A Convex Model for Nonnegative Matrix Factorization and Dimensionality Reduction on Physical Space

Ernie Esser, Michael Möller, Stanley Osher, Guillermo Sapiro, *Senior Member, IEEE*, and Jack Xin

***Abstract*—A collaborative convex framework for factoring a data matrix X into a nonnegative product AS, with a sparse coefficient matrix S, is proposed. We restrict the columns of the dictionary matrix A to coincide with certain columns of the data matrix X, thereby guaranteeing a physically meaningful dictionary and dimensionality reduction. We use $l_{1,\infty}$ regularization to select the dictionary from the data and show that this leads to an exact convex relaxation of l_0 in the case of distinct noise-free data. We also show how to relax the restriction-to-X constraint by initializing an alternating minimization approach with the solution of the convex model, obtaining a dictionary close to but not necessarily in X. We focus on applications of the proposed framework to hyperspectral endmember and abundance identification and also show an application to blind source separation of nuclear magnetic resonance data.**

***Index Terms*—Blind source separation (BSS), dictionary learning, dimensionality reduction, hyperspectral endmember detection, nonnegative matrix factorization (NMF), subset selection.**

I. Introduction

DIMENSIONALITY reduction has been widely studied in the signal processing and computational learning communities. One of the major drawbacks of virtually all popular approaches for dimensionality reduction is the lack of physical meaning in the reduced dimension space. This significantly reduces the applicability of such methods. In this paper, we present a framework for dimensionality reduction, based on matrix factorization and sparsity theory, that uses the data itself (or small variations from it) for the low-dimensional representation, thereby guaranteeing physical fidelity. We propose a new convex method to factor a nonnegative data matrix X into a product AS, for which S is nonnegative and sparse and the columns of A coincide with columns from the data matrix X.

The organization of this paper is as follows. In the remainder of the introduction, we further explain the problem, summarize our approach, and discuss applications and related work. In Section II, we present our proposed convex model for *endmember* (dictionary) computation that uses $l_{1,\infty}$ regularization to select as endmembers a sparse subset of columns of X, such that sparse nonnegative linear combinations of them are capable of representing all other columns. Section III shows that, in the case of distinct noise-free data, $l_{1,\infty}$ regularization is an exact relaxation of the ideal row-0 norm (number of nonzero rows) and furthermore proves the stability of our method in the noisy case. Section IV presents numerical results for both synthetic and real hyperspectral data. In Section V, we present an extension of our convex endmember detection model that is better able to handle outliers in the data. We discuss its numerical optimization, compare its performance to the basic model, and also demonstrate its application to a blind source separation (BSS) problem based on nuclear magnetic resonance (NMR) spectroscopy data.

A. Summary of the Problem and Geometric Interpretation

The underlying general problem of representing $X \approx AS$ with $A, S \geq 0$ is known as nonnegative matrix factorization (NMF). Variational models for solving NMF problems are typically nonconvex and are solved by estimating A and S alternatingly. Although variants of alternating minimization methods for NMF often produce good results in practice, they are not guaranteed to converge to a global minimum.

The problem can be greatly simplified by assuming a partial orthogonality condition on matrix S as is done in [1] and [2]. More precisely, the assumption is that, for each row i of S, there exists some column j such that $S_{i,j} > 0$ and $S_{k,j} = 0$ for $k \neq i$. Under this assumption, NMF has a simple geometric interpretation. Not only should the columns of A appear in the data X up to scaling but the remaining data should be expressible as nonnegative linear combinations of these columns. Therefore, the problem of finding A is to find columns in X, preferably as few as possible, that span a cone containing the rest of the data X. Fig. 1 illustrates the geometry in three dimensions.

The problem we actually want to solve is more difficult than NMF in a couple respects. One reason is the need to deal with noisy data. While NMF by itself is a difficult problem already, the identification of the vectors becomes even more difficult if the data X contain noise and we need to find a low-dimensional cone that contains most of the data (see the lower right image in Fig. 1). Notice that in the noisy case, finding vectors such that all data are contained in the cone they span would lead to a drastic

Manuscript received February 04, 2011; revised October 21, 2011; accepted February 19, 2012. Date of publication March 06, 2012; date of current version June 13, 2012. The work of E. Esser and J. Xin was supported in part by the NSF under Grant DMS-0911277 and Grant DMS-0928427. The work of M. Möller and S. Osher was supported in part by the NSF under Grant DMS-0835863, Grant DMS-0914561, and Grant DMS-0914856, and in part by the ONR under Grant N00014-08-1119. The work of M. Möller was also supported by the German Academic Exchange Service (DAAD). The work of G. Sapiro was supported by the NSF, NGA, ONR, ARO, DARPA, and NSSEFF. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Birsen Yazici.

E. Esser and J. Xin are with the Department of Mathematics, University of California at Irvine, Irvine, CA 92697-3875 USA (e-mail: eesser@math.uci.edu).

M. Möller was with the Department of Mathematics, University of California at Los Angeles, Los Angeles, CA 90095-1555 USA. He is now with the Westfälische Wilhelms University of Münster, 48149 Münster, Germany.

S. Osher is with the Department of Mathematics, University of California at Los Angeles, Los Angeles, CA 90095-1555 USA.

G. Sapiro is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455-0170 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2012.2190081

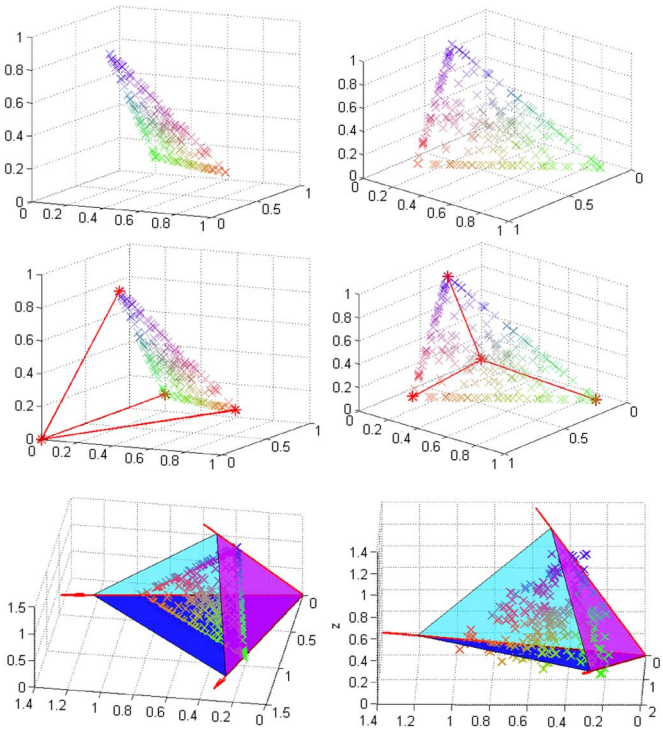

Fig. 1. Geometric interpretation of the endmember detection problem. (First row) Two different viewpoints for a data set in three dimensions. (Second row) Same data set with the vectors that can express any data point as a nonnegative linear combination in red (in online version). (Third row, left) Cone spanned by the three vectors containing all data. (Third row, right) Illustration of the cone with Gaussian noise added to the data; in this case, not all points lie inside the cone anymore.

overestimation of the number of vectors. Arbitrarily small noise at a single data point could already lead to including this vector into the set of cone spanning vectors. Thus, the problem is ill posed and regularization is needed to handle noise. In addition to small noise, there could be also outliers in the data, i.e., columns of X that are not close to being well represented as a nonnegative linear combination of other columns, but that we do not wish to include in A. Such outliers could arise from bad sensor measurements, nonphysical artifacts, or any features that, for some reason, we are not interested in including in our dictionary A. Another complication that requires additional modeling is that, for the applications we consider, matrix S should be also sparse, which means that we want the data to be represented as sparse nonnegative linear combinations of the columns of A.

B. Our Proposed Approach

We obtain the $X = AS$ factorization by formulating the problem as finding a sparse nonnegative T such that $X \approx XT$ and as many rows as possible of T are entirely zero. We want to encourage this so-called *row sparsity* of T in order to select as few as possible data points as dictionary atoms. We do this by using $l_{1,\infty}$ regularization. This type of regularization couples the elements in each row of T and is based on the recent ideas of collaborative sparsity (see, for example, [3] and the references therein). The particular $l_{1,\infty}$ regularization has been studied by Tropp in [4] but without considering nonnegativity constraints and also not in the setting of finding a T such that $X \approx XT$ for physically meaningful dimensionality reduction. A strong motivation for using $l_{1,\infty}$ regularization instead of other row sparsity regularizers such as $l_{1,2}$ is that, in the case of distinct noise-free data, the $l_{1,\infty}$ model is an exact relaxation of minimizing the number of nonzero rows in T such that $X = XT$. This exact relaxation is independent of the coherence of the columns of X. Without the nonnegativity constraint, low coherence is crucial, as shown in [4]. The general setting $X \approx XT$ was proposed by Lin *et al.* in [5] for low-rank approximation with the nuclear norm. However, the nuclear norm does not lead to row sparsity, and thus, there is no obvious way to extract dictionary atoms from the minimizer. Both the nuclear norm and $l_{1,\infty}$ approaches are addressing related but different problems. Our main contribution is to apply the joint sparsity regularizer $l_{1,\infty}$ to the nonnegative factorization setting $X \approx XT$ and thereby implicitly select certain columns of X for the description of all columns of X. We pose this as a convex optimization problem in T. For practical reasons, we will need to perform some preliminary data reduction, explained in Section II, before carrying out convex minimization. The main idea, however, is to minimize over $T \geq 0$ the $l_{1,\infty}$ norm of T plus some convex penalty on $X - XT$. In the simplest case, we penalize $\|X - XT\|_F^2$. We also propose an advanced noise model to handle the case where X contains outliers. Both models also incorporate a weighted l_1 penalty to encourage a sparser T so that from the few columns of X selected to represent the whole data, only a few are used per sample.

C. Applications and Related Work

Although we concentrate on hyperspectral imaging (HSI) and briefly discuss an application to BSS, our method is applicable in numerous areas, from biology to sensor networks.

For instance, one approach to text mining, which is the technique of extracting important information from large text data sets, is to reduce the number of relevant text documents by clustering them into content-dependent categories using NMF (see [6] for details). Our approach could be potentially used to not only cluster the large amount of documents by a similar factorization, but due to the dictionary being a part of the data, it would furthermore lead to a correspondence of every atom in the dictionary to a certain document. This correspondence might help a human analyzer judge the importance of each cluster. Therefore, the physical meaning of the dictionary atoms would have an immediate advantage for further analysis.

As another example, in [7], NMF is applied to spectrograms of many different musical sounds in order to obtain a spectral basis for musical genre classification. Again, the physical fidelity of our approach could be interesting since it would provide a correspondence of each spectral basis element to a certain musical sound.

From the numerous potential applications, we concentrate on two to illustrate the proposed framework. We next describe the challenges in these applications. This early description of these applications will help to further motivate the work.

