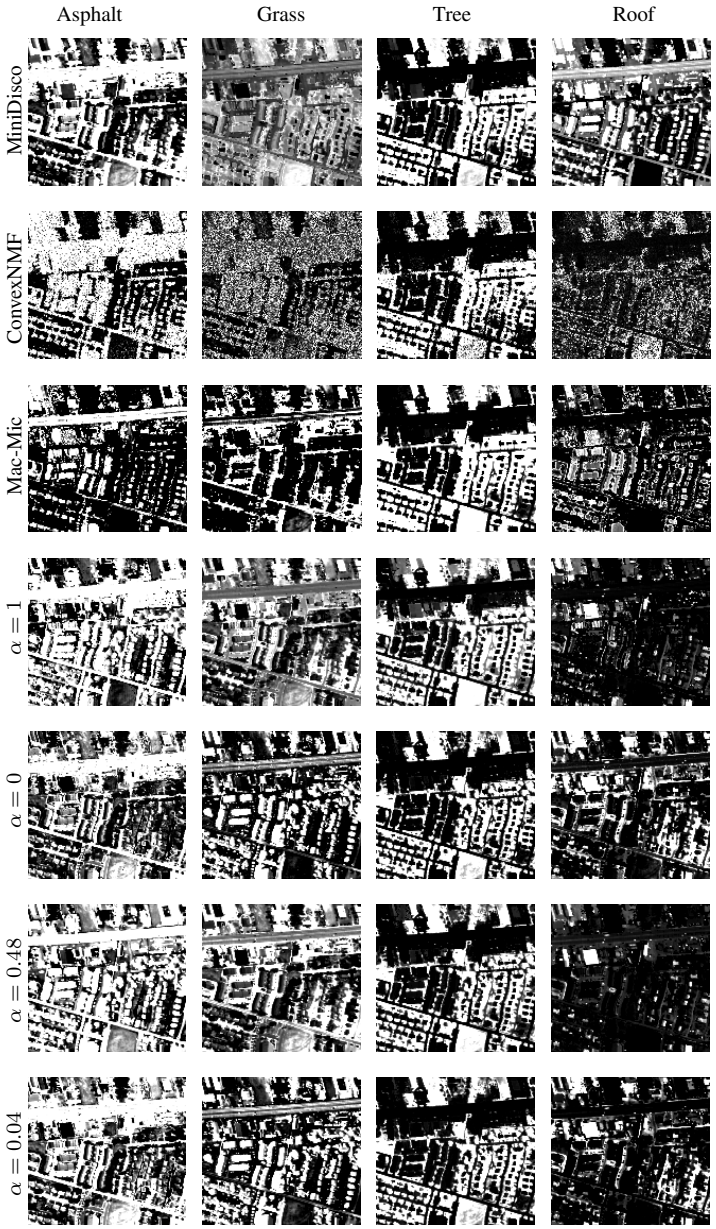

Fig. 9: Estimated abundance maps on the Urban image. Left to right: Abundance maps for Asphalt, Grass, Tree and Roof. Top to bottom: MiniDisco, ConvexNMF, Mac-Mic, and the proposed bi-objective NMF with $\alpha = 1$ (linear NMF), $\alpha = 0$ (nonlinear Gaussian NMF), and Pareto optimal solutions $\alpha = 0.48$ (ℓ_1/ℓ_2-norm), $\alpha = 0.04$ ($\ell_{-\infty}$-norm).

[9] A. Huck, M. Guillaume, and J. Blanc-Talon, "Minimum dispersion constrained nonnegative matrix factorization to unmix hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 6, pp. 2590–2602, Jun. 2010.

[10] J. Chen, C. Richard, and P. Honeine, "Nonlinear estimation of material abundances of hyperspectral images with ℓ_1-norm spatial regularization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, pp. 2654–2665, May 2014.

[11] ——, "Nonlinear unmixing of hyperspectral data based on a linear-mixture/nonlinear-fluctuation model," *IEEE Transactions on Signal Processing*, vol. 61, no. 2, pp. 480–492, Jan. 2013.

