SIAM J. Imaging Sciences
Vol. 17, No. 4, pp. 2362–2391

Dual Simplex Volume Maximization for Simplex-Structured Matrix Factorization*

Maryam Abdolali†, Giovanni Barbarino‡, and Nicolas Gillis‡

Abstract. Simplex-structured matrix factorization (SSMF), closely related to nonnegative matrix factorization, is a fundamental interpretable data analysis model and has applications in hyperspectral unmixing and topic modeling. To obtain identifiable solutions, a standard approach is to find minimum-volume solutions. By taking advantage of the duality/polarity concept for polytopes, we convert minimum-volume SSMF in the primal space to a maximum-volume problem in the dual space. We first prove the identifiability of this maximum-volume dual problem. Then, we use this dual formulation to provide a novel optimization approach which bridges the gap between two existing families of algorithms for SSMF, namely volume minimization and facet identification. Numerical experiments show that the proposed approach performs favorably compared to the state-of-the-art SSMF algorithms.

Key words. simplex-structured matrix factorization, nonnegative matrix factorization, minimum volume, sparsity, polarity/duality, hyperspectral imaging

MSC codes. 15A23, 65F55, 68Q25, 65D18

DOI. 10.1137/24M1650600

1. Introduction.

Matrix factorization (MF) is a fundamental technique for extracting latent low-dimensional factors, with applications in numerous fields, such as data analysis, machine learning, and signal processing. MF aims to decompose a given data matrix, $X \in \mathbb{R}^{m\times n}$, where the n columns represent m-dimensional samples, into the product of two smaller matrices, $W \in \mathbb{R}^{m\times r}$ and $H \in \mathbb{R}^{r\times n}$, called factors, such that $X \approx WH$. Often imposing additional constraints, such as sparsity or nonnegativity, on the factors is crucial, e.g., for interpretation purposes, leading to structured (or constrained) matrix factorization (SMF); see, e.g., [32, 15] and the references therein. A specific problem of the broad family of SMF assumes that each column of H belongs to the unit simplex, that is, for all j,

$$H(:,j) \in \Delta^r := \left\{ x \in \mathbb{R}^r \;\middle|\; x \geq 0, e^\top x = \sum_{i=1}^r x_i = 1 \right\},$$

where e is the vector of all ones of appropriate dimension. Simplex-structured matrix factorization (SSMF) has several applications in machine learning with two prominent examples:

*Received by the editors March 29, 2024; accepted for publication (in revised form) August 13, 2024; published electronically December 10, 2024.
https://doi.org/10.1137/24M1650600
Funding: The work of the second and third authors was supported by the European Union (ERC consolidator, eLinoR) grant 101085607.

†K. N. Toosi University (KNTU), Tehran, Iran (maryam.abdolali@kntu.ac.ir).
‡University of Mons, Mons, Belgium. GB is a member of the Research Group GNCS ruppo Nazionale per il Calcolo Scientificoof INdAM stituto Nazionale di Alta Matematica (giovanni.barbarino@umons.ac.be, https://giovannibarbarino.github.io/, nicolas.gillis@umons.ac.be, https://sites.google.com/site/nicolasgillis).

unmixing hyperspectral images where $H(i,j)$ is the proportion/abundance of the ith material within the jth pixel [20, 6, 28], and topic modeling where $H(i,j)$ is the contribution of the ith topic within the jth document [3, 13, 5].

Contribution and outline of the paper. This paper focuses on the concept of duality and uses the correspondence between primal and dual spaces to provide a new perspective on fitting a simplex to the samples. The main contributions are as follows:

- We present a new formulation for SSMF which is based on the concept of duality. This formulation provides a different perspective on SSMF and bridges the gap between two existing families of approaches: volume minimization and facet-based identification (section 3).
- We study the identifiability of the parameters with this new formulation (section 4).
- We develop an efficient optimization scheme based on block coordinate descent (section 5).
- We provide numerous numerical experiments on both synthetic and real-world data sets, showing that the proposed algorithm competes favorably with the state of the art (section 6).