1) Introduction to HSI: HSI sensors record up to several hundred different frequencies in the visible, near-infrared, and infrared spectra. This precise spectral information provides some insight on the material at each pixel in the image. Due to relatively low spatial resolution and the presence of multiple materials at a single location (e.g., tree canopies above ground or water and water sediments), many pixels in the image contain the mixed spectral signatures of multiple materials. The task of

determining the *abundance values* (presence quantity) of different materials in each pixel is called *spectral unmixing*. This is clearly an ill-posed problem that requires some assumptions and data models.

Unmixing requires a dictionary with the spectral signatures of the possible materials (often denoted as *endmembers*). Since these dictionaries can be difficult to obtain and might depend on the conditions under which they were recorded, it is sometimes desirable to automatically extract suitable endmembers from the image one wants to demix in a process called *endmember detection*. Many different techniques for endmember detection have been proposed, including variational [8]–[10] and Bayesian [11]–[13] models as well as convex geometry-based methods [1], [2], [14]. Related although not yet applied to endmember detection are subset selection methods such as the rank-revealing QR (RRQR) decomposition (e.g., [15] and [16]) and greedy dictionary selection methods [17], [18]. The general principle behind RRQR for subset selection is to find a column permutation of the data matrix such that the first few columns are as well conditioned as possible [16]. In a sense, it looks for the most linearly independent columns of a matrix.[1] However, unlike our approach, QR methods do not take nonnegativity constraints into account. The greedy dictionary selection methods [17], [18] are based on greedy approximations of combinatorial optimization problems for sparse dictionary selection. These methods come with theoretical guarantees given incoherence assumptions about the data matrix. Such assumptions do not hold for our data, but because we have a nonnegativity constraint not assumed there, we can show that our model is theoretically sound without strong incoherence assumptions on the data, as demonstrated by the exact convex relaxation discussed in Section III.

Simultaneously detecting endmembers and computing abundance values can be stated as factoring the data matrix $X \in \mathbb{R}^{m,d}$ into $X \approx AS$, $A, S \geq 0$, with both $A \in \mathbb{R}^{m,n}$ and $S \in \mathbb{R}^{n,d}$ being unknown. In this notation, each column of X is the spectral signature of one pixel in the image. Hence, m is the number of spectral bands, d is the total number of pixels, and each of the n columns of A represents one endmember. Abundance matrix S contains the amounts of each material in A at each pixel in X. The application of NMF to hyperspectral endmember detection can be found for instance in [8] and [9].

Considering that while material mixtures in HSI exist, it is unlikely that pixels contain a mixture of all or many of the materials in A; researchers have recently focused on encouraging sparsity on the abundance matrix S [20], [21]. Motivated by the ideas of dictionary learning for sparse coding, the works [22], [23] were proposed explicitly to look for endmember matrices that lead to sparse abundance maps. We follow a similar idea, although our method will be fundamentally different in two aspects: First, we restrict columns of our dictionary A to appear somewhere in the data X. This is a common working hypothesis for moderate ground sampling distance images and is called *pixel purity* assumption. It corresponds to the partial orthogonality assumption on S previously discussed. In the general context of dictionary learning and NMF, it guarantees the columns of A to be physically meaningful. As aforementioned, the lack of physical interpretation has been a critical shortcoming of standard dimensionality reduction and dictionary learning techniques, and it is not yet addressed in these areas of research. Second, choosing the dictionary columns from the data enables us to propose a convex model and hence avoid the problem of saddle points or local minima.

Different areas of applications use different terminologies for mathematically similar things. Throughout this paper, we will use the terminology of hyperspectral unmixing to explain and motivate our model, although its application is not exclusively in the field of HSI. For instance, an endmember could be any abstract column of the dictionary matrix A with a different physical meaning depending on its context. However, we think it aids the clarity and understanding of the model to use the HSI example throughout the explanations. To show that our model is not limited to the hyperspectral case, we also present results for the problem of BSS, which we will briefly describe in the next section.

2) Introduction to BSS: BSS is the general problem of recovering unknown source signals from measurements of their mixtures, where the mixing process is unknown but often assumed to be linear. Examples include demixing audio signals and NMR spectroscopy. Many BSS problems have the same linear mixing model $X = AS$ that we are using for hyperspectral endmember detection and unmixing, often also with a nonnegativity constraint on S [1]. Here, the rows of S represent the source signals, A is the unknown mixing matrix, and the rows of X are the measured signals. The analogy of the hyperspectral pixel purity assumption can be also applied to some BSS problems [1]. The interpretation is that, for each source, there is some place where among all the sources only that source is nonzero. Thus, up to scaling, the columns of A should appear somewhere in the columns of the data matrix X and we can use the same algorithm we use for endmember detection.

II. Convex Endmember Detection Model

A. Convexification of Matrix Factorization Using the Pixel Purity Assumption

As aforementioned, we are assuming at first that the endmembers can be found somewhere in the data X. This assumption will enable us to propose a convex model for factoring the data X into a product AS, i.e., a problem usually tackled by nonconvex optimization techniques. We assume that there exists an index set I such that the columns X_i of X are endmembers for $i \in I$. Under the assumption of nonnegative linear mixing of signals, this means that any column X_j in X can be written as

$$X_j = \sum_{i \in I} X_i T_{i,j} \tag{1}$$

for coefficients $T_{i,j} \geq 0$. The problem is that the coefficients $T_{i,j}$, as well as the index set I, are unknown. Hence, we start by using all columns in X to describe X itself, i.e., we look for coefficients $T \geq 0$ for which

$$X = XT. \tag{2}$$

Notice that (2) has many solutions. However, we know that the desired representation uses as few columns of X as possible, i.e., only the endmembers. Since not using the jth column in the above formulation corresponds to having the entire jth row of T

[1]Independent of the work described here, L. Balzano, R. Nowak, and W. Bajwa developed a related matrix factorization technique and connected and compared it to RRQR. We thank Balzano *et al.* for pointing out their work [19] and also the possible relationships with RRQR.

be zero, we can reformulate the endmember detection problem as finding a solution to (2) such that as many rows of T as possible are zero. Mathematically

$$\min_{T\geq 0} \|T\|_{\text{row-0}} \text{ such that } XT = X \tag{3}$$

where $\|T\|_{\text{row-0}}$ denotes the number of nonzero rows. The columns of X that correspond to nonzero rows of the minimizer T of (3) are the desired endmembers (which define the lowest dimensional subspace where the data resides). Since problem (3) is not convex, we relax the above formulation by replacing $\|T\|_{\text{row-0}}$ by the convex $l_{1,\infty}$ norm $\|T\|_{1,\infty} = \sum_i \max_j |T_{i,j}|$. The l_1 part of the norm should encourage sparsity. Combined with the maximum norm, we can expect that the vector of maxima among each row becomes sparse, which means that we obtain row sparsity. Notice that we could have chosen other row sparsity encouraging regularizers such as $l_{1,2}$. However, we will prove in Section III that the $l_{1,\infty}$ relaxation is exact in the case of normalized nonrepeating data, which makes it clearly preferable.

As mentioned in the previous section, it is important to take noise into account, and therefore, equality (2) is too restrictive for real data. Hence, we will introduce a parameter that negotiates between having many zero rows (also called *row sparsity*) and good approximation of the data $\|X - XT\|_F^2$ in the Frobenius norm. Furthermore, for T to be a good solution, not only should most of its rows be zero but the nonzero rows should be also sparse since sparsity of the coefficient matrix reflects physically reasonable prior knowledge. For hyperspectral unmixing, the additional sparsity requirement on T reflects the assumption that most pixels are not mixtures of all the selected endmembers, but rather just a few. Thus, we will add an additional weighted l_1 term to incorporate this second type of sparsity (see [24] for a model combining structured and collaborative sparsity with individual sparsity).

B. Data Reduction as Preprocessing

It is already clear from (2) that the problem is too large because the unknown matrix T is a $d \times d$ matrix, where d is the number of pixels. Thus, before proceeding with the proposed convex model, we reduce the size of the problem by using clustering to choose a subset of candidate endmembers Y from the columns of X and a submatrix $X_s \in \mathbb{R}^{m\times d_s}$ of X for the data with $d_s \leq d$. In other words, we want to reformulate the problem as $YT \approx X_s$, where $X_s \in \mathbb{R}^{m\times d_s}$, $Y \in \mathbb{R}^{m\times n_c}$, $T \in \mathbb{R}^{n_c\times d_s}$, $n_c \ll d$, and $d_s \leq d$. We use $X_s = Y$ in our experiments but could also include more or even all of the data. We use k-means with a farthest first initialization to select Y. Other data reduction procedures could be used. An angle constraint $\langle Y_i, Y_j\rangle < 0.995$ ensures that the endmember candidates, namely, the columns of Y, are sufficiently distinct. An acceptable angle constraint could be probably estimated from assumptions about the noise and mixing based on the principle that slightly perturbed versions of the same endmember should not be in Y. However, we found it much easier to simply treat the angle constraint as a parameter, and 0.995 worked well for all the HSI experiments. An upper bound is placed on the number of allowable clusters so that the size of the problem is reasonable. We then propose a convex model for the more manageable problem of finding a nonnegative T such that $YT \approx X_s$, with T having the same sparsity properties described above. Note that we have not convexified the problem by prefixing the dictionary Y. This is done simply to work with manageable dimensions and data sets. Our convex model will still select the endmembers as a subset of this reduced data set Y, namely, the columns of Y that will correspond to nonzero rows of T.

C. Endmember Selection Model

Our model consists of a data fidelity term and two terms that encourage the desired sparsity in T. For simplicity, we consider the data fidelity term

$$\frac{\beta}{2}\|(YT - X_s)C_w\|_F^2 \tag{4}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, β is a positive constant, and $C_w \in \mathbb{R}^{d_s\times d_s}$ is a diagonal matrix that we can use to weight the columns of $(YT - X_s)$ so that it reflects the density of the original data. As mentioned earlier, we encourage rows of T to be zero by penalizing with $\zeta\|T\|_{1,\infty}$, which for nonnegative T is equal to $\zeta\sum_i \max_j(T_{i,j})$, with ζ as a positive constant, so that only a few samples are cooperatively selected as endmembers.[2] This kind of collaborative/structured sparsity regularizer has been proposed in several previous works, for example, [4] and [25].

