NONNEGATIVE LOW-RANK SPARSE COMPONENT ANALYSIS

Jeremy E. Cohen

Univ Rennes, Inria, CNRS, IRISA, France

Nicolas Gillis[*]

University of Mons, Belgium

ABSTRACT

In this paper we consider a variant of the dictionary learning problem where the dictionary has full rank, the coefficients have a fixed sparsity level, and both the coefficients and the dictionary are nonnegative. It is equivalent to k-sparse nonnegative matrix factorization (K-NMF). This model is encountered in source separation where nonnegative linear combinations of a few components generate the data points (samples), such as in hyperspectral images. We first discuss the impact of nonnegativity on the identifiability of low-rank sparse component analysis (LRSCA), building upon recent advances. Then, as a main contribution, we propose two algorithms to train K-NMF: one based on alternating optimization and exact sparse coding, the other based on a nonnegative variant of K-subspace. We show on noiseless simulated data that our methods outperform by a large margin the state of the art. Finally, we apply our methods for the spectral unmixing of a hyperspectral image.

***Index Terms*—** dictionary learning, nonnegative matrix factorization, sparsity, identifiability, subspace clustering

1. INTRODUCTION

Sparse component analysis (SCA) is a dictionary learning model that leverages sparsity to solve an otherwise ill-posed problem. Formally, SCA is the following problem:

Definition 1 (SCA)**.** *Given a matrix $M \in \mathbb{R}^{d\times n}$ and an integer r, find $D \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times n}$ where the columns of B are at least k-sparse with $k < r$ such that $M = DB$.*

SCA has been central in signal processing over the last decade, often under the name *dictionary learning*. In spite of the large use of this model for various applications such as image denoising, inpainting or classification [1], few conditions are available that guarantee uniqueness of matrices D and B satisfying Def. 1 in a deterministic scenario. There are however many results in probabilistic settings where the columns of B are drawn from specific distributions; see [2] and the references therein. Recently, focussing on the low-rank case ($\mathrm{rank}(D) = r,\ r \leq d$), a strong identifiability result was obtained:

[*]Authors acknowledge the support of the F.R.S.-FNRS (incentive grant for scientific research n° F.4501.16), and of the European Research Council (ERC starting grant n° 679515).

Theorem 1 ([3])**.** *An SCA decomposition $M = DB$ with $\mathrm{rank}(D) = r$ is essentially unique (that is, up to permutation and scaling of the columns of D and rows of B) if on each subspace spanned by all but one column of D, there are $\left\lfloor \frac{r(r-2)}{r-k} \right\rfloor + 1$ data points with spark r (that is, any subset of $r-1$ points are linearly independent).*

In this paper, we investigate a constrained version of low-rank SCA (LRSCA) where factor matrices D and B are required to be nonnegative. This leads to the following nonnegative LRSCA problem, which is equivalent to k-sparse nonnegative matrix factorization (K-NMF).

Definition 2 (K-NMF)**.** *Given a nonnegative matrix $M \in \mathbb{R}^{d\times n}_+$ and an integer r, find $D \in \mathbb{R}^{d\times r}_+$ and $B \in \mathbb{R}^{r\times n}_+$ where the columns of B are at least k-sparse with $k < r$ and $\mathrm{rank}(D) = r$ such that $M = DB$.*

In this paper, we discuss the identifiability of K-NMF and dedicated algorithms. The contributions of this paper are the following:

• We show that K-NMF is identifiable under milder conditions than LRSCA for $d = 3$. Generalization to any rank, dimensions and sparsity level is a promising direction for further research.

• We briefly survey the literature for tackling K-NMF: most methods rely either on alternating optimization using sparse nonnegative least squares (sNNLS) or on subspace estimation. We propose two algorithms: an exact combinatorial algorithm for sNNLS that is used in combination of a standard alternating approach, and a nonnegative adaptation of K-subspace.

• We compare the performance of the proposed methods with existing ones on simulated data sets where the impact of the dimension d on the performance is studied. The main two observations are that (i) solving the sNNLS subproblems exactly can significantly impact the quality of the solution, and (ii) the nonnegative adaptation of K-subspace often outperforms alternating optimization methods. Finally, we showcase K-NMF on spectral unmixing of an hyperspectral image (HSI) using our proposed methods.

