Inertial Majorization-Minimization Algorithm for Minimum-Volume NMF

Olivier Vu Thanh[1], Andersen Ang[2,1], Nicolas Gillis[1], Le Thi Khanh Hien[1]
[1] Department of Mathematics and Operational Research, Faculté Polytechnique, Université de Mons
Rue de Houdain 9, 7000 Mons, Belgium
{olivier.vuthanh, manshun.ang, nicolas.gillis, thikhanhhien.le}@umons.ac.be
[2] Department of Combinatorics and Optimization, Faculty of Mathematics, University of Waterloo, Canada

***Abstract*—Nonnegative matrix factorization with the minimum-volume criterion (min-vol NMF) guarantees that, under some mild and realistic conditions, the factorization has an essentially unique solution. This result has been successfully leveraged in many applications, including topic modeling, hyperspectral image unmixing, and audio source separation. In this paper, we propose a fast algorithm to solve min-vol NMF which is based on a recently introduced block majorization-minimization framework with extrapolation steps. We illustrate the effectiveness of our new algorithm compared to the state of the art on several real hyperspectral images and document data sets.**

***Index Terms*—nonnegative matrix factorization, minimum volume, fast gradient method, majorization-minimization, hyperspectral imaging**

I. Introduction

Nonnegative Matrix Factorization (NMF) has been an active field of research since the seminal paper by Lee and Seung [1]. The success of NMF comes from many specific applications since many types of data are nonnegative; for example amplitude spectrograms in audio source separation, images, evaluations in recommendation systems, and documents represented by vectors of word counts; see [2] and the references therein. Compared to other unconstrained factorization models such as PCA/SVD, NMF requires the factors to be nonnegative. This constraint naturally leads to factors that are more easily interpretable [1]. Nonetheless, there are two drawbacks with NMF: computability and identifiability.

Computability. As opposed to PCA/SVD, solving NMF is NP-hard in general [3]. Hence most NMF algorithms rely on standard non-linear optimization schemes without global optimality guarantee.

Identifiability. NMF solutions are typically not unique, that is, they are not unique even after removing the trivial scaling and permutation ambiguities of the rank-one factors; see [4] and the references therein. For NMF to have a unique solution, also known as identifiability, one needs to add additional structure to the sought solution. One way to ensure identifiability is the min-vol criterion, which minimizes the volume of one of the factors. If the sufficiently scattered condition (SSC) is satisfied, then identifiability holds for min-vol NMF [5]–[7].

Identifiability for min-vol NMF is a strong result that has been used successfully in many applications such as topic modeling and hyperspectral imaging [8], and audio source separation [7]. However, min-vol NMF is computationally hard to solve. In this paper, after introducing the considered min-vol NMF model in Section II, we propose a fast method to solve min-vol NMF in Section III. Our method is an application of a recent inertial block majorization-minimization framework called TITAN [9]. Experimental results on real data sets show that the proposed method performs better than the state of the art; see Section IV.

NG and LTKH acknowledge the support by the European Research Council (ERCstarting grant No 679515), the Fonds de la Recherche Scientifique - FNRS and the Fonds Wetenschappelijk Onderzoek - Vlaanderen (FWO) under EOS project O005318F-RG47.
The names of the last three authors are in alphabetical order.

II. Minimum-volume NMF

In the noiseless case, the exact NMF model is $\mathbf{M} = \mathbf{WH}$ where $\mathbf{M} \in \mathbb{R}_+^{m\times n}$ denotes the measured data, $\mathbf{W} \in \mathbb{R}_+^{m\times r}$ (resp. $\mathbf{H} \in \mathbb{R}_+^{r\times n}$) denotes the left factor (resp. the right factor). The idea behind the min-vol criterion, a.k.a. Craig's belief [10], is that the convex hull spanned by the columns of $\mathbf{W}$, denoted $\mathrm{conv}(\mathbf{W})$, should embrace all the data points as tightly as possible. In the absence of noise, min-vol NMF is formulated as follows

$$\min_{\mathbf{W},\mathbf{H}} \quad \det(\mathbf{W}^\top\mathbf{W}) \tag{1a}$$
$$\text{s.t.} \quad \mathbf{M} = \mathbf{WH}, \tag{1b}$$
$$\mathbf{H} \geq \mathbf{0},\ \mathbf{W} \geq \mathbf{0},\ \mathbf{1}^\top\mathbf{H} = \mathbf{1}^\top, \tag{1c}$$

