BOUNDED SIMPLEX-STRUCTURED MATRIX FACTORIZATION

Olivier Vu Thanh
Université de Mons

Nicolas Gillis
Université de Mons

BOUNDED SIMPLEX-STRUCTURED MATRIX FACTORIZATION

Olivier Vu Thanh   *Nicolas Gillis*   *Fabian Lecron*

Department of Mathematics and Operational Research, Faculté Polytechnique,
Université de Mons, Rue de Houdain 9, 7000 Mons, Belgium.

ABSTRACT

In this paper, we propose a new low-rank matrix factorization model, dubbed bounded simplex-structured matrix factorization (BSSMF). Given an input matrix X and a factorization rank r, BSSMF looks for a matrix W with r columns and a matrix H with r rows such that $X \approx WH$ where the entries in each column of W are bounded, that is, they belong to given intervals, and the columns of H belong to the unit simplex, that is, H is column stochastic. BSSMF generalizes nonnegative matrix factorization (NMF), and simplex-structured matrix factorization (SSMF). BSSMF is particularly well suited when the entries of the input matrix X themselves belong to a given interval; for example when the columns of X represent images. In this paper, we first provide identifiability conditions for BSSMF, that is, we provide conditions under which BSSMF admits a unique decomposition, up to trivial ambiguities. Then we propose a fast inertial algorithm for BSSMF. Finally, we illustrate the effectiveness of BSSMF to obtain interpretable features in the MNIST dataset.

***Index Terms*—** simplex-structured matrix factorization, nonnegative matrix factorization, identifiability, inertial first-order methods

1. INTRODUCTION

Low-rank matrix factorizations have recently emerged as very efficient models for unsupervised learning; see, e.g., [1]. The most notable example is principal component analysis (PCA). In the last 20 years, many new more sophisticated models have been proposed, such as sparse PCA that requires one of the factors to be sparse to improve interpretability [2], and low-rank matrix completion, also known as PCA with missing data, to handle missing entries in the input matrix [3]. Among such methods, nonnegative matrix factorization (NMF), popularized by Lee and Seung in 1999 [4], requires the factors of the decomposition to be component-wise nonnegative. More precisely, given an input matrix $X \in \mathbb{R}^{m\times n}$ and a factorization rank r, NMF looks for a nonnegative matrix W with r columns and a nonnegative matrix H with r rows such that $X \approx WH$. NMF has been shown to be useful in many applications, including topic modeling, image analysis, hyperspectral unmixing, and audio source separation; see [5, 6]. The main advantage of NMF compared to previously introduced low-rank models is that the nonnegativity constraints on the factors lead to an easily interpretable part-based decomposition. More recently, simplex-structured matrix factorization (SSMF) was introduced as a generalization of NMF [7]; see also [8]. SSMF does not impose any constraints on W, while it requires H to be column stochastic, that is, $H \geq 0$ and $H^\top e = e$ where e is the vector of all ones of appropriate dimension. SSMF is closely related to various machine learning problems, such as latent Dirichlet allocation, clustering, and the mixed membership stochastic block model; see [9].

The authors acknowledge the support by the European Research Council (ERC starting grant no 679515), and by the Fonds de la Recherche Scientifique - FNRS and the Fonds Wetenschappelijk Onderzoek - Vlanderen (FWO) under EOS Project no O005318F-RG47.

In this paper, we introduce bounded simplex-structured matrix factorization (BSSMF). Given vectors $a \leq b \in \mathbb{R}^m$, we write $W(:,k) \in [a,b]$ if $a_i \leq W(i,k) \leq b_i$ for all k.

