A CONVEX MODEL FOR MATRIX FACTORIZATION AND DIMENSIONALITY REDUCTION ON PHYSICAL SPACE AND ITS APPLICATION TO BLIND HYPERSPECTRAL UNMIXING

Michael Möller ⋆ *Ernie Esser*‡ *Stanley Osher*⋆ *Guillermo Sapiro*† *Jack Xin*‡

‡ Department of Mathematics, University of California Irvine
† Department of Electrical and Computer Engineering, University of Minnesota
⋆ Department of Mathematics, University of California Los Angeles

ABSTRACT

A collaborative convex framework for factoring a data matrix X into a non-negative product AS, with a sparse coefficient matrix S, is introduced. We restrict the columns of the dictionary matrix A to coincide with certain columns of X, thereby guaranteeing a physically meaningful dictionary and dimensionality reduction. As an example, we show applications of the proposed framework on hyperspectral endmember and abundances identification.

***Index Terms*—** Hyperspectral endmember detection, non-negative matrix factorization, dictionary learning, subset selection, dimensionality reduction

1. INTRODUCTION

Dimensionality reduction has been widely studied in the signal processing and computational learning communities in recent years. 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 work we present a framework for dimensionality reduction, based on matrix factorization and sparsity theory, that uses the data itself for the low dimensional representation, thereby guaranteeing the physical fidelity. While the method is applicable in numerous areas, from biology to sensor networks, we concentrate on an application in hyperspectral imaging (HSI). This by itself is an important area that will help to illustrate the proposed framework and demonstrate it in real data. Applications in other disciplines, where the physical meaning of the low dimensional representation is also critical, will be studied in the future.

HSI sensors record up to several hundred different frequencies in the visible, near-infrared and infrared spectrum. 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 *abundances* (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*) in the image. 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, see [1] and references therein. Related although not yet applied to endmember detection are subset selection methods like the rank-revealing QR decomposition (e.g. [2]), which finds the most linearly independent columns of a matrix.[1]

Simultaneously detecting endmembers and computing abundances can be stated as factorizing 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. The abundance matrix S contains the amounts of each material in A at each pixel in X. Due to the physical non-negativity constraint the geometric interpretation of blind unmixing is to find endmembers that span a cone which contains most of the data.

The general problem of representing $X \approx AS$ with $A, S \geq 0$ is known as non-negative matrix factorization (NMF). The application of NMF to hyperspectral endmember detection can for instance be found in [3]. NMF problems are non-convex and typically solved by estimating A and S alternatingly. Although variants of NMF methods often produce good results in practice, they are not guaranteed to converge to a global minimum.

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 [4]. Motivated by the ideas of dictionary learning for sparse coding, the works [5, 6] proposed explicitly to look for endmember matrices that lead to sparse abundance maps. We follow a similar idea, though 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

MM WAS SUPPORTED BY THE GERMAN ACADEMIC EXCHANGE SERVICE (DAAD), EE AND JX BY NSF GRANTS DMS-0911277, PRISM-0948247, SO AND MM BY NSF GRANTS DMS-0835863, DMS-0914561, DMS-0914856 AND ONR GRANT N00014-08-1119, AND GS WAS SUPPORTED BY NSF, NGA, ONR, ARO, AND NSSEFF.

[1] Independent of the work here described, Laura Balzano and Rob Nowak developed a related matrix factorization technique and connected and compared it to RRQR. We thank Laura for pointing out their work and also the possible relationships with RRQR.

for moderate ground sampling distance images and is called *pixel purity* assumption. In the general context of dictionary learning and non-negative matrix factorization it guarantees the columns of A to be physically meaningful. As mentioned above, the lack of physical interpretation has been a critical shortcoming of standard dimensionality reduction and dictionary learning techniques, and 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.

Our method is based on the recent ideas of collaborative sparsity (see for example [7] and the references therein) and will use an energy that extends the one studied in [8], in a similar setting as recently proposed by Lin et al. in [9] for low rank approximation with the nuclear norm.