where $\mathbf{1}$ is a vector of appropriate size containing only ones. The constraint (1c) ensures that every data point lies within the convex hull spanned by the columns of $\mathbf{W}$, that is, $\mathbf{M}(:,j) \in \mathrm{conv}(\mathbf{W})$ for all j. The volume of the convex hull of $\mathbf{W}$ and the origin in the subspace span by columns of $\mathbf{W}$, is proportional to $\det(\mathbf{W}^\top\mathbf{W})$; see for example [5]. Under the sufficiently scattered conditions (SSC), which requires the columns of $\mathbf{M}$ to be sufficiently spread within $\mathrm{conv}(\mathbf{W})$ or, equivalently, that $\mathbf{H}$ is sufficiently sparse, min-vol NMF has an essentially unique solution [5], [6]. A drawback of (1c) is that it requires the entries in each column of $\mathbf{H}$ to sum to one, which is not without loss of generality: it imposes that the columns of $\mathbf{M}$ belong to the convex hull of the columns of $\mathbf{W}$ as opposed to the conical hull when the equality constraints of (1c) are absent; see for example [2, Chapter 4].

It was recently shown that the same model where the constraint $\mathbf{1}^\top\mathbf{H} = \mathbf{1}^\top$ is replaced with $\mathbf{1}^\top\mathbf{W} = \mathbf{1}^\top$ retains identifiability [7]. The sum-to-one constraint on the columns of $\mathbf{W}$, that is, $\mathbf{1}^\top\mathbf{W} = \mathbf{1}^\top$, can be assumed w.l.o.g. via the scaling ambiguity of the rank-one factors $\mathbf{W}(:,k)\mathbf{H}(k,:)$ in any NMF decomposition. Moreover, the model with the constraint on $\mathbf{W}$ was shown to be numerically much more stable as it makes $\mathbf{W}$ better conditioned which is important because computing the derivative of $\det(\mathbf{W}^\top\mathbf{W})$ requires computing the inverse of $\mathbf{W}^\top\mathbf{W}$. We refer the interested reader to [2, Chapter 4.3.3] for a discussion on these models.

In the presence of noise, min-vol NMF is typically formulated via penalization. In this paper, we consider the following min-vol NMF model

$$\begin{aligned} \min_{\mathbf{W},\mathbf{H}} \quad & \frac{1}{2}\|\mathbf{M}-\mathbf{W}\mathbf{H}\|_F^2 + \frac{\lambda}{2}\log\det(\mathbf{W}^\top\mathbf{W}+\delta\mathbf{I}_r) \\ \text{s.t.} \quad & \mathbf{H}\geq\mathbf{0},\ \mathbf{W}\geq\mathbf{0},\ \mathbf{1}^\top\mathbf{W}=\mathbf{1}^\top, \end{aligned} \tag{2}$$

where $\|.\|_F$ is the Frobenius norm, $\lambda > 0$ is a parameter balancing the two terms in the objective function, $\mathbf{I}_r$ is the $r \times r$ identity matrix, and $\delta > 0$ is a small parameter that prevents $\log\det(\mathbf{W}^\top\mathbf{W})$ from going to $-\infty$ if $\mathbf{W}$ is rank deficient [11]. The use of the logarithm of the determinant is less sensible to very disparate singular values of $\mathbf{W}$, leading to better practical performances [8], [12].

Applications. In hyperspectral unmixing (HU), each column of $\mathbf{M}$ contains the spectral reflectance of a pixel, each row of $\mathbf{M}$ corresponds to the reflectance of a spectral band among all pixels, each column of $\mathbf{W}$ is the spectral signature of an endmember (a pure material in the image), and each column of $\mathbf{H}$ contains the proportion of each identified pure material in the corresponding pixel; see [13]. Geometrically, the min-vol NMF in (2) applied to HU consists of finding endmembers such that the convex hull spanned by them and the origin embraces as tightly as possible every pixels in $\mathbf{M}$. This is the so-called Craig's belief [10]. In document classification, $\mathbf{M}$ is a word-by-document matrix so that the columns of $\mathbf{W}$ correspond to topics (that is, set of words found simultaneously in several documents) while the columns of $\mathbf{H}$ allow to assign each documents to the topics it discusses [8].

III. New algorithm for min-vol NMF

As far as we know, all algorithms for min-vol NMF rely on two-block coordinate descent methods that update each block ($\mathbf{W}$ or $\mathbf{H}$) by using some outer optimization algorithm to solve the subproblems formed by restricting the min-vol NMF problem to each block. For example, the state-of-the-art method from [11] uses Nesterov fast gradient method to update each factor matrix, one at a time.