2. Previous works. In this paper, we consider the following SSMF formulation: Given $X \in \mathbb{R}^{m\times n}$ and $r > 0$, solve

$$\min_{W\in\mathbb{R}^{m\times r}, H\in\mathbb{R}^{r\times n}} \|X - WH\|_F^2 \quad \text{such that} \quad H(:,j) \in \Delta^r \text{ for all } j.$$

SSMF is closely related to nonnegative matrix factorization (NMF), which decomposes a nonnegative matrix, $X \geq 0$, as $X \approx WH$ where $W \geq 0$ and $H \geq 0$ [23, 17]. In fact, when one is looking for an exact decomposition, that is, $X = WH$, SSMF generalizes NMF: Given an NMF $X = WH \geq 0$, $W \geq 0$ and $H \geq 0$, normalizing each column of X to have unit ℓ_1 norm, and assuming that the columns of W also have unit ℓ_1 norm (this can be done without loss of generality by using the scaling degree of freedom between the columns of W and rows of H), implies that the columns of $H \geq 0$ also have ℓ_1 norm, since $e^\top = e^\top X = e^\top WH = e^\top H$, and hence H is column stochastic. In other words, any exact NMF can be written as an SSMF.

Geometric interpretation of SSMF and uniqueness/identifiability. For an exact SSMF decomposition, we have

$$X(:,j) = WH(:,j) = \sum_{k=1}^{r} W(:,k)H(k,j),$$

meaning that the columns of X are convex combinations of the column of W. In other words, SSMF aims to find r vectors, $\{W(:,k)\}_{k=1}^r$, such that their convex hull contains the columns of X, that is, for all j

$$X(:,j) \quad \in \quad \text{conv}(W) = \{x \mid x = Wh, h \in \Delta^r\}.$$

We will say that $X = WH$ has a unique SSMF if any other SSMF of X, say, $X = W'H'$, can only be obtained by permutation of the columns of W and rows of H, that is, $X = W'H'$ implies that $W'(:,k) = W(:,\pi_k)$ and $H'(k,:) = H(\pi_k,:)$ for some permutation π of $\{1,2,\ldots,r\}$. Without any further constraints, SSMF is never unique, because we can always enlarge the convex

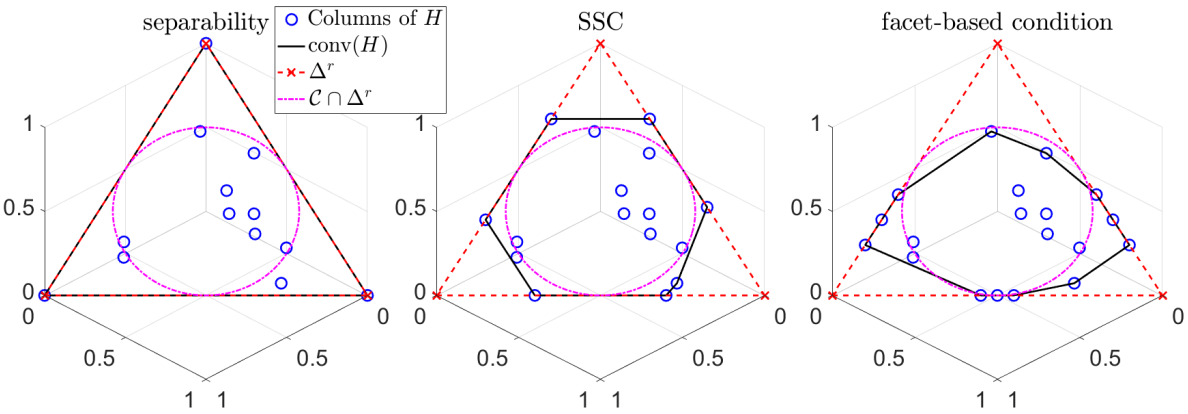


Figure 1. *Comparison of separability (left), the sufficiently scattered condition (SSC) (middle), and the facet-based condition (right) for the matrix H whose columns lie on Δ^r in the case $r = 3$. On the left, separability requires the columns of H to contain the unit vectors, that is, $H(:, \mathcal{K}) = I_r$ for some $\mathcal{K}$. In the middle, the SSC requires $\mathcal{C} \subset \text{cone}(H)$. On the right, the facet-based condition requires $r = 3$ columns of H on each facet of the unit simplex. Figure from [1].*

hull of W to contain more points, and hence obtain equivalent factorizations [18]. It is therefore crucial for SSMF models to include additional constraints or regularizers to obtain identifiable models. There have been three main approaches to achieve this goal: separability, volume minimization, and facet-based identification. They are described in the next three sections.