Definition 1 (BSSMF). *Let $X \in \mathbb{R}^{m\times n}$, r be an integer, and $a, b \in \mathbb{R}^m$ with $a \leq b$. The pair $(W,H) \in \mathbb{R}^{m\times r} \times \mathbb{R}^{r\times n}$ is a BSSMF of X of size r for the interval $[a,b]$ if $X = WH$, $W(:,k) \in [a,b]$ for all k, $H \geq 0$, and $H^\top e = e$.*

BSSMF reduces to SSMF when $a_i = -\infty$ and $b_i = +\infty$ for all i, and to NMF when $a_i = 0$ and $b_i = +\infty$ for all i, after a proper normalization of X; see the discussion in [7].

The contribution of this paper is fourfold: (1) Introduce and motivate BSSMF (Section 2), (2) Provide an identifiability result for BSSMF (Section 3), (3) Propose an efficient algorithm for BSSMF (Section 4), and (4) Show the usefulness of BSSMF to extract meaningful features (Section 5).

2. MOTIVATING BSSMF

The motivation to introduce BSSMF is mostly threefold, as described in the next three paragraphs.

Bounded low-rank approximation. A necessary condition for X to admit a BSSMF $X = WH$ (Definition 1) is that $X(:,j) \in [a,b]$ for all j. In fact, $WH(:,j) \in [a,b]$ for all j since H is column stochastic and $W(:,k) \in [a,b]$ for all k. In other words, the constraints of BSSMF implicitly impose that all columns of WH belong to the interval $[a,b]$. When the data X naturally belongs to intervals (e.g., movie ratings or pixel intensities), imposing the approximation WH to belong to the same interval typically leads to better approximations, taking into account this prior information. Such a model, imposing bounds on the approximation via the constraints $\ell \leq WH \leq L$ for some $\ell, L \in \mathbb{R}$, was for example considered in [10], and applied successfully to recommender systems where ratings typically belong to the interval [1,5] (e.g., Netflix and MovieLens data sets). However, this model does not allow to interpret the basis factor, W, in the same way as the data, while it is also difficult to interpret the factor H. In fact, only imposing bounds on WH typically leads to dense factors W and H, which are not easily interpretable. This leads to the second motivation to introduce BSSMF.

Interpretability. Imposing that the entries in W belong to some interval and that H is column stochastic allows us to easily interpret both factors: the columns of W can be interpreted in the same way as the columns of X (e.g., as movie ratings, or pixel intensities), while the columns of H provide a soft clustering of the columns of X, since $X(:,j) = WH(:,j)$. In other words, the columns of W can be thought of cluster centroids, while the entry $H(k,j)$ indicates the membership of $X(:,j)$ to cluster k.

A closely related model is bounded component analysis (BCA) proposed in [11, 12], where the columns of the matrix W are assumed to belong to compact sets (hyper-rectangle being a special case), while no constraints is imposed on H. Again, without any constraints on H, BCA will generated dense factors which are difficult to interpret. An identifiability theorem is provided in [11, 12] under the strong condition that H is separable, which requires H to contain the identity matrix, up to permutation and scaling, as a submatrix. Moreover, in practice, it requires to balance two terms in the objective function: the data fitting term and the volume of $\text{conv}(W)$, which is not always easy to handle. In this paper, we will provide a much weaker condition for identifiability of BSSMF, while BSSMF is parameter free. This leads to the third motivation to introduce BSSMF.

Identifiability. A drawback of NMF and SSMF is that they are typically not identifiable without using additional constraints or regularization; see Section 3. Identifiability is key in practice as it allows to recover the ground truth that generated the data; see [5]. As we will show in Section 3, BSSMF is identifiable under relatively mild conditions, while *BSSMF does not require parameter tuning*, as opposed to most regularized NMF/SSMF models that are identifiable.

3. IDENTIFIABILITY OF BSSMF

Let us first discuss the identifiability of SSMF and NMF. Without further requirements, SSMF is never identifiable; which follows from a result for semi-NMF which is a factorization model than requires only one factor, H, to be nonnegative [13]. To obtain identifiability for SSMF, one needs either to impose additional constraint on W and/or H such as sparsity [8], or look for a solution minimizing a certain function. In particular, the solution (W, H) that minimizes the volume of the convex hull of the columns of W and the origin, namely $1/r! \det(W^\top W)^{1/2}$, is essentially unique[1] [14, 15] given that H satisfies the so-called sufficiently scattered condition (SSC) defined as follows.