Our proposed algorithm for (2) will be based on the TITAN framework from [9]. TITAN is an inertial block majorization minimization framework for nonsmooth nonconvex optimization. It updates one block at a time while fixing the values of the other blocks, as previous min-vol NMF algorithms. In order to update a block, TITAN chooses a block surrogate function for the corresponding objective function (a.k.a. a majorizer), embeds an inertial term to this surrogate function and then minimizes the obtained inertial surrogate function. When a Lipschitz gradient surrogate is used, TITAN reduces to the Nesterov-type accelerated gradient descent step for each block of variables [9, Section 4.2]. The difference of TITAN compared to previous min-vol NMF algorithms is threefold:

1) The inertial force (also known as the extrapolation, or momentum) is used between block updates. This is a crucial aspect that will make our proposed algorithm faster: when we start the update of a block of variables (here, $\mathbf{W}$ or $\mathbf{H}$), we can use the inertial force (using the previous iterate) although the other blocks have been updated in the mean time.
2) TITAN allows to update the surrogate after each update of $\mathbf{W}$ and $\mathbf{H}$, which was not possible with the algorithm from [11] because it applied fast gradient from convex optimization on a fixed surrogate.
3) It has subsequential convergence guarantee, that is, every limit point of the generated sequence is a stationary point of Problem (2). Note that the state-of-the-art algorithm from [11] does not have convergence guarantees.

Remark. The block prox-linear (BPL) method from [14] can be used to solve (2) since the block functions in $\mathbf{W} \mapsto \frac{1}{2}\|\mathbf{M}-\mathbf{W}\mathbf{H}\|_F^2$ and in $\mathbf{H} \mapsto \frac{1}{2}\|\mathbf{M}-\mathbf{W}\mathbf{H}\|_F^2$ have Lipschitz continuous gradients. However, BPL applies extrapolation to the Lipschitz gradient surrogate of these block functions and requires to compute the proximal point of the regularizer $\frac{\lambda}{2}\log\det(\mathbf{W}^\top\mathbf{W}+\delta\mathbf{I}_r)$, which does not have a closed form. In contrast, TITAN applies extrapolation to the surrogate function of $\mathbf{W} \mapsto f(\mathbf{W},\mathbf{H})$ with a surrogate function for the regularizer $\frac{\lambda}{2}\log\det(\mathbf{W}^\top\mathbf{W}+\delta\mathbf{I}_r)$ (see Section III-A1). This allows TITAN to have closed-form solutions for the subproblems, an acceleration effect, and convergence guarantee.

A. Surrogate functions

An important step of TITAN is to define a surrogate function for each block of variables. These surrogate functions are upper approximation of the objective function at the current iterate. Denote

$$f(\mathbf{W},\mathbf{H}) = \frac{1}{2}\|\mathbf{M}-\mathbf{W}\mathbf{H}\|_F^2 + \frac{\lambda}{2}\log\det(\mathbf{W}^\top\mathbf{W}+\delta\mathbf{I}_r)$$

and suppose we are cyclically updating $(\mathbf{W},\mathbf{H})$. Let us denote $u_{\mathbf{W}_k}(\mathbf{W})$ the surrogate function of $\mathbf{W} \mapsto f(\mathbf{W},\mathbf{H}_k)$ to update $\mathbf{W}_k$, that is,

$$f(\mathbf{W},\mathbf{H}_k) \leq u_{\mathbf{W}_k}(\mathbf{W}) \ \text{ for all } \ \mathbf{W}\in\mathcal{X}_{\mathbf{W}}, \tag{3}$$

where $u_{\mathbf{W}_k}(\mathbf{W}_k) = f(\mathbf{W}_k,\mathbf{H}_k)$ and $\mathcal{X}_{\mathbf{W}}$ is the feasible domain of $\mathbf{W}$. Similarly, let us denote $u_{\mathbf{H}_k}(\mathbf{H})$ the surrogate function of $\mathbf{H} \mapsto f(\mathbf{W}_{k+1},\mathbf{H})$ to update $\mathbf{H}_k$, that is

$$f(\mathbf{W}_{k+1},\mathbf{H}) \leq u_{\mathbf{H}_k}(\mathbf{H}) \ \text{ for all } \ \mathbf{H}\in\mathcal{X}_{\mathbf{H}}, \tag{4}$$

where $u_{\mathbf{H}_k}(\mathbf{H}_k) = f(\mathbf{W}_{k+1},\mathbf{H}_k)$ and $\mathcal{X}_{\mathbf{H}}$ is the feasible domain of $\mathbf{H}$.