2. IDENTIFIABILITY OF K-NMF

In this section, we discuss the identifiability of K-NMF, making use of recent works for the identifiability of LRSCA [3].

We do not formally prove a general identifiability results in this short paper, but instead explain why K-NMF requires milder conditions than LRSCA for identifiability, which is verified for $d = 3$. The rationale goes as follows. Suppose there exist two solutions to the K-NMF problem, and suppose without loss of generality that the data points and basis elements are scaled so that they belong to the unit simplex (that is, $||M(:, j)||_1 = 1$ for all j and $||D(:, k)||_1 = 1$ for all k). Then the data points are located on the the intersection of the convex hull of two polytopes with r vertices. Such an intersection is described by a finite number of low-dimensional subspaces. In order to ensure uniqueness of K-NMF, there must be sufficiently many well-located data points in order for such intersections not to be able to contain them all. Because of nonnegativity, these intersections are more constrained than for LRSCA. Let us show this for $d = 3$; see Fig. 1. In that case, the only non-trivial case is $k = 2$, and a solution of K-NMF can be represented as a triangle whose edges contain the data points. If the solution is not unique, the data points are located on the intersection of two triangles. Since two triangles can intersect on at most six points, K-NMF will be unique if there are at least two data points on each edge and one edge contains at least three data points. On Fig. 1, adding a single data point on any segment makes the factorization unique. In the mean time, without nonnegativity, non-unique LRSCA exists with nine data points, three per subspace spanned by two columns of D [3]. Therefore, regarding identifiability of LRSCA, nonnegativity plays an important role yet unexplored in the literature, as far as we know. Consequently, one way to study the identifiability of K-NMF is to provide sufficient conditions on the data under which such intersection cannot exist unless the two solutions coincide. However the geometry of such an intersection can be quite complex, especially since some facets might belong to both factorizations. Note that K-NMF is identifiable under the same sufficient conditions as LRSCA stated in Theorem 1 since it is a particular instance of LRSCA. We conjecture that it can be identifiable under milder assumptions, as it is the case for $d = 3$ as discussed above. This is an important topic for further research.

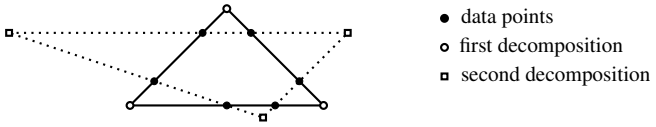

Fig. 1. A case where parameters of K-NMF are not identifiable.

3. ALGORITHMS FOR K-NMF

Let us now discuss the computational aspects of K-NMF. Using the Frobenius norm as an error metric, computing K-NMF can be formulated as the follows

$$\min_{D \in \mathbb{R}_+^{d \times r}, B \in \mathbb{R}_+^{r \times n}} \quad \|M - DB\|_F^2$$
$$\text{such that} \quad \|B(:, j)\|_0 \leq k < r \text{ for } 1 \leq j \leq n. \quad (1)$$

Solving (1) is bound to be difficult, since both NMF and SCA are NP-hard problems [4, 5].

Existing algorithms Traditionally, both NMF and SCA are solved using alternating optimization (AO) on D and B [6, 2]. In the low-rank case, solving (1) with a fixed B amounts to solving a nonnegative least squares (NNLS) problem, which can be done efficiently using for example an active-set method [7] or coordinate descent [8]. However, solving (1) for D fixed is a challenging task known in the litterature as nonnegative sparse coding [9] or sparse NNLS (sNNLS). Similarly to sparse coding, the two most important families of heuristics to obtain a solution to sNNLS are

• Matching pursuit methods that compute scalar products between data points and the known D. We refer to these as Nonnegative Orthogonal Matching Pursuit (NNOMP) variants [10, 11, 12, 13].

• LASSO methods where sparsity is enforced using a ℓ_1 norm penalty, a convex surrogate of the ℓ_0 norm. Various optimization solutions exist such as multiplicative update [14, 15], projected gradient [16] or coordinate descent [8].