2.1. Separability. Separability assumes that the columns of W are among the columns of X [4, 6], that is, the matrix X admits an SSMF of the form $X = WH$ with $W = X(:, \mathcal{K})$ and the index set $\mathcal{K}$ contains r elements, that is, $|\mathcal{K}| = r$. Equivalently, $X = WH$ with $H(:, \mathcal{K}) = I_r$ for some index set $\mathcal{K}$, where I_r the identity matrix of dimension r. See Figure 1 (left) for an illustration. In other words, separability requires that for each basis vector, $W(:, k)$, there exists a data point, $X(:, \mathcal{K}_k)$, such that $W(:, k) = X(:, \mathcal{K}_k)$. This is the so-called pure-pixel assumption in hyperspectral unmixing [7], and the anchor-word assumption in topic modeling [3].

Separability leads to identifiability and simplifies the problem resulting in polynomial-time algorithms, some running in $O(mnr)$ operations, with theoretical guarantees; see [17, Chapter 7] for a comprehensive survey on these algorithms. However, separability is a strong assumption which might not hold in all real-world scenarios.

2.2. Volume minimization. In order to relax separability, one can look for an SSMF, $X = WH$, where the volume of the convex hull of the columns of W is minimized. The first intuitions and empirical evidence came from the hyperspectral imaging literature [11, 29]. Later, minimum-volume SSMF was shown to be identifiable [14, 26] under the SSC[1] introduced in [22].

Definition 2.1 (SSC). *The matrix $H \in \mathbb{R}^{r \times n}$ satisfies the SSC if*

- *its conic hull, defined by $\text{cone}(H) = \{y \mid y = Hx, x \geq 0\}$, contains the second-order cone $\mathcal{C} = \{x \in \mathbb{R}^r_+ \mid e^\top x \geq \sqrt{r-1}||x||_2\}$, and*

[1]There exist several definitions of the SSC, with minor variations. The main condition, $\mathcal{C} \subseteq \text{cone}(H)$, is always required.

- *any real orthogonal matrix* $Q \in \mathbb{R}^{r\times r}$ *satisfying* $\text{cone}(H) \subset \text{cone}(Q)$ *must be a permutation matrix.*

Intuitively, this assumption implies that the columns of the matrix H are well scattered in the unit simplex Δ^r. For example, the SSC implies that there are at least $r-1$ columns of H on each facet of Δ^r, meaning that H has at least $r-1$ zeros per row. See Figure 1 for an illustration, and see [12] and [17, Chapter 4] for more details.

Many algorithms have been designed using volume minimization, starting from [11]. A common approach is to minimize the volume of enclosing vertices W by minimizing the determinant of $W^\top W$ [14, 26]:

$$\min_{W,H} \det(W^\top W) \qquad \text{such that} \qquad X = WH \text{ and } H(:,j) \in \Delta^r \text{ for all } j. \tag{2.1}$$

The rationale behind this model is that $\frac{1}{r!}\sqrt{\det(W^\top W)}$ is equal to the volume of the convex hull of W and the origin, when restricted to the column space of W. Under the SSC, solving (2.1) guarantees to recover the columns of W and H in the SSMF $X = WH$, up to permutation [14, 26]. In the presence of noise, one has to balance the data fitting term and the volume regularizer by minimizing $\|X - WH\|_F^2 + \lambda \det(W^\top W)$ for some well-chosen penalty parameter $\lambda > 0$. Another approach is minimum-volume enclosing simplex (MVES) [10], which attempts to simplify the problem by focusing on the volume of a dimension-reduced transformation of W (via singular value decomposition (SVD)), say, $\bar{W} \in \mathbb{R}^{r\times r-1}$; see section 3.1 for details. MVES works with the transformed matrix

$$\tilde{W} = [\bar{W}(:,1) - \bar{W}(:,r), \ldots, \bar{W}(:,r-1) - \bar{W}(:,r)] \in \mathbb{R}^{(r-1)\times(r-1)} \tag{2.2}$$

and minimizes $|\det(\tilde{W})| = \text{vol}(\text{conv}(\bar{W}))$. This reformulation allows them to solve the subproblem in each column of W via alternating linear optimization. More recently, a more general class of problems was considered in [31], where the columns of H are restricted to belonging to a polytope, which is referred to as polytopic matrix factorization. Instead of minimizing the volume of $\text{conv}(W)$, the determinant of $HH^\top$ is maximized, with identifiability guarantees under a generalized SSC.

In contrast to separable-based algorithms, volume-minimization problems, such as (2.1), are not convex, and hence it is not straightforward to solve them up to global optimality. Hence although volume minimization allows one to theoretically identify SSMF under relaxed conditions, it makes the optimization problems harder to solve than under the separability assumption. Also, robustness to noise is not well understood.