To encourage sparsity of the nonnegative T, we use a weighted l_1 norm, $\langle R_w\sigma C_w, T\rangle$. Here, R_w is a diagonal matrix of row weights. We choose R_w to be the identity in our experiments, but if for example we wanted to encourage selection of endmembers toward the outside of the cone containing the data, we could also choose these weights to be proportional to $\langle Y_j, \bar{Y}\rangle$, where $\bar{Y}$ is the average of the columns of Y. This would encourage the method to prefer endmembers further away from the average $\bar{Y}$. Weighting matrix σ has the same dimension as T, and the weights are chosen to be

$$\sigma_{i,j} = \nu\left(1 - e^{\frac{-\left(1-(Y^TX_s)_{i,j}\right)^2}{2h^2}}\right) \tag{5}$$

for constants h and ν. This means that $\sigma_{i,j}$ is small when the ith column of Y is similar to the jth column of X_s and larger when they are dissimilar. This choice of weights encourages sparsity of T without impeding the effectiveness of the other regularizer. Since the smallest weight in each row occurs at the most similar data, this helps ensure a large entry in each nonzero row of T, which makes the row sparsity term more meaningful. It also still allows elements in a column of T to be easily shifted between rows corresponding to more similar endmember candidates, which can result in a reduction of $\zeta\sum_i \max_j(T_{i,j})$ without significantly affecting the weighted l_1 term. Since the weighted l_1 penalty here is really just a linear term, it cannot be said to directly enforce sparsity, but by encouraging each column of T to sum to something closer to one, the weighted l_1 penalty prefers data to be represented by nearby endmember candidates when possible, and this often results in a sparser matrix T. Overall, the proposed convex model is given by

$$\min_{T\geq 0} \zeta\sum_i \max_j(T_{i,j}) + \langle R_w\sigma C_w, T\rangle + \frac{\beta}{2}\|(YT - X_s)C_w\|_F^2. \tag{6}$$

[2]The aforementioned work by Balzano, Nowak, and Bajwa uses block orthogonal matching pursuit and also mentions the possible use of $\|\cdot\|_2$ instead of $\|\cdot\|_\infty$.

For our experiments, we normalize the columns of X to have unit l_2 norm so that we discriminate based solely on spectral signatures and not intensity.

D. Refinement of Solution

Since we are using a convex model to detect endmembers, it cannot distinguish between identical or very similar endmember candidates, which we will discuss from a more theoretical point of view in Section III. However, the model works very reliably when the columns of Y are sufficiently distinct, which they are by construction. A limitation is that the convex model is unable to choose as endmembers any data not represented in Y. Nonetheless, as shown in Section IV, the results of this approach already compare favorably to other methods. Even if some endmembers are not in Y, which could be due either to data reduction preprocessing or to a failure of the pixel purity assumption, the underlying problem of representing the data in terms of the available endmember candidates is still well posed. The convex model clearly cannot recover endmembers that are not present in Y, but it should still recover those that are present and represent the data as sparse nonnegative linear combinations of a small number of columns of Y. Moreover, it provides an excellent initialization for the alternating minimization approach to NMF, which can be used to further refine the solution if desired. Letting $\tilde{A}$ be the endmembers selected by the convex model, the solution is refined by alternately minimizing

$$\min_{A\geq 0, S\geq 0, \|A_j-\tilde{A}_j\|_2<a_j} \frac{1}{2}\|AS-X\|_F^2+\langle R_w\sigma, S\rangle \qquad (7)$$

and renormalizing the columns of A after each iteration. Here, a_j is the diameter of the jth cluster containing the data near $\tilde{A}_j$ and ensures that the refined endmembers obtained by this alternating approach cannot be too different from those already selected by the convex model, thereby remaining as close as desired to the physical space.

To recover the full abundance matrix S without refining $\tilde{A}$, we can solve the convex minimization problem in (7) for S using the full data matrix $X\in\mathbb{R}^{m\times d}$ and the original endmembers $\tilde{A}\in\mathbb{R}^{m\times n}$ selected by the convex model.

E. Numerical Optimization

We use the alternating direction method of multipliers (ADMM) [26], [27] to solve (6) by finding a saddle point of the augmented Lagrangian

$$\begin{aligned} L_\delta(Z,T,P) = {} & g_{\geq 0}(T)+\zeta\sum_i \max_j(T_{i,j}) \\ & +\langle R_w\sigma C_w, T\rangle+\frac{\beta}{2}\|(YZ-X_s)C_w\|_F^2 \\ & +\langle P, Z-T\rangle+\frac{\delta}{2}\|Z-T\|_F^2 \end{aligned}$$

where $g_{\geq 0}$ is an indicator function for the $T\geq 0$ constraint.

In iteration $k+1$, the algorithm proceeds by minimizing first $L_\delta(Z,T^k,P^k)$ with respect to Z to get Z^{k+1}, then minimizing $L_\delta(T,Z^{k+1},P^k)$ with respect to T to get T^{k+1}, and then updating the Lagrange multiplier P by $P^{k+1}=P^k+\delta(Z^{k+1}-T^{k+1})$. Each minimization step is straightforward to compute, and the algorithm is guaranteed to converge for any $\delta>0$.

Note that it is faster to precompute and store the inverse involved in the update for Z, but this can be overly memory intensive if Y has many columns and $C_w\neq I$. One could use a more explicit variant of ADMM, such as the method in [28], to avoid this difficulty if a larger number of columns of Y is desired. Here, we have restricted this number n_c to be less than 150 and ADMM can be reasonably applied.

In addition, note that the minimization problem for T^{k+1}

$$\begin{aligned} T^{k+1} = {} & \arg\min_T g_{\geq 0}(T)+\zeta\sum_i \max_j(T_{i,j}) \\ & +\frac{\delta}{2}\left\|T-Z^{k+1}-\frac{P^k}{\delta}+\frac{R_w\sigma C_w}{\delta}\right\|_F^2 \qquad (8) \end{aligned}$$

decouples into separate problems for each row T_i. The Legendre transformation of $g_{\geq 0}(T_i)+\zeta\max_j(T_{i,j})$ is the indicator function for the set $C_\zeta=\{P_i\in\mathbb{R}^{d_s}:\|\max(P_i,0)\|_1\leq\zeta\}$. Let $\tilde{T}^{k+1}=Z^{k+1}+(P^k/\delta)-(R_w\sigma C_w/\delta)$. Then, by the Moreau decomposition [29], the minimizer T^{k+1} is given by

$$T^{k+1}=\tilde{T}^{k+1}-\Pi_{C_{\frac{\zeta}{\delta}}}(\tilde{T}^{k+1})$$

where $\Pi_{C_{\zeta/\delta}}(\tilde{T}^{k+1})$ orthogonally projects each row of $\tilde{T}^{k+1}$ onto $C_{\zeta/\delta}$, which is something that can be computed with complexity $O(d_s\log(d_s))$.

The other algorithm parameters we use are $\delta=1$, $\zeta=1$, $\beta=250$, $\nu=50$, and $h=1-\cos(4\pi/180)$. In our experiments, we also choose $X_s=Y$. We then define column weights C_w that weight each column j by the number of pixels in the jth cluster (the cluster centered at Y_j) divided by the total number of pixels d. To refine the solution of the convex model, we note that each alternating step in the minimization of (7) is a convex minimization problem that can be again straightforwardly minimized using ADMM and its variants. The update for abundance S is identical to the split Bregman algorithm proposed for hyperspectral demixing in [20] and [21], and its connection to ADMM is discussed in [30].

III. Connection Between Row-0 and $l_{1,\infty}$ Norms

In this section, we will show that the motivation for our model comes from a close relation between the convex $l_{1,\infty}$ norm and the row-0 norm, which counts the number of nonzero rows.

A. Distinct Noise-Free Data

Let us assume that we have data X, which is completely noise free, normalized, obeys the linear mixing model, and contains a pure pixel for each material. Under slight abuse of notation, let us call this data after removing points that occur more than once in the image X again. As discussed in Section II, the endmember detection problem can be now reformulated as finding the minimizer T of (3), where the true set of endmembers can be then recovered as the columns of X that correspond to nonzero rows of T.

The fact that problem (3) gives a solution to the endmember selection problem is not surprising since (3) is a nonconvex problem and related problems (for instance, in compressed sensing) are usually hard to solve. However, due to the nonnegativity constraint, we will show that $l_{1,\infty}$ minimization is an exact relaxation of the above problem.

For any $T \geq 0$ with $XT = X$, the entries of T are less than or equal to one, i.e., $T_{i,j} \leq 1$.[3] Furthermore, the endmembers can be only represented by themselves, which means that in each row with index i, $i \in I$, we have a coefficient equal to 1. We can conclude that the $l_{1,\infty}$ norm of any $T \geq 0$ with $XT = X$ is

$$\|T\|_{1,\infty} = \sum_{i=1}^{d} \max_j T_{i,j} \tag{9}$$

$$\geq \sum_{i \in I} \max_j T_{i,j} = |I|. \tag{10}$$

However, it is possible to have equality in the above estimate if and only if $T_{i,j} = 0$ for $i \notin I$. In this case, the rows of the nonnegative $l_{1,\infty}$ minimizer of $XT = X$ are only supported on I, which means that it is a minimizer to the row-0 problem (3). Vice versa, any row-0 minimizer $\hat{T}$ has exactly one entry equal to one in any row corresponding to an endmember and zero rows elsewhere, which means that $\|\hat{T}\|_{1,\infty} = |I|$, and hence, $\hat{T}$ also minimizes the $l_{1,\infty}$ norm under the $XT = X$ constraint. We therefore have shown the following lemma.

Lemma III.1: If we remove repeated columns of X and have normalized data, the sets of minimizers of

$$\min_{T \geq 0} \|T\|_{\text{row-0}} \text{ such that } XT = X \tag{11}$$

$$\min_{T \geq 0} \|T\|_{1,\infty} \text{ such that } XT = X \tag{12}$$

are the same.

Notice that while generally there are other regularizations, like for instance $l_{1,2}$, which would also encourage row sparsity, this lemma only holds for $l_{1,\infty}$. For $XT = X$, the $l_{1,\infty}$ norm counts the number of rows and is not influenced by any value that corresponds to a mixed pixel. This property is unique for $l_{1,\infty}$ regularization with a nonnegativity constraint and therefore makes it the preferable choice in the proposed framework.

B. Noise in the Data

Of course the assumptions above are much too restrictive for real data. Therefore, let us look at the case of noisy data of the form $X^\delta = X + N$, where X is the true noise-free data with no repetition of endmembers as in the previous section and N is noise bounded in the Frobenius norm by $\|N\|_F \leq \delta$. We now consider the model

$$J_\alpha(T) = \|XT - X\|_F^2 + \alpha \|T\|_{1,\infty} \text{ such that } T \geq 0. \tag{13}$$

The following lemma shows that for the right noise-dependent choice of regularization parameter α, we converge to the correct solution as the noise decreases.

Lemma III.2: Let $\hat{T}$ be a nonnegative $\|\cdot\|_{1,\infty}$-minimum norm solution of $XT = X$, $X^\delta = X + N$ be noisy data with $\|N\|_F \leq \delta$, and T_α^δ denote a minimizer of energy functional (13) *with regularization parameter α and replacing X by X^δ. If α is chosen such that*

$$\alpha \to 0, \frac{\delta^2}{\alpha} \to 0 \quad \text{as} \quad \delta \to 0 \tag{14}$$

then there exists a convergent subsequence $T_{\alpha_n}^{\delta_n}$ and the limit of each convergent subsequence is a $\|\cdot\|_{1,\infty}$-minimum norm solution of $XT = X$.