On top of these two large families, an adaptation of K-SVD has also been developped for K-NMF [17]. Both NNOMP and LASSO approaches have the drawback of solving the sNNLS subproblems for variable B in a suboptimal way. Of course these methods have theoretical guarantees of providing the optimal solution to sNNLS under some conditions on D [18, 19]. However, even if such conditions are verified for the true D, it is unlikely for such conditions to be satisfied at each step of the AO scheme. Therefore, AO using for instance NNOMP will most likely not solve optimally the subproblems of finding B with known D at every iterations which may lead to rather different solutions than if the subproblems were solved exactly; see Section 4.

Another class of algorithms for K-NMF are based on subspace clustering methods. A well-known subspace clustering algorithm is an adaptation of k-means, sometimes refered to as K-subspace [20]. For the sake of simplicity, we only discuss K-subspace in this paper, but there exist many other subspace clustering methods; see [21] for an overview. K-subspace is also an alternating algorithm, but instead of alternating between B and D as discussed above, it alternates between the estimation of (i) the span of $D \setminus D(:, j)$ for all j and (ii) the positions of the zero values in each column of B. In fact, as discussed in Section 2, in the case $k = r - 1$, data points lie on the union of r hyperplanes generated by all columns of D but one. Therefore, knowing to which hyperplane each data point belongs, it is possible to use any subspace estimation method such as PCA to compute a basis and scores for these data points. Since K-subspace explores the

search space in a different manner than the AO methods discussed above, it may output rather different results, as observed in Section 4. A drawback of K-subspace, on top of working only for $k = r - 1$, is that the estimation quality of D might be more sensitive to noise since D is obtained by intersecting the estimated subspaces. Moreover, as far as we know, there does not exist a variant of K-subspace taking nonnegativity into account.

In the next paragraphs, we propose two algorithms: one based on AO but solving the sNNLS subproblems exactly, and the other taking nonnegativity into account within K-subspace and generalizing for any k. These algorithms are rather straightforward to develop. What is rather interesting is our observations from Section 4: the two proposed methods can outperform standard AO strategies and K-subspace to recover the true underlying identifiable factors D and B.

ESNA: a combinatorial algorithm As for AO, we propose to compute K-NMF by solving (1) alternatively for D and B. As previously, D is updated by solving a nonnegative least squares problem. However, as opposed to most previous methods, when solving for B, we solve the sNNLS subproblem up to global optimality using a brute-force approach (further work includes using more sophisticated combinatorial approaches). In a nutshell, for each pattern of $r - k$ zeros and for each column of B, a NNLS problem is solved, and the best solutions is kept. The computational complexity of this step is $\binom{r}{k}\mathcal{O}(NNLS)(d, n, k)$ where $\mathcal{O}(NNLS)(d, n, k)$ refers to the complexity of solving the NNLS $\min_{X \in \mathbb{R}_+^{k \times n}} ||AX - Y||_F$, with $A \in \mathbb{R}^{d \times k}$ and $Y \in \mathbb{R}^{d \times n}$. We refer to this algorithm as Exact Sparse and Nonnegative Alternating least squares (ESNA). Computing exactly the solution of sNNLS may seem unrealistic, since it is typically coined as a very time consuming process with exponential algorithmic complexity. However, in K-NMF, both k and r are typically small, thus the term $\binom{r}{k}$ may remain sufficiently small in practical applications (e.g., in HSI, r is of the order of 10-20 while k is of the order of 2-5). In Section 4, ESNA is used to unmix a HSI with $r = 6$ and $k = 2$. Also, note that the NNLS subproblems involve k variables instead of r, where k can be an order of magnitude smaller than r.

ESNA has a few interesting features. First, it is the best possible AO algorithm, therefore its performance provide an upper bound for other AO algorithms to reach. Second, it can be seen as a subspace clustering algorithm: finding D means finding the subspaces knowing both data affectation and scores, while computing all possible k-sparse B means finding the best affectation of data to k subspaces in terms of reconstruction error. However, there is no reason to believe that ESNA is the best subspace clustering algorithm, and in the following we suggest another nonnegative subspace clustering solution avoiding exponential complexity. Third, since it is an exact block coordinate descent algorithm, the cost function always decreases.