[12] N. Nguyen, J. Chen, C. Richard, P. Honeine, and C. Theys, "Supervised nonlinear unmixing of hyperspectral images using a pre-image method," in *New Concepts in Imaging: Optical and Statistical Models, In Eds.*

D. Mary, C. Theys, and C. Aime, ser. EAS Publications Series. EDP Sciences, 2013, vol. 59, pp. 417–437.

[13] D. Zhang, Z. Zhou, and S. Chen, "Non-negative matrix factorization on kernels," in *Lecture Notes in Computer Science*, vol. 4099. Springer, 2006, pp. 404–412.

[14] I. Buciu, N. Nikolaidis, and I. Pitas, "Nonnegative matrix factorization in polynomial feature space," *IEEE Transactions on Neural Networks*, vol. 19, no. 6, pp. 1090–1100, Jun. 2008.

[15] H. Lee, A. Cichocki, and S. Choi, "Kernel nonnegative matrix factorization for spectral EEG feature extraction," *Neurocomputing*, vol. 72, no. 1315, pp. 3182 – 3190, 2009, hybrid Learning Machines (HAIS 2007) / Recent Developments in Natural Computation (ICNC 2007).

[16] Y. Li and A. Ngom, "A new kernel non-negative matrix factorization and its application in microarray data analysis," in *2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, San Diego, CA, USA, May. 2012, pp. 371–378.

[17] P. Honeine and C. Richard, "Preimage problem in kernel-based machine learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 77–88, 2011.

[18] F. Zhu, P. Honeine, and M. Kallas, "Kernel non-negative matrix factorization without the pre-image problem," in *IEEE workshop on Machine Learning for Signal Processing*, Reims, France, Sep. 2014.

[19] ——, "Kernel nonnegative matrix factorization without the curse of the pre-image," *ArXiv*, http://arxiv.org/abs/1501.05684 2014.

[20] J. Chen, C. Richard, and P. Honeine, "Estimating abundance fractions of materials in hyperspectral images by fitting a post-nonlinear mixing model," in *Proc. IEEE Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing*, Jun. 2013.

[21] ——, "Nonlinear unmixing of hyperspectral images based on multi-kernel learning," in *Proc. IEEE Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing*, Jun. 2012.

[22] Y. Altmann, N. Dobigeon, S. McLaughlin, and J.-Y. Tourneret, "Residual component analysis of hyperspectral images - application to joint nonlinear unmixing and nonlinearity detection." *IEEE Transactions on Image Processing*, pp. 2148–2158, 2014.

[23] J. Nascimento and J. Bioucas Dias, "Vertex component analysis: a fast algorithm to unmix hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 4, pp. 898–910, Apr. 2005.

[24] R. Heylen, M. Parente, and P. Gader, "A review of nonlinear hyperspectral unmixing methods," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 7, no. 6, pp. 1844–1868, June 2014.

[25] N. Dobigeon, J.-Y. Tourneret, C. Richard, J. Bermudez, S. McLaughlin, and A. Hero, "Nonlinear unmixing of hyperspectral images: Models and algorithms," *Signal Processing Magazine, IEEE*, vol. 31, no. 1, pp. 82–94, Jan 2014.

[26] A. Halimi, Y. Altmann, N. Dobigeon, and J. Tourneret, "Nonlinear unmixing of hyperspectral images using a generalized bilinear model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 11, pp. 4153–4162, Nov. 2011.

[27] Y. Altmann, A. Halimi, N. Dobigeon, and J.-Y. Tourneret, "Supervised nonlinear spectral unmixing using a postnonlinear mixing model for hyperspectral imagery," *Image Processing, IEEE Transactions on*, vol. 21, no. 6, pp. 3017–3025, 2012.

[28] N. Yokoya, J. Chanussot, and A. Iwasaki, "Nonlinear unmixing of hyperspectral data using semi-nonnegative matrix factorization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 2, pp. 1430–1437, Feb. 2014.

[29] B. Hapke, "Bidirectional reflectance spectroscopy: 1. theory," *Journal of Geophysical Research: Solid Earth (1978–2012)*, vol. 86, no. B4, pp. 3039–3054, 1981.

[30] R. Close, P. Gader, and J. Wilson, "Hyperspectral unmixing using macroscopic and microscopic mixture models," *Journal of Applied Remote Sensing*, vol. 8, no. 1, pp. 083 642–083 642, 2014.

[31] D. Heinz and C. Chang, "Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 3, pp. 529–545, Mar. 2001.