2. CONVEX ENDMEMBER DETECTION MODEL

With the restriction that the endmembers in A be part of the data X, the underlying problem is then to find a nonnegative matrix, T, such that $XT \approx X$, most rows of T are zero and the columns of T are additionally sparse. In other words, we want the columns of X, which are the spectral signatures at each pixel, to be well represented by sparse nonnegative linear combinations of the same sparse subset of columns of X. This subset of columns corresponds to the nonzero rows of T and will be the selected endmembers. 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. We normalize the columns of X to have unit l_2 norm so that we discriminate based solely on spectral signatures and not intensity.

It's already clear from this formulation 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 dimension of the problem by using clustering to choose a subset of candidate endmembers A from the columns of X and a submatrix $X_s \in \mathbb{R}^{m \times d_s}$ of X for the data with $d_s \leq d$. We use $X_s = A$ 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 A. An angle constraint $\langle A_i, A_j \rangle < .995$ ensures the endmember candidates are sufficiently distinct, and an upper bound n_c is placed on the number of allowable clusters so that T is at most a $n_c \times d_s$ matrix and 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 $AT \approx X_s$, with T having the same sparsity properties described above. Note that we have not convexified the problem by pre-fixing the dictionary A. This is done simply to work with manageable dimensions and datasets. Our convex model will still select the endmembers as a subset of this reduced A.

2.1. The 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 $\frac{\beta}{2}\|(AT - X_s)\sqrt{C_w}\|_F^2$, 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 we can use to weight the columns of $(AT - X_s)$ so that it reflects the density of the original data. So that only a few samples are cooperatively selected as endmembers, we encourage rows of T to be zero by penalizing $\zeta \sum_i \|T_i\|_\infty$ with ζ a positive constant. Due to the nonnegativity constraint, this is the same as $\zeta \sum_i \max_j(T_{i,j})$.[2] This kind of collaborative/structured sparsity regularizer has been proposed in several previous works, for example [8]. 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 we could also choose these weights to be proportional to $\langle A_j, \bar{A} \rangle$, where $\bar{A}$ is the average of the columns of A, which would encourage selection of endmembers towards the outside of the cone containing the data. The weighting matrix σ has the same dimension as T, and the weights are chosen to be $\sigma_{i,j} = \nu(1 - e^{\frac{-(1-(A^T X_s)_{i,j})^2}{2h^2}})$ for constants h and ν. This means that $\sigma_{i,j}$ is small when the ith column of A 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. 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

$$\min_{T \geq 0} \zeta \sum_i \max_j(T_{i,j}) + \langle R_w \sigma C_w, T \rangle + \frac{\beta}{2}\|(AT - X_s)\sqrt{C_w}\|_F^2. \tag{1}$$

2.2. Refinement of solution

Since we are using a convex model to detect endmembers, it can't distinguish between identical or very similar endmember candidates. But the model works very reliably when the columns of A 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 A. Nonetheless, as is shown in Section 3, the results of this approach already compare favorably to other methods. 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

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

and renormalizing the columns of A after each iteration. Here, r_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 close to the physical space.

2.3. Numerical optimization

We use the alternating direction method of multipliers (ADMM) [10, 11] to solve (1) using the parameters $n_c = 150$, $\zeta = 1$, $\beta = 250$, $\nu = 50$, and $h = 1 - \cos(4\pi/180)$. In our experiments, we also choose $X_s = A$. 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 A_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 (2) is a convex minimization problem that can again be minimized using ADMM and its variants. The update for the abundance S is identical to the split Bregman algorithm proposed for hyperspectral demixing in [4], and its connection to ADMM is discussed in [12].

[2] The work mentioned above by Balzano and Nowak uses $\|\cdot\|_2$ instead of $\|\cdot\|_\infty$.