NOLRAK: a nonnegative subspace clustering algorithm While running experiments, we observed that K-subspace sometimes was able to identify the true underlying solution while ESNA was failing. Therefore, we propose a second algorithm adapted from K-subspace taking nonnegativity into account: NOnnegative Low-RAnk K-subspace (NOLRAK). Just like K-subspace, NOLRAK alternates between the affectation of each data points and the computation of a basis for each subspace. However, unlike K-subspace, it estimates nonnegative coefficients and uses the fact that the subspaces are generated using the columns of D, and thus does not resort to intersecting estimated subspaces. Also, each data point is not affected to only one hyperplane, but rather to $r - k$, to account for the knowledge of the sparsity level k. NOLRAK iterates the two following steps:

Step 1. Knowing D, for each data point m_i ($1 \leq i \leq n$) and for each hyperplane $\mathcal{H}_j = \text{span}(D \backslash D(:, j))$, the affectation of m_i to $r - k$ hyperplanes is obtained by computing the distances $\|m_i - \text{Proj}_{\mathcal{H}_j}(m_i)\|_2$ for $1 \leq j \leq r$ and selecting the $r-k$ smallest values. This fixes $r-k$ zeros for the ith column of B. In other words, we assign each data point to the $r - k$ closest hyperplanes.

Step 2. Knowing the positions of $r - k$ zeros in each column of B, the values of D and the other coefficients of B are obtained by computing the NMF of M with a fixed support for B. For this step, we use the so-called accelerated hierarchical alternating least squares (A-HALS) algorithm for NMF [22]. This step is a generalization of the subspace basis estimation in K-subspace, but uses NMF instead of PCA.

When using A-HALS which runs in $O(dnr)$ operations per iteration (like most NMF algorithms [23]), the computational complexity of NOLRAK per iteration is also in $O(dnr)$. Although NOLRAK requires computing a NMF at each iteration, it scales linearly in the dimension of the problem. Hence, it can be applied to large-scale data sets (as opposed to ESNA which can only be used for small r and/or small k).

4. NUMERICAL EXPERIMENTS

In this section, we perform some numerical experiments to assess the performance of ESNA and NOLRAK compared to existing approaches.

Simulated experiments We are interested in the following question: "For an exact K-NMF problem that satisfies the identifiability conditions of LRSCA (Theorem 1), are any of the algorithms discussed above able to exactly recover the dictionary D?" Due the space limitation, we only consider small-scale problems: the rank r is set to 4 and the sparsity level k to 3 or 2. The number of data points is set to $n = 200$ for $k = 3$ and $n = 300$ for $k = 2$. We generate entries of D and B using uniform distributions in $[0, 1]$. The columns of the matrix D are normalized, and B is sparsified so that identifiability is ensured (for $k = 3$, the four subspaces are randomly populated, while for $k = 2$, the six segments are

populated, each with 50 points). Performance are computed over $N = 100$ realizations[1]. The grid parameter is the sample dimension d, since a larger d leads to less correlation in D, thus heuristics such as NNOMP are expected to perform better as d increases. Initialization is random with optimal scaling, but we tried a separable NMF initialization [24] with similar results. Figure 2 presents the performance in terms of the recovery of D (meaning that the relative MSE $\frac{||D-D_{\text{computed}}||_F}{||D||_F} < 10^{-4}$) and average relative MSE of the following methods: ESNA, AO with SuNNOMP [13], truncated active set [11] (a.set NNOMP), LASSO with coordinate descent [25] (LASSO-HALS) and NMF with HALS [22] (NMF-HALS), NOLRAK and K-subspace [20] (KSub). For LASSO-HALS, we used a grid on the regularization parameter λ to choose it, more details are available alongside the code.

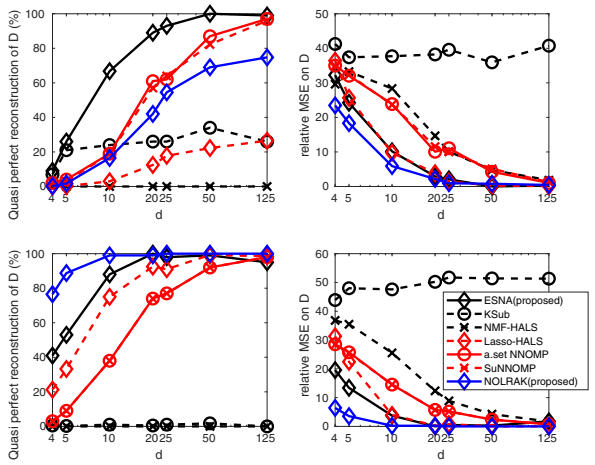

Fig. 2. Average performance of K-NMF algorithms to recover D up to 10^{-4} relative MSE (left) and relative MSE (right) for $k = 2$ (top) and $k = 3$ (bottom).