2.3. Facet-based identification. Instead of looking for columns of W whose convex hull contains the columns of X, one can instead look for a set of facets (a facet is an affine hyperplane $\{x \mid a^\top x = b\}$ delimiting the associated half space $\{x \mid a^\top x \le b\}$ for some vector a and scalar b) enclosing a region where the columns of X lie. Two main algorithms in this category are maximum-volume inscribed ellipsoid (MVIE) [27] and greedy facet-based polytope identification (GFPI) [1]:

- MVIE [27] identifies the enclosing r facets by a two-step approach: (i) generate all facets of $\text{conv}(X)$, and (ii) find the maximum-volume ellipsoid inscribed in the generated facets. Under the SSC, this ellipsoid touches every facet of $\text{conv}(W)$ which leads

to the identification of r facets of the simplex and, subsequently, the vertices in W. Although MVIE is guaranteed to recover W in the noiseless case under the SSC, it relies on the computationally expensive algorithm of facet enumeration limiting the algorithm values of r up to around 10 and is sensitive to noise and outliers.

- GFPI [1] uses duality to map the facet identification problem in the primal space into the corresponding vertex identification problem in the dual space. Using duality, GFPI prioritizes facets with most samples on them. GFPI formulates this problem as a mixed integer program which identifies the facets sequentially. GFPI has several significant advantages over other approaches, including the ability to handle rank-deficient matrices, outliers, and input data that violates the SSC. Moreover, it is identifiable under a typically weaker condition than the SSC, namely the facet-based condition which requires r data points on each facet of $\text{conv}(W)$ (and some other minor conditions generically satisfied); see Figure 1 for an illustration and [1] for more details.

3. Proposed model: SSMF based on maximum volume in the polar. In this section, we introduce our novel approach which relies on facet-based identification. It is based on a novel efficient vertex enumeration in the dual space. In contrast to GFPI, the proposed approach does not rely on greedy sequential identification of vertices (corresponding facets in the primal space) but identifies the facets simultaneously by maximizing their volume in the dual space.

Our proposed approach is based on duality/polarity (we use both words interchangeably in this paper). In order to recover the columns of W, which are the vertices of the simplex enclosing the columns of X, we focus on extracting the facets of its convex hull. The facets are implicitly obtained by calculating the vertices of the corresponding dual simplex. Before explaining this in section 3.2, we first reduce the dimension to work with full-dimensional polytopes. This reduction requires $\text{conv}(X)$ to have dimension $r-1$, which requires the dimension of its affine hull to be $r-1$, which we will assume throughout the paper.

Assumption 3.1. The affine hull of X, $\{y \mid y = Xh \text{ where } e^\top h = 1\}$, has dimension $r-1$.

If $\text{rank}(H) = r$, which is implied by the SSC, and if the affine hull of W has dimension $r-1$, then the affine hull of X has dimension $r-1$. Note that $\text{rank}(W) = r$ implies that the affine hull of W has dimension $r-1$. However, we could also have the case $\text{rank}(W) = r-1$ if 0 belongs to the affine hull of W (e.g., in two dimensions, $\text{conv}(W)$ is a triangle containing the origin).

3.1. Preprocessing: Translation and dimensionality reduction. In this paper, like in many other SSMF approaches, e.g., MVES [10] and GFPI [1] (see also [28]), we will use a preprocessing of the data to reduce it to an $(r-1)$-dimensional space. This has several motivations:

- The convex hull of the columns of W, $\text{conv}(W)$, is an $r-1$ dimensional simplex, under Assumption 3.1.
- In the noiseless case, the preprocessing does not change the geometry and properties of the problem. In the presence of noise, it allows one to reduce the noise level. We will use the truncated SVD which is suitable for Gaussian noise (it corresponds to the maximum likelihood estimator); other techniques could be used in the presence of outliers and non-Gaussian noise.

- The notion of polarity is simpler to grasp for full-dimensional polytopes: the polar of $\text{conv}(W)$ will also be an $(r-1)$-dimensional polytope.

The preprocessing has two steps.

Step 1: Translation around the origin. Let us choose a point, v, in the relative interior of $\text{conv}(X)$. For example, one can choose the sample mean, $v = \bar{x} = \frac{1}{n}\sum_{j=1}^{n} X(:,j)$. We will discuss in section 4 the importance of this choice, which will need to be part of the optimization problem to obtain identifiability. For example, if v is chosen too close to the border of $\text{conv}(W)$, the polar of $\text{conv}(W)$ in the dual space (which will be defined below) will have a large volume and be ill-conditioned. As we will see, the optimal choice is to take the center of the columns of W, that is, $v = \bar{w} = \frac{1}{r}\sum_{k=1}^{r} W(:,k)$, but W is unknown and needs to be estimated.