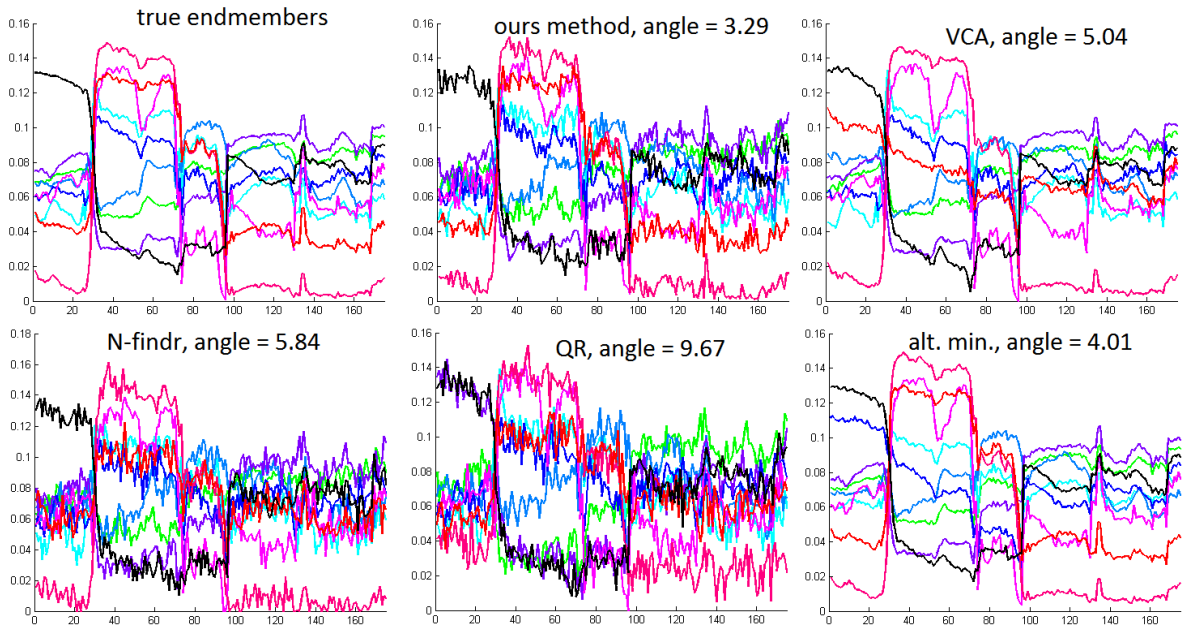

Fig. 1: Comparison of endmember recostruction methods

3. NUMERICAL RESULTS

3.1. Supervised endmember detection

For comparison purposes we extracted nine endmembers from the standard Indian pines dataset (publicly available at `https://engineering.purdue.edu/biehl/MultiSpec/hyperspectral.html`) 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, 10 data points for each combination of three different, 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 [13], vertex component analysis (VCA) [14] with code from [15], 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. Since the success of non-convex 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 angle by which the reconstructed endmember vectors deviate from the true ones in Table 1. For the tests we adjusted the parameters of our method to obtain 9 endmembers.

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 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 non-convex methods like alternating minimization and VCA can outperform our method on some examples giving angles as low as 1.76. However, due to the non-convexity they can sometimes find results which are far off the true solution with angles of 6.95 or even 8.17 degrees. A big benefit of our convex framework is that we consistently produce good results. The difference between the best and the worst reconstruction angle for our method is 0.15 degrees with and 0.17 without the refinement, which underlines its high stability. Figure 1 shows the original endmembers as well as an example reconstructions by each method with the corresponding angle of deviation.

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

Table 1: Angle of deviation from true endmembers

3.2. 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`). Figure 2 shows the RGB image of the urban scene, the spectral signatures of the endmembers our method extracted, and the abundance maps of each endmember. We can see that we get a very sparse image representation, segmenting into 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.

4. FUTURE RESEARCH

We have presented a convex method for endmember detection in hyperspectral images. However, the general framework can be applied