Definition 2 (SSC). *The matrix $H \in \mathbb{R}^{r\times n}_+$ is sufficiently scattered if the following two conditions are satisfied:*

[SSC1] $\mathcal{C} = \{x \in \mathbb{R}^r_+ \mid e^\top x \geq \sqrt{r-1}\|x\|_2\} \subseteq \text{cone}(H)$.

[SSC2] There does not exist any orthogonal matrix Q such that $\text{cone}(H) \subseteq \text{cone}(Q)$, except for permutation matrices.

SSC1 requires the columns of H to contain the cone $\mathcal{C}$, which is tangent to every facet of the nonnegative orthant. This implies that H has some degree of sparsity, in fact it requires at least $r-1$ zeros in each row of H [6, Theorem 4.28]; see also [5]. SSC2 is a mild regularity condition which is typically satisfied when SSC1 is satisfied. In practice, because of noise and model misfit, optimization models need to balance the data fitting term which measures the discrepancy between X and WH, and the volume regularization for $\text{conv}(W)$. This requires the tuning of a penalty parameter, which is a nontrivial process [16, 17].

As opposed to SSMF, NMF decompositions can be identifiable without the use of additional constraints or regularizers. One of the most relaxed sufficient condition for identifiability is based on the SSC.

Theorem 1. *[18, Theorem 4] If $W^\top \in \mathbb{R}^{r\times m}_+$ and $H \in \mathbb{R}^{r\times n}_+$ are sufficiently scattered, then the NMF (W, H) of $X = WH$ is essentially unique.*

In practice, it is not likely for both $W^\top$ and H to satisfy the SSC. Typically H will satisfy the SSC, as it is sparse. However, in many applications, $W^\top$ will not satisfy the SSC; in particular in applications where W is not sparse, e.g., in imaging or recommender systems. This is why regularized NMF models have been introduced, including sparse and minimum-volume NMF; see [6, Chapter 4].

A main motivation to introduce BSSMF is that it is identifiable under weaker conditions than NMF. We now state our main identifiability result for BSSMF.

Theorem 2. *Let $a \leq b$, $W \in \mathbb{R}^{m\times r}$ and $H \in \mathbb{R}^{r\times n}$ be such that $W(:,k) \in [a,b]$ for all k, $H \geq 0$, and $H^\top e = e$. If $\binom{W-ae^\top}{be^\top - W}^\top \in \mathbb{R}^{r\times 2m}$ and $H \in \mathbb{R}^{r\times n}$ satisfy the SSC, then the BSSMF (W, H) of $X = WH$ of size $r = \text{rank}(X)$ for the interval $[a,b]$ is essentially unique.*

Proof. The proof[2] follows from the identifiability result of NMF (Theorem 1) and the following observation:

[1]A decomposition is essentially unique if it is unique up to permutation of the rank-one terms $W(:,k)H(k,:)$, and up to their scalings $(W(:,k)/\alpha_k)(\alpha H(k,:))$ for $\alpha_k > 0$.

[2]Due to the lack of space, we only provide here the sketch of the proof, the formal proof will be available with the long version of this manuscript.