For the sake of clarity, we moved the proof of this lemma to the Appendix. Lemma III.2 shows that our regularization is stable, since for decreasing noise and appropriate choice of parameter α, we converge to a nonnegative $\|\cdot\|_{1,\infty}$-minimum norm solution of $XT = X$, which, as we know from the first part, gives the true solution to the endmember detection problem *when the columns of X are distinct*. While identical points are easy to identify and eliminate in the noise-free data, determining which points belong to the same endmember in the noisy data can be very difficult. This is the reason for our first data reduction step. Lemma III.2 tells us that for our method to be closely related to the row-0 approach, we have to avoid having several noisy versions of the same endmember in X. We therefore impose an angle constraint, as described in Section II-B, while clustering the data, which basically corresponds to an upper bound on the noise and states up to which angle signals might correspond to the same point.

IV. Numerical Results for HSI

In this section, we present numerical results on endmember detection for supervised and real hyperspectral data and compare our results to existing detection algorithms. Since the goal of this paper is to present a general new convex framework for matrix factorization that is applicable to multiple areas, the hyperspectral numerical results are intended mainly as a proof of concept. It is therefore encouraging that our method is competitive with some of the established methods it is compared to in the examples below.

A. Application to Blind Hyperspectral Unmixing

1) Supervised Endmember Detection: For comparison purposes, we extracted nine endmembers from the standard Indian pines data set[4] by averaging over the corresponding signals in the ground truth region. Then, we created 50 data points for each endmember, 30 data points for each combination of two different endmembers, 10 data points for each combination of three different endmembers, and additionally 30 data points as mixtures of all endmembers. Finally, we add Gaussian noise with zero mean and standard deviation 0.006, make sure our data is positive, and normalize it. We evaluate our method in a comparison to N-findr [2], vertex component analysis (VCA) [14], [31] with code from [32], an NMF method using the alternating minimization scheme of our refinement step with random initial conditions, and the QR algorithm. For the latter, we simply used MATLAB's QR algorithm to calculate a permutation matrix Π such that $X\Pi = QR$ with decreasing diagonal values in R and chose the first nine columns of $X\Pi$ as endmembers. Many other methods could be also applied here, including the additional variational, Bayesian, and greedy dictionary selection methods referenced in the introduction. However, comparisons to the simple well-established N-findr, VCA, QR, and alternating minimization algorithms suffice to demonstrate that our proposed method works. Since the success of nonconvex

[3]This is a simple fact based on the normalization and nonnegativity, $1 = \|X_k\| = \|\sum_i T_{i,k} X_i\| \overset{T,X \geq 0}{\geq} \max_i \|T_{i,k} X_i\| = \max_i T_{i,k}$

[4]Available at https://engineering.purdue.edu/ biehl/MultiSpec/hyperspectral.html.

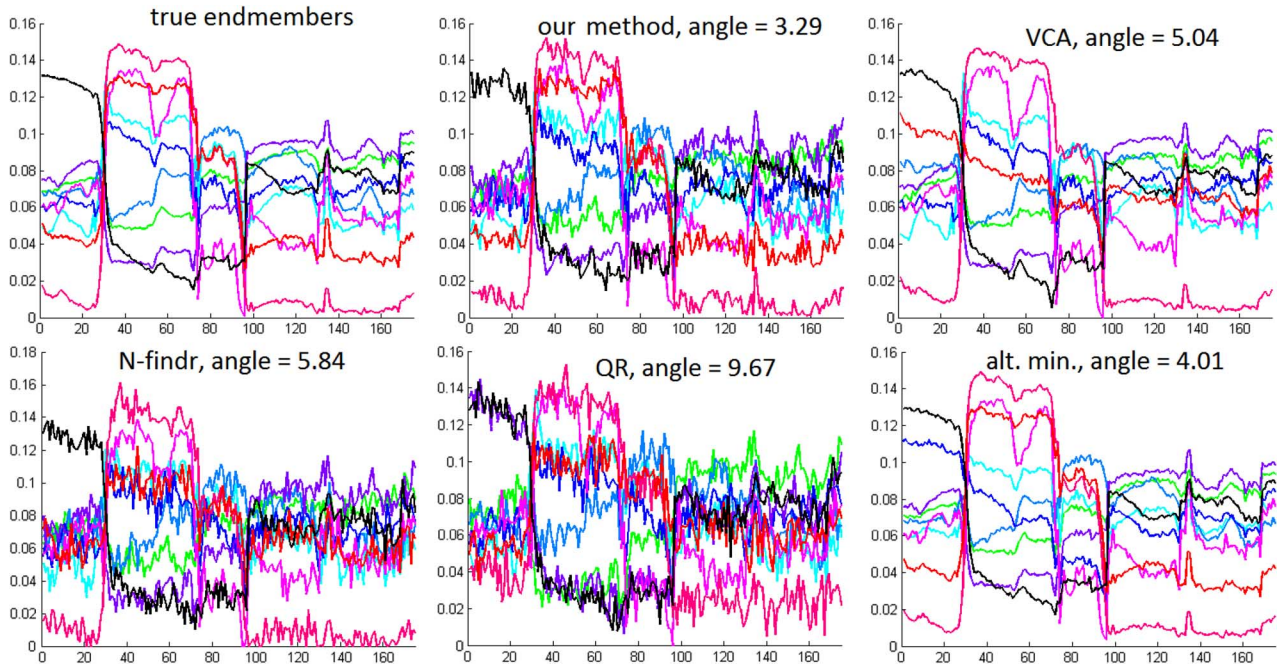

Fig. 2. Comparison of endmember reconstruction methods.

TABLE I
DEVIATION ANGLE FROM TRUE ENDMEMBERS

Method	Evaluation on 15 test runs		
	Avg. α	Min. α	Max. α
Ours refined	**3.37**	3.30	**3.42**
Ours without refinement	3.93	3.84	4.01
VCA	4.76	1.78	6.95
N-findr	10.19	7.12	13.79
QR	9.87	4.71	12.74
Alt. Min.	4.50	**1.76**	8.17

methods depends on the initial conditions or on random projections, we run 15 tests with the same general settings and record the average, maximum, and minimum angles by which the reconstructed endmember vectors deviate from the true ones (see Table I). Our method, unlike the others in Table I, does not require the number of endmembers to be specified in advance. For the tests, we adjusted the parameters of our method to obtain nine endmembers. Fig. 2 shows the original endmembers and example reconstruction by each method with the corresponding angle of deviation.

We can see that our method gives the best average performance. Due to a high noise level, methods that rely on finding the cone with maximal volume or on finding most linearly independent vectors will select outliers as the endmembers and do not yield robust results. Looking at the minimal and maximal α, we see the effect predicted. The nonconvex methods such as alternating minimization and VCA can outperform our method on some examples given angles as low as 1.76°. However, due to nonconvexity, they can sometimes find results that are far off the true solution with deviation angles of 6.95° or even 8.17°. A big benefit of our convex framework is that we consistently produce good results. The difference between the best and worst reconstruction angle deviations for our method is 0.15° with and 0.17° without refinement, which underlines its high stability.

The computational complexity of our method is highly dependent on the amount of data reduction. There is a tradeoff here since more data reduction speeds up the computation but limits the number of endmember candidates. For instance, in the above experiment, our algorithm (including the clustering with 10 k-means iterations and the refinement step) takes about 100 s. This runtime comes from the clustering selecting 250 endmember candidates due to a rather large noise level, which can make many similar signals look much more distinct than they really are. Reducing the noise level by a factor of 1/2, the clustering only selects between 90 and 100 endmember candidates, leading to a much faster runtime of about 13 s. In comparison, the alternating minimization approach needs about 45 s in the case of strong noise and about 29 s in the case of moderate noise. In our approach, the upper bound on the number of columns of Y can be adjusted according to how much of an issue computational resources are.

2) Results on Real Hyperspectral Data: To show how our method performs on real hyperspectral data, we use the urban image (publicly available at www.tec.army.mil/hypercube). Fig. 3 shows the RGB image of the urban scene, the spectral signatures of the endmembers our method extracted, and the abundance maps of each endmember, i.e., each row of T written back into an image.

The abundance maps are used in HSI to analyze the distribution of materials in the scene. First of all, our method managed to select six endmembers from the image, which is a very reasonable dimension reduction for hyperspectral image analysis. We can see that the abundance maps basically segment the image into clusters of different material categories such as concrete, house roofs, soil or dirt, grass, and two different types of vegetation, which all seem to be reasonable when visually comparing our results to the RGB image. The spectral signatures our method selected are relatively distinct but do not look noisy. Furthermore, the abundance maps seem to be sparse, which was a goal of our method and reflects the physically reasonable assumption that only very few materials should be present at each pixel.

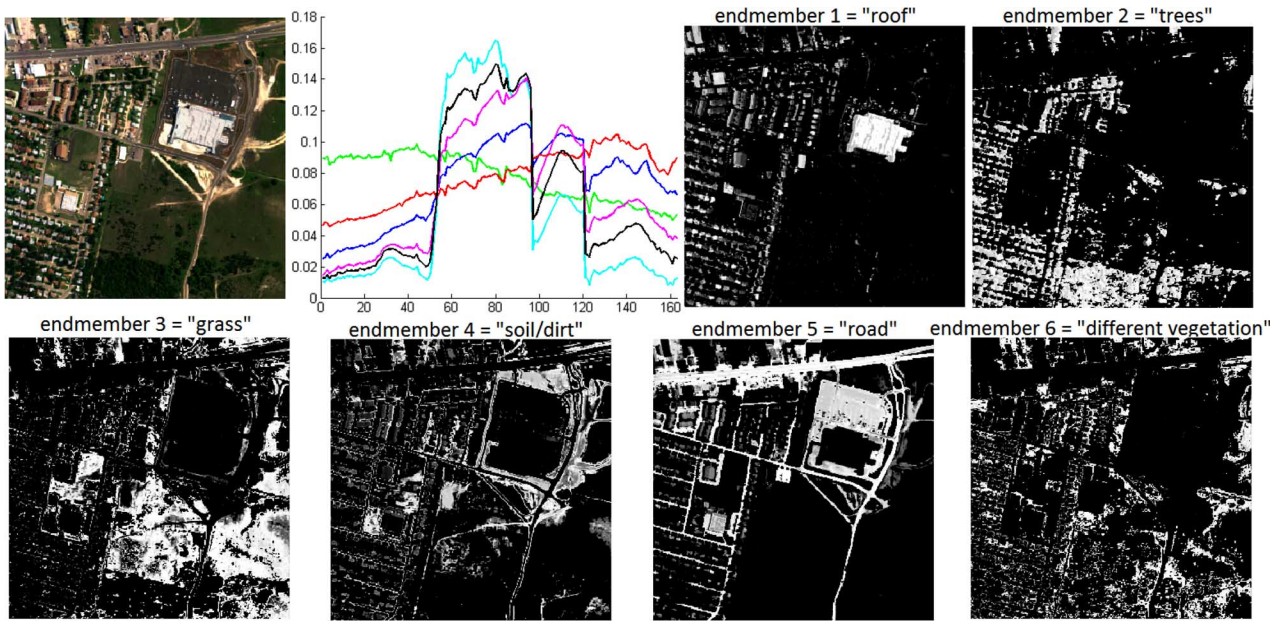

Fig. 3. Results on the urban hyperspectral image. (First row, left) RGB image to compare the fraction planes to. (First row, middle-left) Spectral signatures of the endmembers our method found. (First row, middle- right; first row, right; and second row) Abundance maps of the six endmembers.