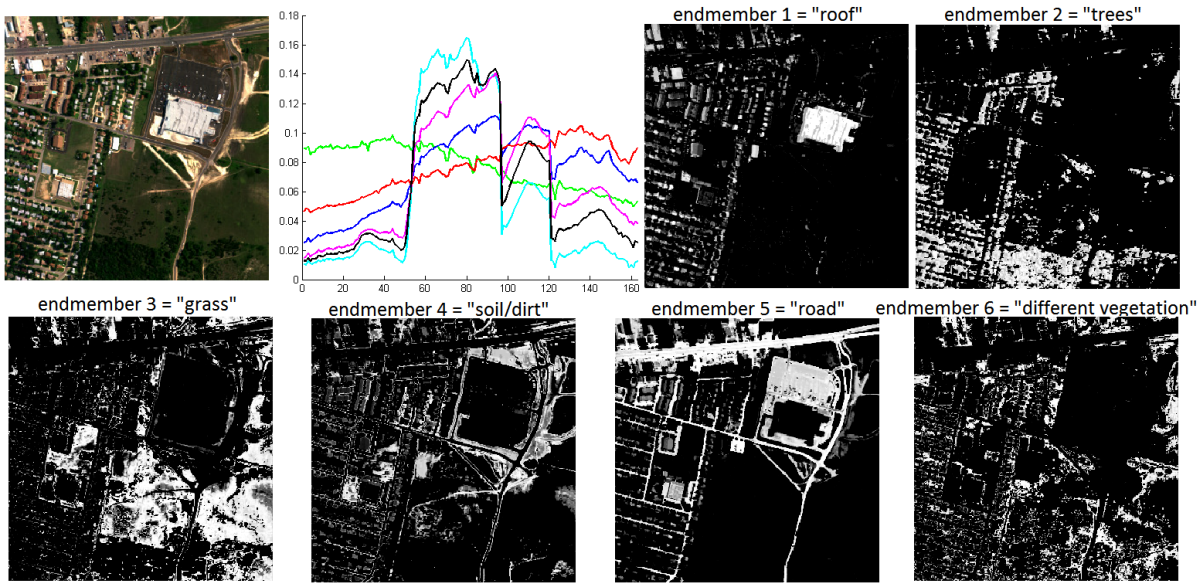

Fig. 2: Results on the urban hyperpectral image

to a much wider class of problems. For any non-negative matrix factorization or dictionary learning method, it could be interesting to find physically meaningful dictionary atoms by restricting the atoms to appear in the data. Hence, one could develop a concept of non-negative principal component analysis restricting the principal components to be part of the data. 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.

Acknowledgments - We would like to acknowledge Laura Balzano and Rob Nowak, who are working on a similar selection framework without non-negativity constraints, partially motivated by problems in sensor networks, and had provided important comments and feedback.

5. REFERENCES

[1] A. Zare, "Hyperspectral endmember detection and band selection using Bayesian methods," Ph.D. Thesis, 2008.

[2] C. Boutsidis, M. W. Mahoney, and P. Drineas, "An improved approximation algorithm for the column subset selection problem," in *SODA '09: Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2009, pp. 968–977.

[3] 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, 2007.

[4] A. Szlam, Z. Guo, and S. Osher, "A split Bregman method for non-negative sparsity penalized least squares with applications to hyperspectral demixing," Tech. Rep., 2010, UCLA CAM Report [10-06].

[5] J. Greer, "Sparse demixing," *SPIE proceedings on Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XVI*, vol. 7695, pp. 76951O–76951O–12, 2010.

[6] A. Castrodad, Z. Xing, J. Greer, E. Bosch, L. Carin, and G. Sapiro, "Learning discriminative sparse models for source separation and mapping of hyperspectral imagery," Submitted September 2010, http://www.ima.umn.edu/preprints/oct2010/oct2010.html.

[7] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," Tech. Rep., 2009, arXiv:0904.3523v2.

[8] J. A. Tropp, "Algorithms for simultaneous sparse approximation: part II: Convex relaxation," *Signal Process.*, vol. 86, no. 3, pp. 589–602, 2006.

[9] Z. Lin, Y. Yu, and G. Liu, "Robust subspace segmentation by low-rank representation," 2010, Accepted in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*.

[10] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite-element approximations," *Comp. Math. Appl.*, vol. 2, pp. 17–40, 1976.

[11] 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.

[12] E. Esser, "Applications of Lagrangian-based alternating direction methods and connections to split Bregman," Tech. Rep., 2009, UCLA CAM Report [09-31].

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

[14] J. M. P. Nascimento and J. M. Bioucas-Dias, "Blind hyperspectral unmixing," in *Proceedings of the SPIE Conference on Image and Signal Processing for Remote Sensing XIII*, 2007, vol. 6748.

[15] "Open source matlab hyperspectral toolbox," 2010. Version 0.04. http://matlabhyperspec.sourceforge.net/.