To help understanding the figure, note that the reconstruction errors on the factor D are distributed in two clusters: either a large error (missidentification, above a few percent of relative error), or machine precision. The only exception is LASSO-HALS which never reaches machine precision. We observe that AO solving inexactly the sNNLS subproblems (namely, NMF-HALS, LASSO-HALS, SuNNOMP and a.set NNOMP) do not manage to compute K-NMF reliably. On the other hand, both ESNA and NOLRAK perform well, and NOLRAK performs better than ESNA for $k = 2$ despite being computationally cheaper. Note that when $k = r - 1 = 3$, independently of d, K-subspace can sometimes identify D when other alternating methods fail.

Spectral Unmixing The Urban hyperspectral image (HSI)[2] contains 162 spectral bands with 307×307 pixels for each band. Noisy bands were removed beforehand. It is a rather simple and well understood data set: it is mainly composed of $r = 6$ types of materials (grass, trees, two types of roof tops, road and dirt) as reported in [26]. The scene represents a wallmart, its parking lot and the surrounding area (it can be found on Google maps with the keywords "copperas cove texas walmart"). In the K-NMF model, each column of the matrix $M \in \mathbb{R}^{162\times 94249}$ is the reflectance spectrum measured for a pixel, each column of $D \in \mathbb{R}^{162\times 6}$ is the spectral signature of a pure material and each column of $B \in \mathbb{R}^{6\times 94249}$ gives the abundance of the pure materials in the corresponding pixel. This HSI has relatively high spatial resolution hence most pixels contain no more than two materials so that using $k = 2$ makes sense. ESNA gives a solution with relative error $\frac{||M-DB||_F}{||M||_F} = 5.08\%$ which, in this case, performs better than NOLRAK with relative error of 7.91%. Fig. 3 displays the abundance maps (the matricized rows of B) which localize the pure materials relatively well.

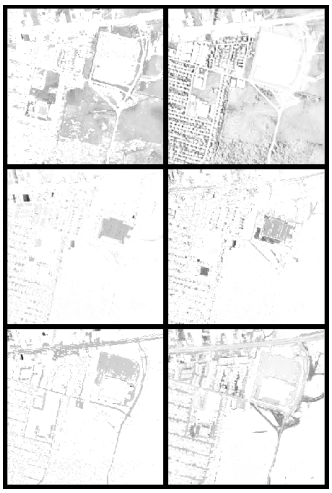

Fig. 3. Abundance maps extracted by ESNA on the Urban HSI. From left to right, top to bottom: grass, trees, roof tops 1 and 2, road, dirt.

5. CONCLUSION

In this paper, we have first discussed the identifiability of K-NMF showing that it requires milder assumption than LRSCA to obtain essentially unique decompositions. Then, we have proposed two algorithms, and shown their superiority compared to the state of the art. In particular, we have observed that (i) solving exactly the sNNLS subproblems within AO impacts significantly the outcome of the algorithm and allows to obtain significantly better solutions in many cases, and (ii) algorithms based on subspace clustering are a promising direction for research as they outperform AO in some cases. Far from closing the question on identifiability of K-NMF and on how to compute such decompositions, our results should be interpreted as an invitation to further study the theoretical properties and algorithms of K-NMF.

[1]Most parameters in the proposed experiment can be toyed with in the code available online on github at `cohenjer/Tensor_codes`

[2]Available at `http://www.agc.army.mil/`.

6. REFERENCES

[1] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 27–38, 2011.

[2] R. Gribonval, R. Jenatton, and F. Bach, "Sparse and spurious: dictionary learning with noise and outliers," *IEEE Trans. on Information Theory*, vol. 61, no. 11, pp. 6298–6319, 2015.