TABLE II
COMPARISON OF DIFFERENT ENDMEMBER DETECTION METHODS IN TERMS OF ERROR AND SPARSITY OF THE DATA PROJECTION ONTO THE CONE SPANNED BY THE ENDMEMBERS ON THE URBAN IMAGE. THE CORRESPONDING ENDMEMBER SIGNATURES ARE SHOWN IN FIG. 4

	Ours	N-findr	VCA	QR	Alt. Min.
$\|A_m S_m - X\|^2$	533.9	4185.8	2516.5	1857.6	**454.3**
$\|S_m\|_0/(d \cdot n)$	**0.40**	0.60	0.47	0.60	0.41

As a next step, we compare our results to the ones obtained by N-findr, VCA, QR, and alternating minimization. Unfortunately, we have no ground truth for the urban image, which is why we will look at the nonnegative least squares (NNLS) unmixing results based on the endmembers A_m that each method found. Notice that geometrically, NNLS gives the projection of the data onto the cone spanned by the endmembers. If S_m denotes the NNLS solution of each method, the error $\|A_m S_m - X\|_F^2$ gives some insight on how much data is contained in the cone and, therefore, how well the endmembers describe the data. However, as discussed earlier, due to noise, we might not be interested in detecting the pixels furthest outside to be endmembers, although they might describe the data better. Thus, we will also report the sparsity of each cone projection S_m. Since any outside point will be projected onto a face or an edge of the cone, the sparsity will give some insight into whether the endmember vectors are well located. The more the endmembers are toward the inside of a point cluster, the more points we expect to be projected onto an edge of the cone rather than a face. Thus, a high sparsity relates to a reasonable position of an endmember. Table II shows the relative number of nonzero coefficients, i.e., $\|S_m\|_0$ divided by the total number of entries in S_m, as well as the projection error for N-findr, QR, VCA, alternating minimization, and our method. Fig. 4 shows the corresponding endmember signals that each method found.

We can see from the projection error that the sparse alternating minimization approach and our method found much better representation for the data than N-findr, VCA, and QR, with the alternating minimization performing slightly better than our approach. Furthermore, the sparsity of the projection is also higher, which indicates more reasonable choices for the endmembers. Looking at the spectral signatures in Fig. 4, we can speculate why this is the case. QR and N-findr chose very distinct signatures, which are probably far outliers in the data set. Some of the VCA endmembers look even more extreme and take negative values, which is clearly unphysical and does not allow these endmembers to be interpreted as any material. On the other hand, the endmembers of the alternating minimization approach and of our method are very similar and look comparably smooth. The average angle of deviation between our method and the alternating minimization approach is only 3.4°, which lets us conclude that they basically converged to the same answer, which is encouraging considering the fact that these endmembers describe the rest of the data more than three times more accurately than the endmembers found by other methods. Furthermore, any signal our method selected differs at most by 0.036° from an actual point in the data and is therefore physically meaningful.

V. EXTENDED MODEL

We also propose an extended version of the model that takes into account the normalization of the data in order to better distinguish between noise and outliers. We show this slightly more complicated functional can be still efficiently solved using convex optimization.

A. Error Model

We continue to work with the reduced set of endmember candidates Y and a subset of the data X_s, which in practice we take to be Y. Instead of penalizing $\|(YT - X_s)C_w\|_F^2$, we impose the linear constraint $YT - X_s = V - X_s \operatorname{diag}(e)$, where T, V, and e are the unknowns. T has the same interpretation as before, $V \in \mathbb{R}^{m \times d_s}$ models the noise, and $e \in \mathbb{R}^{d_s}$ models the sparse outliers. Since the columns of X_s are normalized to have unit l_2 norm, we would also like most of the columns of YT to be approximately normalized. In modeling the noise V, we therefore restrict it from having large components in the direction of X_s. To approximate this with a convex constraint $V \in D$, we

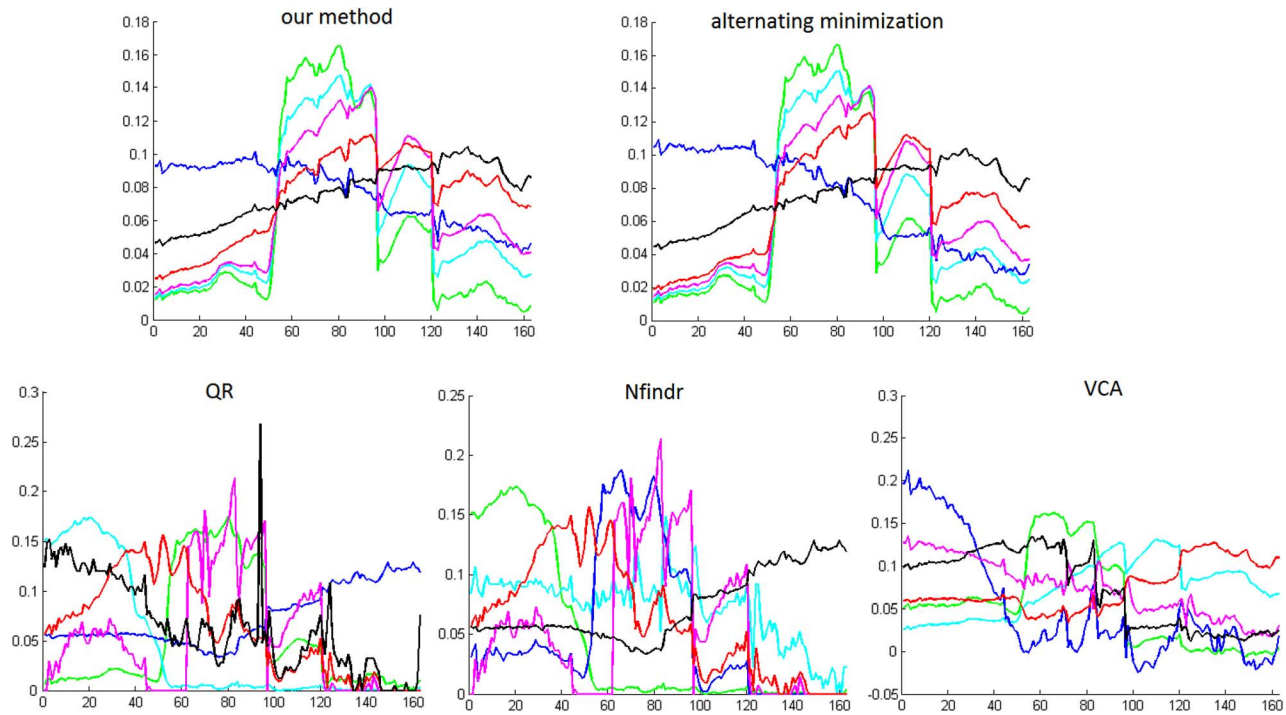

Fig. 4. Spectral signatures of endmembers extracted by different methods. (Top row) Results of our method and the alternating minimization approach. (Bottom row) Endmembers found by N-findr, QR, and VCA.

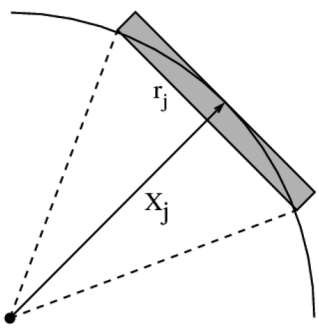

Fig. 5. Region of possible values for $X_j + V_j$.

restrict each column V_j to lie in a hockey-puck-shaped disk D_j. Decompose $V_j = V_j^{\perp} + V_j^{\|}$, where $V_j^{\|}$ is the orthogonal projection of V_j onto the line spanned by X_j and $V_j^{\perp}$ is the radial component of V_j perpendicular to $V_j^{\|}$. Then, given $0 \leq r_j < 1$, we restrict $\|V_j^{\perp}\|_2 \leq r_j$ and $\sqrt{1-r_j^2} - 1 \leq V_j^{\|} \leq 0$. The orthogonal projection onto this set is straightforward to compute since it is a box constraint in cylindrical coordinates. This constraint set for V_j is shown in Fig. 5 in the case when $e_j = 0$.

We also allow for a few columns of the data to be outliers. These are columns of X that we do not expect to be well represented as a small error plus a sparse nonnegative linear combination of other data but that we also do not want to consider as endmembers. Given some $\gamma \geq 0$, this sparse error is modeled as $-X_s \mathrm{diag}\, e$ with e restricted to the convex set $E = \{e : e \geq 0 \text{ and } \sum_j (C_w e)_j \leq \gamma\}$. Since E is the nonnegative region of a weighted l_1 ball, the orthogonal projection onto E can be computed with $O(d_s \log(d_s))$ complexity. Here, since the weights w_j sum to one by definition, γ can be roughly interpreted as the fraction of data we expect to be outliers. For nonoutlier data X_j, we want $e_j \approx 0$, and for outlier data, we want $e_j \approx 1$. In the latter outlier case, regularization on matrix T should encourage the corresponding column T_j to be close to zero; hence, $\|YT_j\|_2$ is encouraged to be small rather than close to one.

Keeping the $l_{1,\infty}$ regularization, the nonnegativity constraint, and the weighted l_1 penalty from (6), the overall extended model is given by

$$\min_{T \geq 0, V_j \in D_j, e \in E} \zeta \sum_i \max_j (T_{i,j}) + \langle R_w \sigma C_w, T \rangle \quad \text{such that} \quad YT - X_s = V - X_s \mathrm{diag}(e). \tag{15}$$

The structure of this model is similar to the robust principal component analysis model proposed in [33] although it has a different noise model and uses $l_{1,\infty}$ regularization instead of the nuclear norm.