If $X = WH$ is a BSSMF of X for the interval $[a, b]$, then

$$be^\top - X = (be^\top - W)H \quad \text{and} \quad X - ae^\top = (W - ae^\top)H$$

are NMF decompositions. □

The condition that $\binom{W-ae^\top}{be^\top-W}^\top$ satisfies the SSC is much weaker than requiring $W^\top$ to satisfies the SSC in NMF. For example, in recommender systems, with $W(i,j) \in [1,5]$ for all (i,j), many entries of W are expected to be equal to 1 or to 5 (the maximum and minimum ratings), so that $\binom{W-ae^\top}{be^\top-W}^\top$ will be sparse, and hence likely to satisfy the SSC [5, 6]. On the other hand, W is positive and hence $X = WH$ will not admit an essentially unique NMF (a necessary condition for NMF to be unique is that the sparsity pattern of each column of W is not contained in the sparsity pattern of the other columns [19]). In Section 5, we will illustrate this fact on handwritten digits.

4. INERTIAL BLOCK-COORDINATE DESCENT

We formulate the BSSMF optimization problem as follows

$$\min_{W,H} \quad g(W,H) := \frac{1}{2}\|X - WH\|_F^2$$
$$\text{s.t.} \quad W(:,k) \in [a,b] \text{ for all } k, H \geq 0, \text{ and } H^\top e = e, \quad (1)$$

We now apply TITAN, an inertial block majorization minimization algorithm proposed in [20], to solve (1). TITAN updates one block, W or H, at a time while fixing the value of the other block. In order to update W (resp. H), TITAN chooses a block surrogate function for W (resp. H), embeds an inertial term to this surrogate function and then minimizes the obtained inertial surrogate function. The gradient of g with respect to W, $\nabla_W g(W,H) = (WH - X)H^\top$, is Lipschitz continuous in W with the Lipschitz constant $\|H\|_2^2$. Similarly, $\nabla_H g(W,H)$ is Lipschitz continuous in H with constant $\|W\|_2^2$. Hence, we choose the Lipschitz gradient surrogate for both W and H and choose the Nesterov-type acceleration as analysed in [20, Section 4.2.1] and [20, Remark 4.1], see also [20, Section 6.1] and [21] for similar applications. We describe TITAN for solving (1) in Algorithm 1 and name it by TITANized BSSMF. As proved in [20, Theorem 3.2], TITANized BSSMF guarantees a subsequential convergence, that is, every limit point of the generated sequence is a stationary point of Problem (1). Note that, without the inertial terms (that is, taking $\beta_W = \beta_H = 0$ in Algorithm 1), TITAN reduces to an alternating gradient descent method equivalent to PALM [22]. To initialize (W,H) in Algorithm 1, we use the successive nonnegative projection algorithm (SNPA) that sets $W = X(:,\mathcal{K})$ for some well-chosen index set $\mathcal{K}$ [23].

5. NUMERICAL EXPERIMENTS

When applied on an pixel-by-image matrix X (each column of X is a vectorized handwritten digit), NMF allows to automatically extract common features within the columns of W [4]. In this section, we use one of the most widely used data sets for handwritten digits: MNIST ($60,000$ images, 28×28 pixels) [24]. For BSSMF to make more sense, we preprocess X so that the intensities of the pixels in each digit belong to the interval $[0,1]$ (first remove from $X(:,j)$ its minimum entry, then divide by the maximum entry minus the minimum entry). In our experiments, NMF and BSSMF are given the same initialization provided by SNPA and the same algorithm is used: TITAN which is also very effective for NMF [20]. All experiments are run in MATLAB R2021a on a PC with an Intel(R) Core(TM) i7-9750H CPU @ 2.60GHz and 16GiB RAM. The code is available on https://gitlab.com/vuthanho/bounded-simplex-structered-matrix-factorization/.