1) Surrogate function and update of $\mathbf{W}$*:* Denote $\mathbf{A} = \mathbf{W}^\top\mathbf{W}+\delta\mathbf{I}_r$, $\mathbf{B}_k = \mathbf{W}_k^\top\mathbf{W}_k+\delta\mathbf{I}_r$ and $\mathbf{P}_k = (\mathbf{B}_k)^{-1}$. Since $\log\det$ is concave, its first-order Taylor expansion around $\mathbf{B}_k$ leads to $\log\det(\mathbf{A}) \leq \log\det(\mathbf{B}_k) + \langle(\mathbf{B}_k)^{-1}, \mathbf{A}-\mathbf{B}_k\rangle$. Hence,

$$f(\mathbf{W},\mathbf{H}_k) \leq \widetilde{f}_{\mathbf{W}_k}(\mathbf{W}) := \frac{1}{2}\|\mathbf{M}-\mathbf{W}\mathbf{H}_k\|_F^2 + \frac{\lambda}{2}\langle\mathbf{P}_k, \mathbf{W}^\top\mathbf{W}\rangle + C_1, \quad (5)$$

where C_1 is a constant independent of $\mathbf{W}$. Note that the gradient of $\mathbf{W} \mapsto \widetilde{f}_{\mathbf{W}_k}(\mathbf{W})$, being equal to

$$(\mathbf{W}\mathbf{H}_k-\mathbf{M})\mathbf{H}_k^\top + \lambda\mathbf{W}\mathbf{P}_k,$$

is $L_{\mathbf{W}}^k$-Lipschitz continuous with $L_{\mathbf{W}}^k = \|\mathbf{H}_k\mathbf{H}_k^\top + \lambda\mathbf{P}_k\|_2$. Hence, from (5) and the descent lemma (see [15, Section 2.1]),

$$f(\mathbf{W},\mathbf{H}_k) \leq u_{\mathbf{W}_k}(\mathbf{W}) := \langle\nabla\widetilde{f}_{\mathbf{W}_k}(\mathbf{W}_k), \mathbf{W}\rangle + \frac{L_{\mathbf{W}}^k}{2}\|\mathbf{W}-\mathbf{W}_k\|_F^2 + C_2, \quad (6)$$

where C_2 is a constant depending on $\mathbf{W}_k$. We use the surrogate $u_{\mathbf{W}_k}(\mathbf{W})$ defined in (6) to update $\mathbf{W}_k$. As TITAN recovers Nesterov-type acceleration for the update of each block of variables [9, Section 4.2], we have the following update for $\mathbf{W}$:

$$\begin{aligned}\mathbf{W}_{k+1} &= \operatorname*{argmin}_{\mathbf{W}\in\mathcal{X}_{\mathbf{W}}}\langle\nabla\widetilde{f}_{\mathbf{W}_k}(\overline{\mathbf{W}}_k), \mathbf{W}\rangle + \frac{L_{\mathbf{W}}^k}{2}\|\mathbf{W}-\overline{\mathbf{W}}_k\|_F^2,\\ &= \mathscr{P}\left(\overline{\mathbf{W}}_k + \frac{(\mathbf{M}-\overline{\mathbf{W}}_k\mathbf{H}_k)\mathbf{H}_k^\top - \lambda\overline{\mathbf{W}}_k\mathbf{P}}{L_{\mathbf{W}}^k}\right),\end{aligned} \quad (7)$$

where $\mathscr{P}$ performs column wise projections onto the unit simplex as in [16] in order to satisfy the constraint on $\mathbf{W}$ in (2), and where $\overline{\mathbf{W}}_k$ is an extrapolated point, that is, the current point $\mathbf{W}_k$ plus some momentum,

$$\overline{\mathbf{W}}_k = \mathbf{W}_k + \beta_{\mathbf{W}}^k(\mathbf{W}_k-\mathbf{W}_{k-1}), \quad (8)$$

where the extrapolation parameter $\beta_{\mathbf{W}}^k$ is chosen as follows

$$\beta_{\mathbf{W}}^k = \min\left[\frac{\alpha_k-1}{\alpha_{k+1}}, 0.9999\sqrt{\frac{L_{\mathbf{W}}^{k-1}}{L_{\mathbf{W}}^k}}\right], \quad (9)$$

$\alpha_0 = 1$, $\alpha_k = (1+\sqrt{1+4\alpha_{k-1}^2})/2$. This choice of parameter satisfies the conditions to have a subsequential convergence of TITAN, see Section III-C.