B. Numerical Optimization

Since the convex functional for the extended model (15) is slightly more complicated, it is convenient to use a variant of ADMM that allows the functional to be split into more than two parts. The method proposed by He *et al.* in [34] is appropriate for this application. Again, introduce a new variable Z and constraint $Z = T$. In addition, let P_1 and P_2 be Lagrange multipliers for constraints $Z - T = 0$ and $YZ - V - X_s + X_s \mathrm{diag}(e) = 0$, respectively. Then, the augmented Lagrangian is given by

$$\begin{aligned} L_\delta(Z, T, V, e, P_1, P_2) &= g_{\geq 0}(T) + g_D(V) + g_E(e) \\ &\quad + \zeta \sum_i \max_j (T_{i,j}) + \langle R_w \sigma C_w, T \rangle + \langle P_1, Z - T \rangle \\ &\quad + \langle P_2, YZ - V - X_s + X_s \mathrm{diag}(e) \rangle + \frac{\delta}{2} \|Z - T\|_F^2 \\ &\quad + \frac{\delta}{2} \|YZ - V - X_s + X_s \mathrm{diag}(e)\|_F^2 \end{aligned}$$

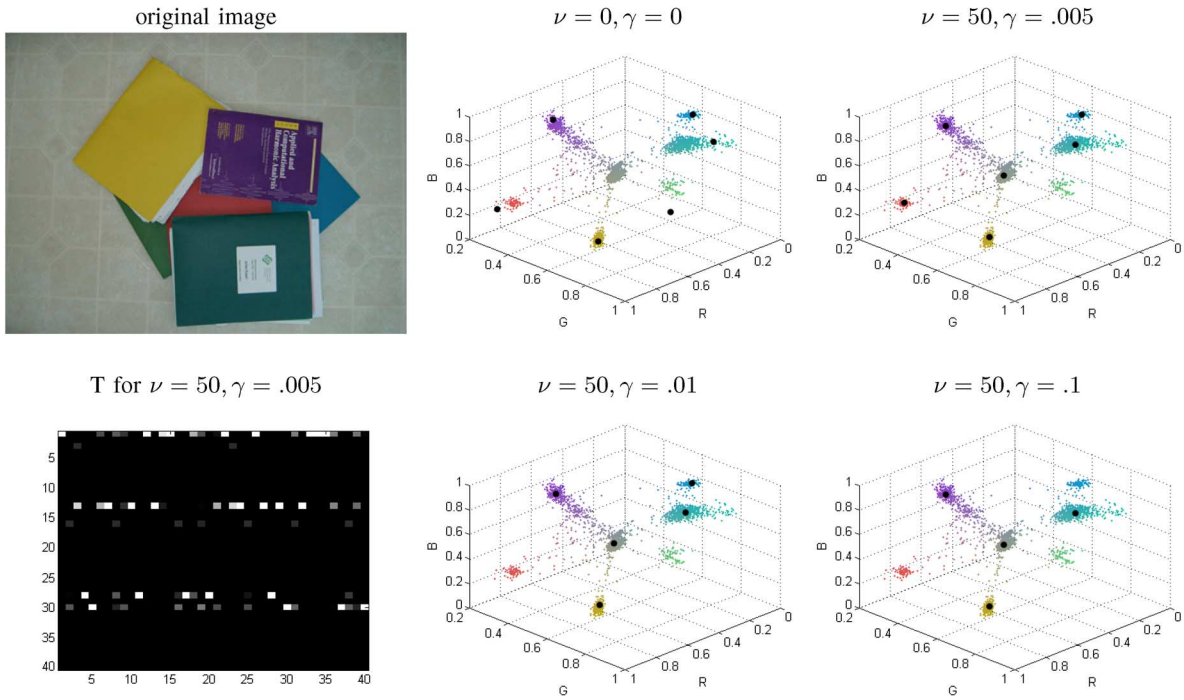

Fig. 6. Results of the extended model applied to the RGB image. (Top left) RGB image we apply the blind unmixing algorithm to. (Top middle) 3-D plot of the data points in the image in their corresponding color (in online version). (Black dots) Endmembers detected without allowing outliers ($\gamma = 0$) and without encouraging particular sparsity on the coefficients ($\nu = 0$). (Top right) With allowing some outliers the method removed an endmember in the one of the outside clusters, but included the middle cluster due to the encouraged sparsity. (Bottom left) Endmember coefficients for the parameter choice $\nu = 50, \gamma = 0.005$, where the brightness corresponds to the coefficient value. We can see that the coefficient matrix is sparse. (Bottom middle) Increasing the allowed outliers the red cluster endmember is removed (in online version). Increasing the outliers even further leads to decreasing the number of endmembers to four.

where g_D and g_E are indicator functions for the $V \in D$ and $e \in E$ constraints.

Using the ADMM-like method in [34], a saddle point of the augmented Lagrangian can be found by iteratively solving the subproblems with parameters $\delta > 0$ and $\mu > 2$, shown in the equations at the bottom of this page.

Each of these subproblems can be efficiently solved. There are closed formulas for the Z^{k+1} and V^{k+1} updates, and the e^{k+1} and T^{k+1} updates both involve orthogonal projections that can be efficiently computed.

C. Effect of Extended Model

A helpful example for visualizing the effect of the extended model (15) is to apply it to an RGB image. Although low dimensionality makes this significantly different from hyperspectral data, it is possible to view a scatter plot of the colors and how modifying the model parameters affects the selection of endmembers. The NMR data in Section V-E is 4-D; hence, low-dimensional data is not inherently unreasonable.

For the following RGB experiments, we use the same parameters as described in Section II-E and use the same k-means with farthest first initialization strategy to reduce the size of initial matrix Y. We do not however perform the alternating minimization refinement step. Due to the different algorithm used to solve the extended model, there is an additional numerical parameter μ, which for this application must be greater than two according to [34]. We set μ equal to 2.01. There are also model parameters r_j and γ for modeling the noise and outliers. To model the small-scale noise V, we set $r_j = \eta + m_j$, where η is fixed at .07 and m_j is the maximum distance from data in cluster j to the

$$Z^{k+1} = \arg\min_{Z} \left\| \begin{bmatrix} I \\ Y \end{bmatrix} Z - \begin{bmatrix} T^k \\ V^k - X_s \mathrm{diag}(e^k) + X_s \end{bmatrix} + \frac{1}{\delta} \begin{bmatrix} P_1^k \\ P_2^k \end{bmatrix} \right\|_F^2$$

$$T^{k+1} = \arg\min_{T} g_{\geq 0}(T) + \zeta \sum_i \max_j (T_{i,j}) + \frac{\delta\mu}{2} \left\| T - T^k - \frac{1}{\mu}(Z^k - T^k) - \frac{P_1}{\delta\mu} + \frac{R_w \sigma C_w}{\delta\mu} \right\|_F^2$$

$$V^{k+1} = \arg\min_{V} g_D(V) + \frac{\delta\mu}{2} \left\| V - V^k - \frac{1}{\mu}\left(YZ^{k+1} - V^k + X_s \mathrm{diag}(e^k) - X_s\right) - \frac{P_2^k}{\delta\mu} \right\|_F^2$$

$$e^{k+1} = \arg\min_{e} g_E(e) + \frac{\delta\mu}{2} \sum_{j=1}^{d_s} \left(e_j - e_j^k + \frac{1}{\delta\mu} \sum_{i=1}^{m} (X_s)_{i,j} \left(P_2^k + \delta\left(YZ^{k+1} - V^k + X_s \mathrm{diag}(e^k) - X_s\right)\right)_{i,j} \right)^2$$

$$P_1^{k+1} = P_1^k + \delta(Z^{k+1} - T^{k+1})$$

$$P_2^{k+1} = P_2^k + \delta\left(YZ^{k+1} - V^{k+1} - X_s + X_s \mathrm{diag}(e^{k+1})\right)$$

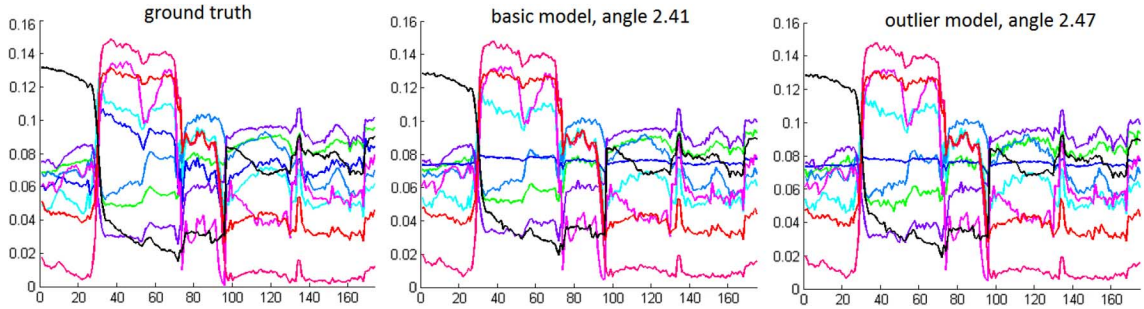

Fig. 7. Comparison between the basic and outlier methods on Indian pines data with Gaussian noise.

cluster center Y_j. To model the sparse error e, we use several different values of γ, which should roughly correspond to the fraction of data we are allowed to ignore as outliers. We will also use different values of ν, setting $\nu = 50$ to encourage a sparser abundance matrix and setting $\nu = 0$ to remove the weighted l_1 penalty from the model. Fig. 6 shows the data, which is a color image of multicolored folders; the selected endmembers (black dots) from four different experiments; and the sparse abundance matrix T for one of the experiments.

Note that in the $\nu = 0$ experiment, sparsity of T is not encouraged; thus, the dense gray cluster in the center of the cone, which corresponds to the floor in the image, is not selected as an endmember. Additionally, the selected endmembers are toward the very outside of the cone of data. In all the other experiments where $\nu = 50$, the floor is selected as an endmember although it is in the center of the cone of data. This results in a much sparser matrix T. Moreover, the selected endmembers tend to be in the center of the clusters of colors that we see in the scatter plot of the folder data. Finally, we note that, as we increase parameter γ, fewer endmembers are selected and some of the smaller outlying color clusters are ignored.

D. Comparison Between the Base and Outlier Models

To illustrate the difference between the outlier and base models, we use the same data set as for the supervised endmember detection experiment and first repeat the experiment from Section IV-A1 with 30 data points for each endmember, 20 data points for each combination of two different endmembers, 10 data points for each combination of three different endmembers, and additionally 30 data points as mixtures of all endmembers. We add Gaussian noise with zero mean and standard deviation 0.005; run the outlier model with $\zeta = 1$, $\eta = 0.08$, $\gamma = 0.01$, and $\nu = 40$; and run the basic model with $\zeta = 1.3$, $\nu = 40$, and $\beta = 250$.Fig. 7 shows the results for both methods with their average angle of deviation from the true endmembers.

We can see that both models give good results close to the ground truth. Due to the Gaussian noise we added, the best possible choice of columns of X deviates by an average angle of 3.29° from the true endmembers. With the refinement step, both methods could achieve an average angle below this value. The main remaining deviation is mainly due to the methods selecting an almost straight line rather than the corresponding true endmember. The similar results show that the extended model can be made to perform like the basic one. The extended model has the advantage, though, of being able to ignore a specified fraction of outliers.

As a second experiment, we simulate outliers in the data. First, we again create a mixed pixel data set with nine endmembers as above but without adding Gaussian noise. Next, we create a spike signal and add the spike itself, as well as about 3% data that is mixed with the spike signal. This shall simulate a small fraction of the data being outliers. Again, we run the basic and outlier models on this data set and obtain the results shown in Fig. 8. The upper left image shows the true nine endmembers plus the spike signal we used to create the outliers with. As we can see in the image on the upper right, which shows the result of the basic model, the algorithm selected the spike as an endmember. This is very reasonable because although only 3% of the data contains parts of the spike, it is a strong outlier and hence expensive for the fidelity term in the Frobenius norm to exclude. The shown 10 detected endmembers deviate only by 3.6° from the nine true endmembers and the spike.