Algorithm 1 TITANized BSSMF

Input: $X \in \mathbb{R}^{m\times n}$, $W_0 \in \mathbb{R}_+^{m\times r}$, $H_0 \in \mathbb{R}_+^{r\times n}$, $a \leq b \in \mathbb{R}^m$

$\alpha_1 = 1$, $\alpha_2 = 1$, $W_{old} = W_0, H_{old} = H_0$, $L_W^{prev} = L_W = \|H_0H_0^\top\|_2$, $L_H^{prev} = L_H = \|W_0^\top W_0\|_2$
repeat
  while stopping criteria not satisfied **do**
    $\alpha_0 = \alpha_1, \alpha_1 = (1 + \sqrt{1+4\alpha_0^2})/2$
    $\beta_W = \min\left[\,(\alpha_0 - 1)/\alpha_1, 0.9999\sqrt{L_W^{prev}/L_W}\,\right]$
    $\overline{W} \leftarrow W + \beta_W(W - W_{old})$, $W_{old} \leftarrow W$
    $W \leftarrow \left[\overline{W} + \frac{(X-\overline{W}H)H^\top}{L_W}\right]_{[a,b]}$
    $L_W^{prev} \leftarrow L_W$
  end while
  $L_H \leftarrow \|W^\top W\|_2$
  while stopping criteria not satisfied **do**
    $\alpha_0 = \alpha_2, \alpha_2 = (1 + \sqrt{1+4\alpha_0^2})/2$
    $\beta_H = \min\left[\,(\alpha_0 - 1)/\alpha_2, 0.9999\sqrt{L_H^{prev}/L_H}\,\right]$
    $\overline{H} \leftarrow H + \beta_H(H - H_{old})$, $H_{old} \leftarrow H$
    $H \leftarrow \left[\overline{H} + \frac{W^\top(X-W\overline{H})}{L_H}\right]_\Delta$
    $L_H^{prev} \leftarrow L_H$
  end while
  $L_W \leftarrow \|HH^\top\|_2$
until some stopping criteria is satisfied

Notation: $[.]_S$ denotes the projection onto the set S, column wise; $\Delta = \{x | x \geq 0, e^\top x = 1\}$ is the unit simplex.

Interpretability. Let us take a toy example with $n = 500$ randomly selected digits and $r = 10$, in order to visualize the natural interpretability of BSSMF. Fig. 1a shows the features learned by NMF which look like parts of digits. On the other hand, the features learned by BSSMF in Fig. 1b look mostly like real digits, because of the bound constraint and the sum-to-one constraints. We distinguish numbers (like 7, 3 and 6). From a clustering point of view, this is of much interest because a column of H which is near a ray of the unit

(a) NMF (b) BSSMF

Fig. 1: Basis matrix W for $r = 10$ for MNIST with 500 digits.

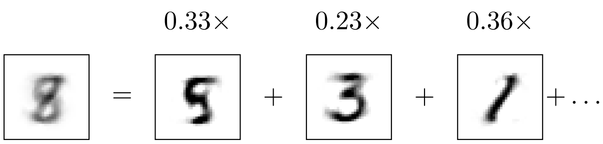
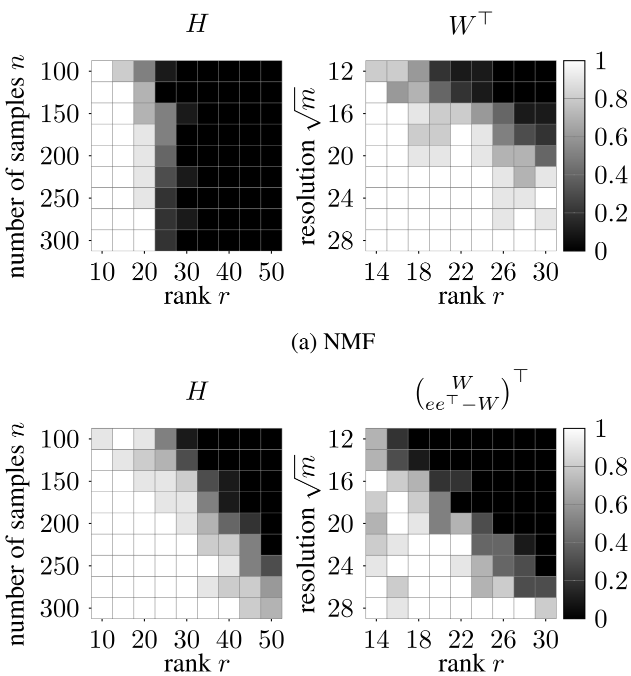


Fig. 2: Decomposition of an eight by BSSMF with $r = 10$.