2) Surrogate function and update of $\mathbf{H}$*:* Since

$$\nabla_{\mathbf{H}} f(\mathbf{W}_{k+1},\mathbf{H}) = \mathbf{W}_{k+1}^\top(\mathbf{W}_{k+1}\mathbf{H}-\mathbf{M}),$$

the gradient of f according to $\mathbf{H}$ is $L_{\mathbf{H}}^k$-Lipschitz continuous with $L_{\mathbf{H}}^k = \|\mathbf{W}_{k+1}^\top\mathbf{W}_{k+1}\|_2$. Hence, we use the following Lipschitz gradient surrogate to update $\mathbf{H}_k$:

$$u_{\mathbf{H}_k}(\mathbf{H}) = \langle\nabla_{\mathbf{H}} f(\mathbf{W}_{k+1},\mathbf{H}_k),\mathbf{H}\rangle + \frac{L_{\mathbf{H}}^k}{2}\|\mathbf{H}-\mathbf{H}_k\|_F^2 + C_3, \quad (10)$$

where C_3 is a constant depending on $\mathbf{H}_k$. We derive our update rule for $\mathbf{H}$ by minimizing the surrogate function from Equation (10) embedded with extrapolation,

$$\begin{aligned}\mathbf{H}_{k+1} &= \operatorname*{argmin}_{\mathbf{H}\in\mathcal{X}_H}\langle\nabla_{\mathbf{H}} f(\mathbf{W}_{k+1},\overline{\mathbf{H}}_k),\mathbf{H}\rangle + \frac{L_{\mathbf{H}}^k}{2}\|\mathbf{H}-\overline{\mathbf{H}}_k\|_F^2,\\ &= \left[\overline{\mathbf{H}}_k + \frac{1}{L_{\mathbf{H}}^k}\mathbf{W}_{k+1}^\top(\mathbf{M}-\mathbf{W}_{k+1}\overline{\mathbf{H}}_k)\right]_+,\end{aligned} \quad (11)$$

where $[\,.\,]_+$ denotes the projector setting all negative values to zero, and $\overline{\mathbf{H}}_k$ is the extrapolated $\mathbf{H}_k$:

$$\overline{\mathbf{H}}_k = \mathbf{H}_k + \beta_{\mathbf{H}}^k(\mathbf{H}_k-\mathbf{H}_{k-1}), \quad (12)$$

where, as for the update of $\mathbf{W}$,

$$\beta_{\mathbf{H}}^k = \min\left[\frac{\alpha_k-1}{\alpha_{k+1}}, 0.9999\sqrt{\frac{L_{\mathbf{H}}^{k-1}}{L_{\mathbf{H}}^k}}\right]. \quad (13)$$

B. Algorithm

Note that the update of $\mathbf{W}$ in (7) and $\mathbf{H}$ in (11) was described when the cyclic update rule is applied. Since TITAN also allows the essentially cyclic rule [9, Section 5], we can update $\mathbf{W}$ several times before switching updating $\mathbf{H}$, and vice versa. This leads to our proposed method TITANized min-vol, see Algorithm 1 for the pseudo code. The stopping criteria in lines 4 and 15 is the same as in [11]. The way λ and δ are computed is also identical to [11]. Let us mention that technically the main difference with [11] resides in how the extrapolation is embedded. In [11] the Nesterov sequence is restarted and evolves in each inner loop to solve each subproblem corresponding to each block. In our algorithm, the extrapolation parameter $\beta_{\mathbf{W}}$ (and $\beta_{\mathbf{H}}$) for updating each block $\mathbf{W}$ (and $\mathbf{H}$) is updated continuously without restarting. It means we are accelerating the global convergence of the sequences rather than trying to accelerate the convergence for the subproblems. Moreover, TITAN allows to update the surrogate function at each step, while the algorithm from [11] can only update it before each subproblem is solved, as it relies on Nesterov's acceleration for convex optimization.

C. Convergence guarantee

In order to have a convergence guarantee, TITAN requires the update of each block to satisfy the nearly sufficiently decreasing property (NSDP), see [9, Section 2]. By [9, Section 4.2.1], the update for $\mathbf{H}$ of TITANized min-vol satisfies the NSDP condition since it uses a Lipschitz gradient surrogate for $\mathbf{H} \mapsto f(\mathbf{W},\mathbf{H})$ combined with the Nesterov-type extrapolation; and the bounds of the extrapolation parameters in the update of $\mathbf{H}$ are derived similarly as in [9, Section 6.1]. However, it is important noting that the update for $\mathbf{W}$ of TITANized min-vol does not directly use a Lipschitz gradient surrogate for $\mathbf{W} \mapsto f(\mathbf{W},\mathbf{H})$. We thus need to verify NSDP condition for the update of $\mathbf{W}$ by another method that is presented in the following.