The second row of Fig. 8 shows the nine true endmembers on the left and the result of the outlier model on the right. As we can see, the outlier model was able to handle the small fraction of signals containing the spike and only selected the nine true endmembers. The average angle of deviation in this case is 2.7°, and we can confirm that the model behaves like we expected it to.

E. Application to BSS of NMR Data

To illustrate how our model can be applied to BSS problems, we use it to recover the four NMR source spectra from [1, Fig. 4] from four noise-free mixtures. The four sources are shown in Fig. 9. Let $S_0 \in \mathbb{R}^{4\times 5000}$ be the sources, and let the mixtures X_0 be generated by $X_0 = A_0 S_0$ with

$$A_0 = \begin{bmatrix} 0.3162 & 0.6576 & 0.3288 & 0.5000 \\ 0.3162 & 0.3288 & 0.6576 & 0.5000 \\ 0.6325 & 0.1644 & 0.1644 & 0.5000 \\ 0.6325 & 0.6576 & 0.6576 & 0.5000 \end{bmatrix}.$$

We will use the outlier model to recover the mixing matrix A from X_0. Unlike the hyperspectral examples, some columns of X_0 here can be nearly zero if all sources are simultaneously zero at the same spectral index. We can see in Fig. 9 that this is indeed the case. Since our algorithm uses normalized data, we first remove columns of X_0 whose norm is below some threshold, which we take to be $0.01 \max_j \|X_j\|$. We then normalize the remaining columns to get X. This simple approach suffices for this example, but in general, the parameters r_j for the $V \in D$ constraint could be also modified to account for columns of X_0 that have significantly different norms.

A minor difficulty in applying our method to this BSS problem is that we know that A should have four columns

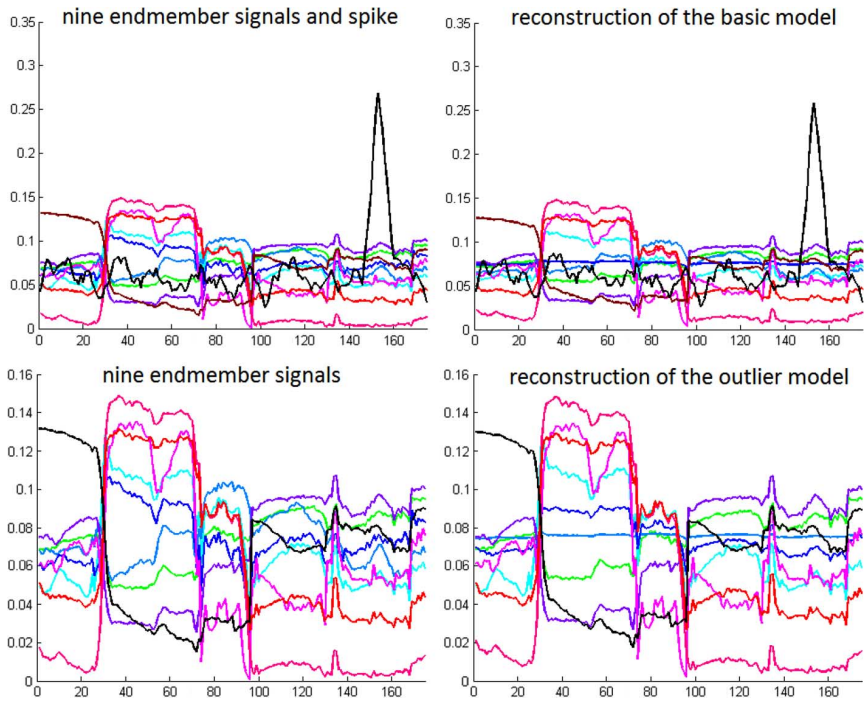

Fig. 8. Comparison between the basic and outlier methods on Indian pines data with outliers.

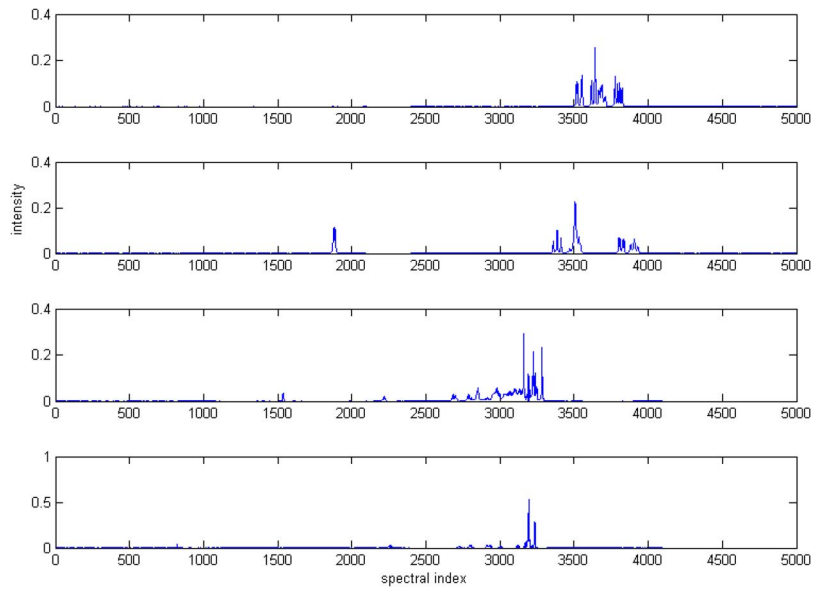

Fig. 9. Four NMR source spectra from [1, Fig. 4].

but there is no way to constrain the algorithm to produce a dictionary with exactly four elements. We therefore adjust parameters until the dimension of the result is correct. This is straightforward to do and could be automated. For example, to choose a smaller dictionary, we can reduce ν and/or increase γ.

The parameters used here are identical to those used in the RGB experiments of Section V-C except $\gamma = 0.01$ and $\nu = 5$. In addition, for the data reduction step, the angle constraint was increased to $\langle Y_i, Y_j \rangle < 0.998$. The computed mixing matrix after permutation is

$$A = \begin{bmatrix} 0.3267 & 0.6524 & 0.3327 & 0.4933 \\ 0.3180 & 0.3300 & 0.6544 & 0.5110 \\ 0.6228 & 0.1757 & 0.1658 & 0.4836 \\ 0.6358 & 0.6593 & 0.6585 & 0.5114 \end{bmatrix}.$$

Note that the columns of A are normalized because the algorithm selects dictionary elements from a normalized version of the data X. Since for this simple problem, A is invertible, it is straightforward to recover the sources by $S = \max(0, A^{-1}X_0)$. More generally, we can recover S by minimizing the convex functional in (7) with respect to S using the unnormalized data matrix X_0 and the computed endmembers A.

VI. Future Research

We have presented a convex method for factoring a data matrix X into a product AS with $S \geq 0$ under the constraint that the columns of A appear somewhere in the data X. This type of factorization ensures the physical meaning of the dictionary A, and we have successfully applied it to hyperspectral endmember detection and BSS in NMR. For nonrepeating noise-free data,

the $l_{1,\infty}$ regularization was proven to be an exact relaxation of the row-0 norm. We further proposed an extended model that can better handle outliers. Possible future application areas include computational biology, sensor networks, and in general, dimensionality reduction and compact representation applications where the physical interpretation of the reduced space is critical. It will be also interesting to try and extend our convex model to the class of problems discussed in [35] for which the pixel purity or nonoverlapping assumption is approximately but not exactly satisfied.

APPENDIX
PROOF OF LEMMA III.2

Since T_α^δ is a minimizer of J_α, we can conclude

$$\begin{aligned}\left\|T_\alpha^\delta\right\|_{1,\infty} &\le \frac{1}{\alpha} J_\alpha\left(T_\alpha^\delta\right)\\ &\le \frac{1}{\alpha} J_\alpha(\hat{T})\\ &= \|\hat{T}\|_{1,\infty} + \frac{1}{\alpha}\left\|(X+N)\hat{T} - (X+N)\right\|_F^2\\ &= \|\hat{T}\|_{1,\infty} + \frac{1}{\alpha}\left\|\underbrace{(X\hat{T}-X)}_{=0} + (N\hat{T}-N)\right\|_F^2\\ &\le \|\hat{T}\|_{1,\infty} + \frac{1}{\alpha}\|N\|_F^2\|\hat{T}-\mathrm{Id}\|_F^2\\ &= \|\hat{T}\|_{1,\infty} + \frac{\delta^2}{\alpha}\|\hat{T}-\mathrm{Id}\|_F^2. \end{aligned} \tag{16}$$

Since $\delta^2/\alpha \to 0$, T_α^δ is bounded in the $\|\cdot\|_{1,\infty}$ norm and therefore has a convergent subsequence. Let $T_{\alpha_n}^{\delta_n}$ denote such a convergent subsequence, and let $\bar{T}$ be its limit. Because $T_{\alpha_n}^{\delta_n} \ge 0$, we also have $\bar{T} \ge 0$. We can now use the above estimate for showing

$$\begin{aligned}\|\bar{T}\|_{1,\infty} &= \lim_n \left\|T_{\alpha_n}^{\delta_n}\right\|_{1,\infty}\\ &\le \lim_n \left[\|\hat{T}\|_{1,\infty} + \frac{\delta_n^2}{\alpha_n}\|\hat{T}-\mathrm{Id}\|_F^2\right]\\ &= \|\hat{T}\|_{1,\infty}. \end{aligned} \tag{17}$$

Furthermore, we have

$$\begin{aligned}\|X\bar{T} - X\|_F^2 &= \lim_n \left\|X^{\delta_n}T_{\alpha_n}^{\delta_n} - X^{\delta_n}\right\|_F^2\\ &\le \lim_n J_{\alpha_n}\left(T_{\alpha_n}^{\delta_n}\right)\\ &\le \lim_n J_{\alpha_n}(\hat{T})\\ &\le \lim_n \|X^{\delta_n}\hat{T} - X^{\delta_n}\|_F^2 + \alpha_n\|\hat{T}\|_{1,\infty}\\ &\le \lim_n \delta_n^2\|\hat{T}-\mathrm{Id}\|_F^2 + \alpha_n\|\hat{T}\|_{1,\infty} = 0. \end{aligned} \tag{18}$$

Now, by the above estimate (18), we know that $X\bar{T} = X$. Furthermore, by the estimate (17) and taking into account that $\hat{T}$ was a nonnegative $\|\cdot\|_{1,\infty}$-minimum norm solution of $XT = X$, we have shown that the limit of our convergent subsequence is also a nonnegative $\|\cdot\|_{1,\infty}$-minimum norm solution of $XT = X$. ■

ACKNOWLEDGMENT

The authors would like to thank L. Balzano and R. Nowak who are working on a similar selection framework without nonnegativity constraints, partially motivated by problems in sensor networks, and had provided important comments and feedback. The authors would also like to thank N. Gillis, J. Greer, and T. Wittman for helpful discussions about endmember detection strategies and Y. Sun for his advice on blind source separation problems.