(a) NMF

(b) BSSMF

Fig. 3: Ratio, over 10 runs, of the factors generated by NMF in Fig. 3a and by BSSMF in Fig. 3b that satisfy the necessary condition for SSC1 (white squares indicate that all matrices meet the necessary condition, black squares that none do).

simplex can directly be associated with the corresponding digit from W. In this toy example, due to r being small, an 8 cannot be seen. Nonetheless, an eight can be reconstructed as the sum of a 5, a 3 and an italic 1; see Fig. 2 for an example. Note that since BSSMF is more constrained than NMF, its reconstruction error is larger than that of NMF. For our example ($r = 10$), BSSMF has relative error $\|X - WH\|_F / \|X\|_F$ of 61.56%, and NMF of 59.04%.

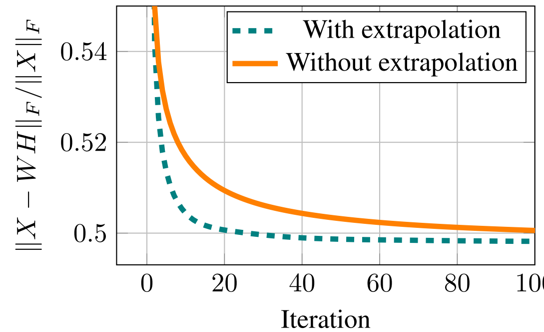


Fig. 4: BSSMF solved via Algorithm 1 with and without the extrapolation (MNIST data set for $r = 50$ and $n = 1000$).

Identifiability. As it is NP-hard to check the SSC [18], we will perform an experiment where only a necessary condition for SSC1 is verified, namely [6, Alg. 4.2]. To see when H satisfies this condition, we first vary n from 100 to 300 for m fixed (=28×28). For $W^\top$, we fix n to 300, and downscale the resolution m from 28×28 to 12×12 with a linear interpolation (`imresize3` in MATLAB), and the rank r is varied from 12 to 30. Recall that both factors need to satisfy the SSC to correspond to an essentially unique factorization. In Fig. 3a, we see that $W^\top$ of NMF often satisfies the necessary condition. This is due to NMF learning "parts" of objects, which are sparse by nature. On the contrary, even for relatively large n, H is too dense to satisfy the necessary condition. For $r \geq 30$, the factor H generated by NMF never satisfies the condition. Meanwhile, in Fig. 3b we see that H of BSSMF always satisfies the condition when $n \geq 225$ for $r = 30$ and more generally, if n and m are large enough, both H and $\left(\frac{W}{ee^\top - W}\right)^\top$ satisfy the necessary condition. This substantiates that BSSMF provides essentially unique factorizations more often than NMF does.

Extrapolation. Let us show the importance of the extrapolation sequence presented in Section 4. We fix r to 50 and n to 1000. The test is run 20 times and is performed on BSSMF with and without extrapolation. The average is then computed among the runs and displayed on Fig 4. As expected, BSSMF with extrapolation converges to a stationary point much faster than BSSMF without extrapolation, as observed for similar problems [20, 21].

6. CONCLUSION

We have proposed and motivated a new matrix factorization model, namely BSSMF, proved its identifiability under the SSC (Theorem 1), proposed an efficient inertial algorithm to solve it (Alg. 1), and illustrated our findings on the MNIST data set (Section 5). In particular, we showed that BSSMF is more often identifiable than NMF, while leading to more easily interpretable factors.

Acknowledgment. We are grateful to Hien Le for helpful discussions on the design of Algorithm 1 using TITAN.