The function $u_{\mathbf{W}_k}(\mathbf{W})$ is a Lipschitz gradient surrogate of $\tilde{f}_{\mathbf{W}_k}(\mathbf{W})$ and we apply the Nesterov-type extrapolation to

Algorithm 1 TITANized min-vol

initialize $\mathbf{W}_0$ and $\mathbf{H}_0$,
$\alpha_1 = 1$, $\alpha_2 = 1$, $\mathbf{W}_{old} = \mathbf{W}_0, \mathbf{H}_{old} = \mathbf{H}_0$, $L_{\mathbf{W}}^{prev} = \|\mathbf{H}_0\mathbf{H}_0^\top + \lambda(\mathbf{W}_0^\top\mathbf{W}_0 + \delta\mathbf{I}_r)^{-1}\|_2$, $L_{\mathbf{H}}^{prev} = \|\mathbf{W}_0^\top\mathbf{W}_0\|_2$
repeat
  while stopping criteria not satisfied **do**
    $\alpha_0 = \alpha_1, \alpha_1 = (1+\sqrt{1+4\alpha_0^2})/2$
    $\mathbf{P} \leftarrow (\mathbf{W}^\top\mathbf{W} + \delta\mathbf{I}_r)^{-1}$
    $L_{\mathbf{W}} \leftarrow \|\mathbf{H}\mathbf{H}^\top + \lambda\mathbf{P}\|_2$
    $\beta_{\mathbf{W}} = \min\left[\ (\alpha_0-1)/\alpha_1, 0.9999\sqrt{L_{\mathbf{W}}^{prev}/L_{\mathbf{W}}}\ \right]$
    $\overline{\mathbf{W}} \leftarrow \mathbf{W} + \beta_{\mathbf{W}}(\mathbf{W} - \mathbf{W}_{old})$
    $\mathbf{W}_{old} \leftarrow \mathbf{W}$
    $\mathbf{W} \leftarrow \mathscr{P}\left[\overline{\mathbf{W}} + \frac{(\mathbf{M}\mathbf{H}^\top - \overline{\mathbf{W}}(\mathbf{H}\mathbf{H}^\top + \lambda\mathbf{P}))}{L_{\mathbf{W}}}\right]$
    $L_{\mathbf{W}}^{prev} \leftarrow L_{\mathbf{W}}$
  end while
  $L_{\mathbf{H}} \leftarrow \|\mathbf{W}^\top\mathbf{W}\|_2$
  while stopping criteria not satisfied **do**
    $\alpha_0 = \alpha_2, \alpha_2 = (1+\sqrt{1+4\alpha_0^2})/2$
    $\beta_{\mathbf{H}} = \min\left[\ (\alpha_0-1)/\alpha_2, 0.9999\sqrt{L_{\mathbf{H}}^{prev}/L_{\mathbf{H}}}\ \right]$
    $\overline{\mathbf{H}} \leftarrow \mathbf{H} + \beta_{\mathbf{H}}(\mathbf{H} - \mathbf{H}_{old})$
    $\mathbf{H}_{old} \leftarrow \mathbf{H}$
    $\mathbf{H} \leftarrow \left[\overline{\mathbf{H}} + \frac{\mathbf{W}^\top(\mathbf{M} - \mathbf{W}\overline{\mathbf{H}})}{L_{\mathbf{H}}}\right]_+$
    $L_{\mathbf{H}}^{prev} \leftarrow L_{\mathbf{H}}$
  end while
until some stopping criteria is satisfied

obtain the update in (7). Note that the feasible set of $\mathbf{W}$ is convex. Hence, it follows from [9, Remark 4.1] that

$$\tilde{f}_{\mathbf{W}_k}(\mathbf{W}_k) + \frac{L_{\mathbf{W}}^k(\beta_{\mathbf{W}}^k)^2}{2}\|\mathbf{W}_k - \mathbf{W}_{k-1}\|_F^2 \geq \tilde{f}_{\mathbf{W}_k}(\mathbf{W}_{k+1}) + \frac{L_{\mathbf{W}}^k}{2}\|\mathbf{W}_{k+1} - \mathbf{W}_k\|_F^2. \quad (14)$$

Furthermore, we note that $\tilde{f}_{\mathbf{W}_k}(\mathbf{W}_k) = f(\mathbf{W}_k, \mathbf{H}_k)$, and $\tilde{f}_{\mathbf{W}_k}(\mathbf{W}_{k+1}) \geq f(\mathbf{W}_{k+1}, \mathbf{H}_k)$. Therefore, from (14) we have

$$f(\mathbf{W}_k, \mathbf{H}_k) + \frac{L_{\mathbf{W}}^k(\beta_{\mathbf{W}}^k)^2}{2}\|\mathbf{W}_k - \mathbf{W}_{k-1}\|_F^2 \geq f(\mathbf{W}_{k+1}, \mathbf{H}_k) + \frac{L_{\mathbf{W}}^k}{2}\|\mathbf{W}_{k+1} - \mathbf{W}_k\|_F^2,$$

which is the required NSDP condition of TITAN. Consequently, the choice of $\beta_{\mathbf{W}}^k$ in (9) satisfy the required condition to guarantee subsequential convergence [9, Proposition 3.1].