REFERENCES

[1] W. Naanaa and J.-M. Nuzillard, "Blind source separation of positive and partially correlated data," *Signal Process.*, vol. 85, no. 9, pp. 1711–1722, Sep. 2005.
[2] M. E. Winter, "N-FINDR: An algorithm for fast autonomous spectral end-member determination in hyperspectral data," in *Proc. SPIE—Imaging Spectrometry V*, 1999, vol. 3753, pp. 266–275.
[3] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," INRIA, Le Chesnay, France, Tech. Rep. arXiv:0904.3523v2, 2009.
[4] J. A. Tropp, "Algorithms for simultaneous sparse approximation: Part II: Convex relaxation," *Signal Process.*, vol. 86, no. 3, pp. 589–602, Mar. 2006.
[5] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. ICML*, 2010, pp. 663–670.
[6] W. Xu, X. Liu, and Y. Gong, "Document clustering based on nonnegative matrix factorization," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval*, 2003, pp. 267–273.
[7] A. Holzapfel and Y. Stylianou, "Musical genre classification using nonnegative matrix factorization-based features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 424–434, Feb. 2008.
[8] 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.
[9] M. Berry, M. Browne, A. Langville, P. Pauca, and R. J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," *Comput. Stat. Data Anal.*, vol. 52, no. 1, pp. 155–173, 2007.
[10] V. P. Pauca, J. Piper, and R. J. Plemmons, "Nonnegative matrix factorization for spectral data analysis," *Linear Algebra Appl.*, vol. 416, no. 1, pp. 29–47, Jul. 2006.
[11] A. Zare, "Hyperspectral endmember detection and band selection using Bayesian methods" Ph.D. dissertation, Univ. Florida, Gainesville, FL, 2008.
[12] S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret, "Separation of non-negative mixture of non-negative sources using a Bayesian approach and MCMC sampling," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4133–4145, Nov. 2006.
[13] N. Dobigeon, S. Moussaoui, M. Coulon, J.-Y. Tourneret, and A. O. Hero, "Joint Bayesian endmember extraction and linear unmixing for hyperspectral imagery," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4355–4368, Nov. 2009.
[14] J. M. P. Nascimento and J. M. Bioucas-Dias, "Blind hyperspectral unmixing," in *Proc. SPIE Conf. Image Signal Process. Remote Sens. XIII*, 2007, vol. 6748, p. 67 480J.
[15] C. Boutsidis, M. W. Mahoney, and P. Drineas, "An improved approximation algorithm for the column subset selection problem," in *Proc. 20th Annu. ACM-SIAM SODA*, 2009, pp. 968–977.
[16] T. F. Chan and P. C. Hansen, "Some applications of the rank revealing QR factorization," *SIAM J. Sci. Stat. Comput.*, vol. 13, no. 3, pp. 727–741, May 1992.
[17] A. Krause and V. Cevher, "Submodular dictionary selection for sparse representation," in *Proc. ICML*, 2010, pp. 567–574.
[18] A. Das and D. Kempe, "Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection," in *Proc. ICML*, 2011, pp. 1057–1064.
[19] L. Balzano, R. Nowak, and W. U. Bajwa, "Column subset selection with missing data," in *Proc. NIPS Workshop Low-Rank Methods Large-Scale Mach. Learn.*, Whistler, BC, Canada, 2010.
[20] A. Szlam, Z. Guo, and S. Osher, A split Bregman method for nonnegative sparsity penalized least squares with applications to hyperspectral demixing UCLA CAM, Los Angeles, CA, Tech. Rep. 10-06, 2010.
[21] Z. Guo, T. Wittman, and S. Osher, "L1 unmixing and its application to hyperspectral image enhancement," in *Proc. SPIE Conf. Algorithms Technol. Multispectral, Hyperspectral, Ultraspectral Imag. XV*, 2008, vol. 7334, pp. 73 341M-1–73 341M-9.

[22] J. Greer, "Sparse demixing," in *Proc. SPIE Algorithms Technol. Multispectral, Hyperspectral, Ultraspectral Imag. XVI*, 2010, vol. 7695, pp. 76 951O-1–76 951O-12.
[23] A. Castrodad, Z. Xing, J. Greer, E. Bosch, L. Carin, and G. Sapiro, "Learning discriminative sparse representations for modeling, source separation, and mapping of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sensing*, vol. 49, no. 11, pp. 4263–4281, Nov. 2011.
[24] P. Sprechmann, I. Ramrez, G. Sapiro, and Y. Eldar, "C-HiLasso: A collaborative hierarchical sparse modeling framework," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4183–4198, Sep. 2011.
[25] J. Mairal, R. Jenatton, G. Obozinski, and F. Bach, "Network flow algorithms for structured sparsity," in *Proc. NIPS*, 2010, pp. 1558–1566.
[26] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite-element approximations," *Comput. Math. Appl.*, vol. 2, no. 1, pp. 17–40, 1976.
[27] R. Glowinski and A. Marrocco, "Sur lapproximation par elements finis dordre un, et la resolution par penalisation-dualite dune classe de problemes de Dirichlet nonlineaires," *Rev. Francaise dAut Inf. Rech. Oper.*, vol. R-2, pp. 41–76, 1975.
[28] X. Zhang, M. Burger, and S. Osher, "A unified primal–dual algorithm framework based on Bregman iteration," UCLA CAM, Los Angeles, CA, Tech. Rep. 09-99, 2009.
[29] J. J. Moreau, "Proximité et dualité dans un espace hilbertien," *Bull. Soc. Math. France*, vol. 93, pp. 273–299, 1965.
[30] E. Esser, "Applications of Lagrangian-based alternating direction methods and connections to split Bregman," UCLA CAM, Los Angeles, CA, Tech. Rep. 09-31, 2009.
[31] J. M. P. Nascimento and J. M. Bioucas-Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 4, pp. 898–910, Apr. 2005.
[32] Open Source MATLAB Hyperspectral Toolbox. ver. 0.04, 2010 [Online]. Available: http://matlabhyperspec.sourceforge.net/
[33] E. Candes, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis," 2009 [Online]. Available: http://arxiv.org/PS cache/arxiv/pdf/0912/0912.3599v1.pdf
[34] B. He, M. Tao, and X. Yuan, "A splitting method for separate convex programming with linking linear constraints," Tech. Rep., 2011 [Online]. Available: http://www.optimization-online.org/DB_FILE/2010/06/ 2665.pdf
[35] Y. Sun, C. Ridge, F. del Rio, A. J. Shaka, and J. Xin, "Postprocessing and sparse blind source separation of positive and partially overlapped data," *Signal Process.*, vol. 91, no. 8, pp. 1838–1851, Aug. 2011.

Ernie Esser received the B.S. degrees in mathematics and applied and computational mathematical sciences from the University of Washington, Seattle, in 2003 and the Ph.D. degree in math from the University of California at Los Angeles in 2010.

He is currently a Postdoctoral Researcher at the University of California, Irvine. His main research interests are convex optimization, image processing, and inverse problems.

Michael Möller received the Diploma in mathematics from the Westfälische Wilhelms University of Münster, Münster, Germany, in 2009, where he currently working toward the Ph.D. degree.

He spent two years as a Visiting Researcher at the University of California at Los Angeles. His research interests include convex optimization, compressed sensing, image processing, and inverse problems.

Stanley Osher received the B.S. degree in mathematics from Brooklyn College, Brooklyn, NY, in 1962, and the M.S. and Ph.D. degrees in mathematics from New York University, New York, in 1964 and 1966, respectively.

Since 1977, he has been with the University of California, Los Angeles (UCLA), where he is currently a Professor in the Department of Mathematics and the Director of Special Projects at the Institute for Pure and Applied Mathematics. His current interests mainly involve information science, which includes image processing, compressed sensing, and machine learning. His website is www.math.ucla.edu/~sjo. His recent papers are on the website for the UCLA Computational and Applied Mathematics Reports at http://www.math.ucla.edu/applied/cam/index.shtml.

Dr. Osher is a member of the National Academy of Sciences and the American Academy of Arts and Sciences, and he is one of the top 25 most highly cited researchers in both mathematics and computer science. He has received numerous academic honors and has cofounded three successful companies, each based largely on his own (joint) research.

Guillermo Sapiro (M'95–SM'03) was born in Montevideo, Uruguay, on April 3, 1966. He received the B.Sc. (*summa cum laude*), M.Sc., and Ph.D. degrees from the Technion-Israel Institute of Technology, Haifa, Israel, in 1989, 1991, and 1993, respectively.

After postdoctoral research at the Massachusetts Institute of Technology, Cambridge, he became a member of the Technical Staff at the research facilities of HP Laboratories, Palo Alto, CA. He is currently with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, where he holds the position of Distinguished McKnight University Professor and Vincentine Hermes-Luh Chair in Electrical and Computer Engineering. He works on differential geometry and geometric partial differential equations both in theory and applications in computer vision, computer graphics, medical imaging, and image analysis. He has recently co-edited a special issue of IEEE TRANSACTIONS ON IMAGE PROCESSING in this topic and a second one in the *Journal of Visual Communication and Image Representation*. He is the founding Editor-in-Chief of the SIAM *Journal on Imaging Sciences*. He has authored and coauthored numerous papers in this area and has written a book published by Cambridge University Press in January 2001.

Dr. Sapiro is a member of the Society of Industrial and Applied Mathematics (SIAM). He was the recipient of the Gutwirth Scholarship for Special Excellence in Graduate Studies in 1991, the Ollendorff Fellowship for Excellence in Vision and Image Understanding Work in 1992, the Rothschild Fellowship for Postdoctoral Studies in 1993, the Office of Naval Research Young Investigator Award in 1998, the Presidential Early Career Awards for Scientist and Engineers (PECASE) in 1998, the National Science Foundation Career Award in 1999, and the National Security Science and Engineering Faculty Fellowship in 2010.

Jack Xin received the B.S. degree in computational mathematics from Peking University, Beijing, China, in 1985 and the M.S. and Ph.D. degrees in applied mathematics from New York University, New York, in 1988 and 1990, respectively.

He was a Postdoctoral Fellow at the University of California at Berkeley and Princeton University, Princeton, NJ, in 1991 and 1992, respectively. He was an Assistant Professor and Associate Professor of mathematics at the University of Arizona, Tucson, from 1991 to 1999. He was a Professor of mathematics from 1999 to 2005 at the University of Texas, Austin. Since 2005, he has been a Professor of mathematics in the Department of Mathematics, Center for Hearing Research, Institute for Mathematical Behavioral Sciences, and Center for Mathematical and Computational Biology at University of California, Irvine. His research interests include applied analysis, computational methods and their applications in nonlinear and multiscale dynamics in fluids, and signal processing.

Dr. Xin is a Fellow of the John S. Guggenheim Foundation.