7. REFERENCES

[1] N. Vaswani, Y. Chi, and T. Bouwmans, "Rethinking PCA for modern data sets: Theory, algorithms, and applications," *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1274–1276, 2018.

[2] A. d'Aspremont, L. El Ghaoui, M.I. Jordan, and G. Lanckriet, "A direct formulation for sparse PCA using semidefinite programming," *SIAM Review*, vol. 49, no. 3, pp. 434–448, 2007.

[3] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[4] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[5] X. Fu, K. Huang, N.D. Sidiropoulos, and W.-K. Ma, "Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications," *IEEE Signal Process. Mag.*, vol. 36(2), pp. 59–80, 2019.

[6] N. Gillis, *Nonnegative Matrix Factorization*, SIAM, 2020.

[7] R. Wu, W.-K. Ma, and X. Fu, "A stochastic maximum-likelihood framework for simplex structured matrix factorization," in *ICASSP*, 2017, pp. 2557–2561.

[8] M. Abdolali and N. Gillis, "Simplex-structured matrix factorization: Sparsity-based identifiability and provably correct algorithms," *SIAM J. Math. Data Sci.*, vol. 3, no. 2, pp. 593–623, 2021.

[9] A. Bakshi, C. Bhattacharyya, R. Kannan, D.P. Woodruff, and S. Zhou, "Learning a latent simplex in input sparsity time," in *ICLR*, 2021.

[10] R. Kannan, M. Ishteva, and H. Park, "Bounded matrix factorization for recommender system," *Knowledge and Information Systems*, vol. 39, no. 3, pp. 491–511, 2014.

[11] S. Cruces, "Bounded component analysis of linear mixtures: A criterion of minimum convex perimeter," *IEEE Trans. Signal Process.*, vol. 58(4), pp. 2141–2154, 2010.

[12] A.T. Erdogan, "A class of bounded component analysis algorithms for the separation of both independent and dependent sources," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5730–5743, 2013.

[13] N. Gillis and A. Kumar, "Exact and heuristic algorithms for semi-nonnegative matrix factorization," *SIAM J. Matrix Anal. Appl.*, vol. 36, no. 4, pp. 1404–1424, 2015.

[14] X. Fu, W.-K. Ma, K. Huang, and N.D. Sidiropoulos, "Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain," *IEEE Trans. Signal Process.*, vol. 63(9), pp. 2306–2320, 2015.

[15] C.-H. Lin, W.-K. Ma, W.-C. Li, C.-Y. Chi, and A. Ambikapathi, "Identifiability of the simplex volume minimization criterion for blind hyperspectral unmixing: The no-pure-pixel case," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5530–5546, 2015.

[16] A.M.S. Ang and N. Gillis, "Algorithms and comparisons of nonnegative matrix factorizations with volume regularization for hyperspectral unmixing," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 12, no. 12, pp. 4843–4853, 2019.

[17] L. Zhuang, C.-H. Lin, M.A.T. Figueiredo, and J.M. Bioucas-Dias, "Regularization parameter selection in minimum volume hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9858–9877, 2019.

[18] K. Huang, N.D. Sidiropoulos, and A. Swami, "Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition," *IEEE Trans. Signal Process.*, vol. 62, no. 1, pp. 211–224, 2013.

[19] S. Moussaoui, D. Brie, and J. Idier, "Non-negative source separation: range of admissible solutions and conditions for the uniqueness of the solution," in *ICASSP*, 2005.

[20] L.T.K. Hien, D.N. Phan, and N. Gillis, "An inertial block majorization minimization framework for nonsmooth nonconvex optimization," 2020, arXiv:2010.12133.

[21] O. Vu Thanh, A.M.S. Ang, N. Gillis, and L.T.K. Hien, "Inertial majorization-minimization algorithm for minimum-volume NMF," in *EUSIPCO*, 2021.

[22] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Math. Prog.*, vol. 146, no. 1, pp. 459–494, 2014.

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

[24] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.