On the other hand, we note that the error function $\mathbf{W} \mapsto e_1(\mathbf{W}) := u_{\mathbf{W}_k}(\mathbf{W}) - f(\mathbf{W}, \mathbf{H}_k)$ is continuously differentiable and $\nabla_{\mathbf{W}} e_1(\mathbf{W}_k) = \mathbf{0}$; similarly for the error function $\mathbf{H} \mapsto e_2(\mathbf{H}) := u_{\mathbf{H}_k}(\mathbf{H}) - f(\mathbf{W}_{k+1}, \mathbf{H})$. Hence, it follows from [9, Lemma 2.3] that the Assumption 2.2 in [9] is satisfied. Applying [9, Theorem 3.2], we conclude that every limit point of the generated sequence is a stationary point of Problem (2). It is worth noting that as TITANized min-vol does not apply restarting step, [9, Theorem 3.5] for a global convergence is not applicable.

IV. Numerical Experiments

In this section we compare TITANized min-vol to [11], an accelerated version of the method from [8] (for $p = 2$), on two NMF applications: hyperspectral unmixing and document clustering, which are dense and sparse data sets, respectively. All tests are performed on MATLAB R2018a, on a PC with an Intel® Core™ i7 6700HQ and 24GB RAM. The code is available from https://github.com/vuthanho/titanized-minvol.

The data sets used are shown in Table I. For each data set, each algorithm is launched with the same random initializations, for the same amount of CPU time. In order to derive some statistics, for both hyperspectral unmixing and document clustering, 20 random initializations are used (each entry of $\mathbf{W}$ and $\mathbf{H}$ are drawn from the uniform distribution in [0,1]). The CPU time used for each data set is adjusted manually, and corresponds to the maximum displayed value on the respective time axes in Fig. 1; see also Table II.

data set	m	n	r
Urban	162	94249	6
Indian Pine	200	21025	16
Pavia Univ.	103	207400	9
San Diego	158	160000	7
Terrain	166	153500	5
20 News	61188	7505	20
Sports	14870	8580	7
Reviews	18483	4069	5

TABLE I: data sets used in our experiments and their respective dimensions

For display purposes, for each data set, we compare the average of the scaled objective functions according to time, that is, the average of $(f(\mathbf{W}, \mathbf{H}) - e_{\min})/\|\mathbf{M}\|_F$ where $e_{\min}$ is the minimum obtained error among the 20 different runs and among both methods. The results are presented in Fig. 1. On both hyperspectral and document data sets, TITANized min-vol converges on average faster than [11] except for the San Diego data set (although TITANized min-vol converges initially faster). For most tested data sets, min-vol [11] cannot reach the same error as TITANized min-vol within the allocated time. In particular, TITANized min-vol achieves a lower error in 94 out of the 100 runs for the hyperspectral images (5 images with 20 random initialization each), and 55 out of 60 for the document data sets (3 sets of documents with 20 random initialization each).

We also reported in Table II TITANized min-vol's lead time over [11] when the latter reaches its minimum error after the maximum allotted CPU time. The lead time is the time saved by TITANized min-vol to achieve the error of the method from [11] using the maximum allotted CPU time. On average, TITANized min-vol is twice faster than [11], with an average gain of CPU time above 50%.

To summarize, our experimental results show that TITANized min-vol has a faster convergence speed and smaller final solutions than [11].

data set	Our method's lead time (s)	CPU time for [11]	Saved CPU time
Urban	44	60	73%
Indian Pines	25	30	83%
Pavia Univ.	68	90	76%
San Diego	NaN	120	0%
Terrain	44	60	73%
20News	221	300	74%
Reviews	26	30	80%
Sports	15	30	50%

TABLE II: TITANized min-vol's lead time over min-vol [11] to obtain the same minimum error.

V. Conclusion and Discussion

We developed a new algorithm to solve min-vol NMF (2) based on the inertial block majorization-minimization framework of [9]. This framework, under some conditions that hold for our method, guarantees subsequential convergence. Experimental results show that this acceleration strategy performs better than the state-of-the-art accelerated min-vol NMF algorithm from [11]. Future works will focus on different types of acceleration such as Anderson's acceleration [17], and on different constraints on $\mathbf{W}$ and/or $\mathbf{H}$ to address some specific applications.

References

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

[2] N. Gillis, *Nonnegative Matrix Factorization*. SIAM, 2020. [Online]. Available: https://doi.org/10.1137/1.9781611976410

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

[4] 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, pp. 59–80, 2019.

[5] 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, pp. 2306–2320, 2015.

[6] 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.

[7] V. Leplat, N. Gillis, and M. S. Ang, "Blind audio source separation with minimum-volume beta-divergence NMF," *IEEE Trans. Signal Process.*, vol. 68, pp. 3400–3410, 2020.