[3] J. E. Cohen and N. Gillis, "Identifiability of low-rank sparse component analysis," *arXiv preprint arXiv:1808.08765*, 2018.

[4] B. Natarajan, "Sparse approximate solutions to linear systems," *SIAM Journal on Computing*, vol. 24, no. 2, pp. 227–234, 1995.

[5] S. A. Vavasis, "On the complexity of nonnegative matrix factorization," *SIAM J. Optim.*, vol. 20, no. 3, pp. 1364–1377, 2010.

[6] N. Gillis, "Successive nonnegative projection algorithm for robust nonnegative blind source separation," *SIAM J. Imaging Sci.*, vol. 7, no. 2, pp. 1420–1450, 2014.

[7] J. Kim and H. Park, "Fast nonnegative matrix factorization: An active-set-like method and comparisons," *SIAM Journal on Scientific Computing*, vol. 33, no. 6, pp. 3261–3281, 2011.

[8] C-J. Hsieh and I. S. Dhillon, "Fast coordinate descent methods with variable selection for non-negative matrix factorization," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1064–1072.

[9] P. O. Hoyer, "Non-negative sparse coding," in *IEEE Workshop on Neural Networks for Signal Processing*. IEEE, 2002, pp. 557–565.

[10] A.M. Bruckstein, M. Elad, and M. Zibulevsky, "Sparse non-negative solution of a linear system of equations is unique," in *3rd Int. Symp. on Communications, Control and Signal Processing*. IEEE, 2008, pp. 762–767.

[11] R. Peharz and F. Pernkopf, "Sparse nonnegative matrix factorization with l0-constraints," *Neurocomputing*, vol. 80, pp. 38–46, 2012.

[12] M. Yaghoobi, D. Wu, and M.E. Davies, "Fast non-negative orthogonal matching pursuit," *Journal Club*, vol. 2015, no. 04, pp. 13, 2015.

[13] T.T. Nguyen, C. Soussen, J. Idier, and E-H. Djermoune, "An optimized version of non-negative omp," in *XXVIe Colloque GRETSI*, 2017.

[14] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, 2004.

[15] M. Mørup, K.H. Madsen, and L.K. Hansen, "Approximate ℓ_0 constrained non-negative matrix and tensor factorization," in *IEEE International Symposium on Circuits and Systems*. IEEE, 2008, pp. 1328–1331.

[16] J. Rapin, J. Bobin, A. Larue, and J-L. Starck, "Sparse and non-negative BSS for noisy data," *IEEE Trans. on Signal Processing*, vol. 61, no. 22, pp. 5620–5632, 2013.

[17] M. Aharon, M. Elad, and A.M. Bruckstein, "K-SVD and its non-negative variant for dictionary design," in *Wavelets XI*. International Society for Optics and Photonics, 2005, vol. 5914, p. 591411.

[18] J.A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. on Information theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[19] J.A. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Trans. on Information Theory*, vol. 52, no. 3, pp. 1030–1051, 2006.

[20] D. Wang, C. Ding, and T. Li, "K-subspace clustering," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2009, pp. 506–521.

[21] R. Vidal, "Subspace clustering," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 52–68, 2011.

[22] N. Gillis and F. Glineur, "Accelerated multiplicative updates and hierarchical als algorithms for nonnegative matrix factorization," *Neural Computation*, vol. 24, no. 4, pp. 1085–1105, 2012.

[23] N. Gillis, "The Why and How of Nonnegative Matrix Factorization," in *Regularization, Optimization, Kernels, and Support Vector Machines*, J.A.K. Suykens, M. Signoretto, and A. Argyriou, Eds., pp. 257–291. Chapman & Hall/CRC, 2014.

[24] N. Gillis and S.A. Vavasis, "Fast and robust recursive algorithmsfor separable nonnegative matrix factorization," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 36, no. 4, pp. 698–714, 2014.

[25] N. Gillis, "Sparse and unique nonnegative matrix factorization through data preprocessing," *J. Mach. Learn. Res.*, vol. 13, no. Nov, pp. 3349–3386, 2012.

[26] Z. Guo, T. Wittman, and S. Osher, "L1 unmixing and its application to hyperspectral image enhancement," in *Proc. SPIE Conference on Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XV*, 2009.