[32] M. Winter, "N-FINDR: an algorithm for fast autonomous spectral end-member determination in hyperspectral data: an algorithm for fast autonomous spectral end-member determination in hyperspectral data," *Proc. of SPIE: Imaging Spectrometry V*, vol. 3753, no. 10, 1999.

[33] C. Ding, T. Li, and M. I. Jordan, "Convex and Semi-Nonnegative Matrix Factorizations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 45–55, Nov. 2010.

[34] B. Pan, J. Lai, and W. Chen, "Nonlinear nonnegative matrix factorization based on Mercer kernel construction," *Pattern Recognition*, vol. 44, no. 10-11, pp. 2800 – 2810, 2011.

[35] D. Dranishnikov, P. Gader, A. Zare, and T. Glenn, "Unmixing using a combined microscopic and macroscopic mixture model with distinct endmembers," in *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*. IEEE, 2013, pp. 1–5.

[36] J. Lampinen, "Multiobjective nonlinear pareto-optimization," *Preinvestigation Report, Lappeenranta University of Technology*, 2000.

[37] K. Miettinen, "Introduction to multiobjective optimization: Noninteractive approaches," in *Multiobjective Optimization*. Springer, 2008, pp. 1–26.

[38] I. Das and J. Dennis, "A closer look at drawbacks of minimizing weighted sums of objectives for pareto set generation in multicriteria optimization problems," *Structural optimization*, vol. 14, no. 1, pp. 63–69, 1997.

[39] J. Ryu, S. Kim, and H. Wan, "Pareto front approximation with adaptive weighted sum method in multiobjective simulation optimization," in *Simulation Conference (WSC), Proceedings of the 2009 Winter*, Dec. 2010, pp. 623–633.

[40] S. A. Vavasis, "On the complexity of nonnegative matrix factorization," *SIAM Journal on Optimization*, vol. 20, no. 3, pp. 1364–1377, 2009.

[41] C. Lin, "Projected gradient methods for non-negative matrix factorization," Neural Computation, Tech. Rep., 2007.

[42] M. E. Daube-Witherspoon and G. Muehllehner, "An iterative image space reconstruction algorthm suitable for volume ect," *IEEE Transactions on Medical Imaging*, vol. 5, no. 2, pp. 61–66, June 1986.

[43] C.-J. Lin, "On the convergence of multiplicative update algorithms for nonnegative matrix factorization," *Neural Networks, IEEE Transactions on*, vol. 18, no. 6, pp. 1589–1596, Nov 2007.

[44] E. Gonzales and Y. Zhang, "Accelerating the Lee-Seung algorithm for non-negative matrix factorization," Department of Computational and Applied Mathematics, Rice University, Tech. Rep., 2005.

[45] X. Blasco, J. Herrero, J. Sanchis, and M. Martnez, "A new graphical visualization of n-dimensional pareto front for decision-making in multiobjective optimization," *Information Sciences*, vol. 178, no. 20, pp. 3908 – 3924, 2008, special Issue on Industrial Applications of Neural Networks 10th Engineering Applications of Neural Networks 2007.

[46] P. Honeine and C. Richard, "Geometric unmixing of large hyperspectral images: a barycentric coordinate approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 6, pp. 2185–2195, Jun. 2012.

[47] L. Miao and H. Qi, "Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 3, pp. 765–777, March 2007.

[48] J. M. Bioucas-Dias and J. M. P. Nascimento, "Hyperspectral subspace identification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 8, pp. 2435–2445, Aug 2008.

[49] S. Jia and Y. Qian, "Spectral and spatial complexity-based hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 12, pp. 3867–3879, Dec. 2007.

[50] M. Fong and Z. Hu, "Hyperactive: Hyperspectral image analysis toolkit," *UCLA Dept of Math, http://www. math. ucla. edu/wittman/hyper/hypermanual. pdf, accessed Apr.*, vol. 3, 2011.

[51] K. Deb and D. Kalyanmoy, *Multi-Objective Optimization Using Evolutionary Algorithms*. New York, NY, USA: John Wiley & Sons, Inc., 2001.

[52] E. Zitzler and L. Thiele, "Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach," *IEEE Transactions on Evolutionary Computation*, vol. 3, no. 4, pp. 257–271, Nov. 1999.