[8] X. Fu, K. Huang, B. Yang, W.-K. Ma, and N. D. Sidiropoulos, "Robust volume minimization-based matrix factorization for remote sensing and document clustering," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6254–6268, 2016.

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

[10] M. D. Craig, "Minimum-volume transforms for remotely sensed data," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 3, pp. 542–552, 1994.

[11] V. Leplat, A. M. S. Ang, and N. Gillis, "Minimum-volume rank-deficient nonnegative matrix factorizations," in *ICASSP*, 2019, pp. 3402–3406.

[12] 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.

[13] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 5, no. 2, pp. 354–379, 2012.

[14] Y. Xu and W. Yin, "A globally convergent algorithm for nonconvex optimization based on block coordinate update," *Journal of Scientific Computing*, vol. 72, no. 2, pp. 700–734, Aug 2017.

[15] Y. Nesterov, *Lectures on Convex Optimization*, 2nd ed. Springer Publishing Company, Incorporated, 2018.

[16] L. Condat, "Fast projection onto the simplex and the ℓ_1 ball," *Mathematical Programming*, vol. 158, no. 1, pp. 575–585, 2016.

[17] D. G. Anderson, "Iterative procedures for nonlinear integral equations," *Journal of the ACM (JACM)*, vol. 12, no. 4, pp. 547–560, 1965.

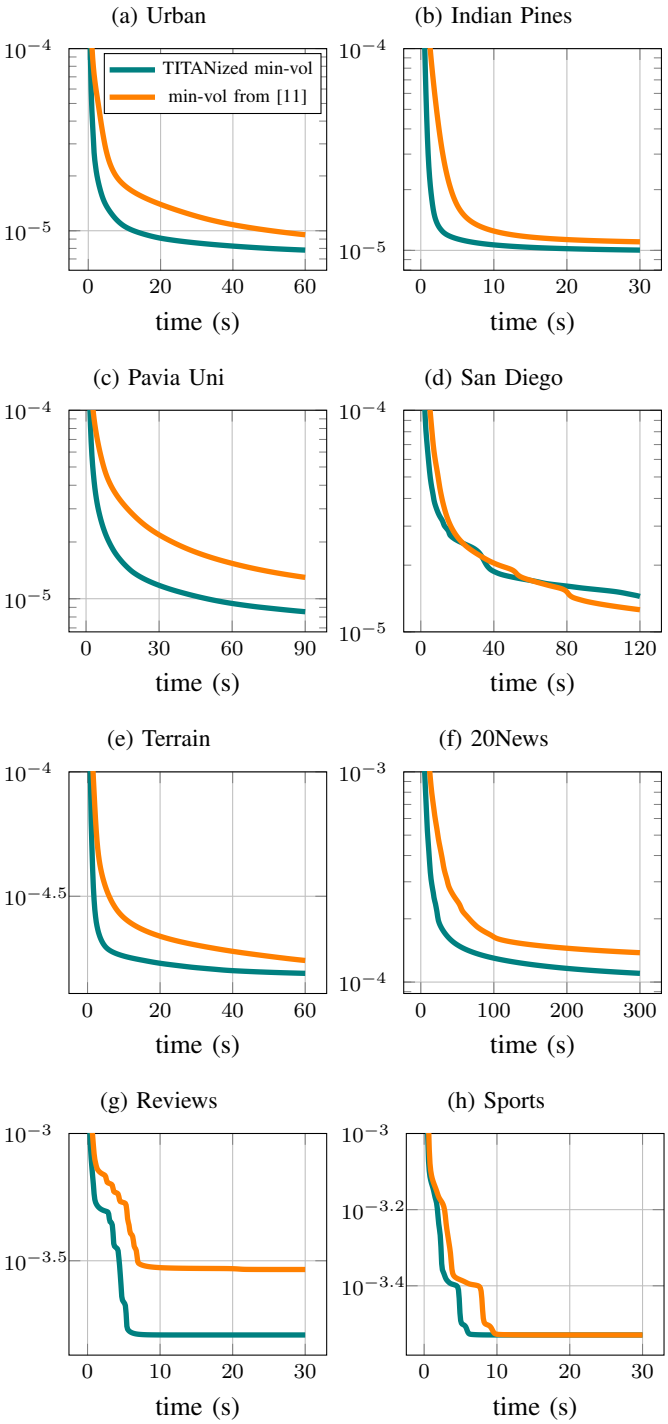

Fig. 1: Evolution w.r.t. time of the average of $(f(\mathbf{W},\mathbf{H}) - e_{\min})/\|\mathbf{M}\|_F$ for the different data sets.