The first step for preprocessing the data is to remove v from each sample to obtain $\hat{X} = X - ve^\top$. Let $X = WH$ be an SSMF of X where $e^\top H = e^\top$ and $H \geq 0$. This first step simply amounts to translating the SSMF problem, since

$$\hat{X} = X - ve^\top = WH - ve^\top = [W - ve^\top]H = \hat{W}H \text{ with } \hat{W} = W - ve^\top.$$

Since v is in the relative interior of $\text{conv}(X)$, the vector of zeros is in the relative interior of $\text{conv}(\hat{X})$: $v = Xh$ for some $h > 0$ and $e^\top h = 1$ implying

$$0 = Xh - v = [X - ve^\top]h = \hat{X}h.$$

This shows that the column space of $\hat{X}$ has dimension $r-1$, under Assumption 3.1.

Step 2: Dimensionality reduction. The second step is to project the centered samples $\hat{X}$ onto the $(r-1)$-dimensional column space of $\hat{X}$ using the truncated SVD. Let $U\Sigma V^\top$ be the truncated SVD of $\hat{X}$ where $U \in \mathbb{R}^{m\times(r-1)}$, $\Sigma \in \mathbb{R}^{(r-1)\times(r-1)}$, and $V \in \mathbb{R}^{n\times(r-1)}$. The projected samples $Y \in \mathbb{R}^{(r-1)\times n}$ are obtained by $Y = U^\top \hat{X} = \Sigma V^\top$. The second step of the preprocessing simply premultiplies $\hat{X}$ by $U^\top$ to obtain

$$Y = U^\top \hat{X} = (U^\top \hat{W})H = PH \quad \text{with } P = U^\top \hat{W} = U^\top [W - ve^\top] \in \mathbb{R}^{(r-1)\times r}.$$

This is also an equivalent SSMF of smaller dimension, with the same matrix H. In the presence of noise, this preprocessing can help filter the noise. Note that in the presence of non-Gaussian noise, one might project using other norms, that is, not use the SVD which is based on the ℓ_2 norm but low-rank matrix approximations minimizing other norms, e.g., [9, 19].

3.2. Polar representation. We have now transformed the original rank-r SSMF problem of matrix $X \in \mathbb{R}^{m\times n}$ into an equivalent SSMF problem of a rank-$(r-1)$ matrix $Y \in \mathbb{R}^{r-1\times n}$.

Let us show how to construct a polar formulation of this problem. Any feasible solution (P, H) of SSMF for Y satisfies $Y = PH$ where $P \in \mathbb{R}^{r-1\times r}$ and $H(:,j) \in \Delta^r$ for all j. By the geometric interpretation of SSMF (see section 2), $\text{conv}(Y) \subseteq \text{conv}(P)$. Let us define the polar of a set.

Definition 3.2 (polar). *Given any set $\mathcal{S} \subseteq \mathbb{R}^d$, its polar, denoted $\mathcal{S}^*$, is defined as*

$$\mathcal{S}^* := \left\{\theta \in \mathbb{R}^d \,|\, \theta^\top x \leq 1 \;\; \textit{for all } x \in \mathcal{S}\right\}.$$

Polars have many interesting properties [34]:

- If $\mathcal{S}_1 \subseteq \mathcal{S}_2$, then $\mathcal{S}_2^* \subseteq \mathcal{S}_1^*$. Moreover, for any bounded $\mathcal{S}$ its polar $\mathcal{S}^*$ contains the origin in its interior.
- For any invertible matrix $M \in \mathbb{R}^{d\times d}$, $(M\mathcal{S})^* = M^{-\top}\mathcal{S}^*$.
- Suppose that $\mathcal{S}$ is a polytope containing the origin in its interior. If $\mathcal{S}$ has $r \geq d+1$ vertices, then $\mathcal{S}^*$ is a polytope with r facets and vice versa. If $\mathcal{S}$ is also a simplex, that is, an $(r-1)$-dimensional polytope in $\mathbb{R}^{r-1}$ with r vertices and r facets, then $\mathcal{S}^*$ is a simplex.
- For a polytope $\mathcal{S}$ containing the origin in its interior, $(\mathcal{S}^*)^* = \mathcal{S}$.
- The polar of the unit ball is itself, and for any matrix $Q \in \mathbb{R}^{r-1\times r}$ such that $\left[\begin{smallmatrix} Q \\ e^\top/\sqrt{r} \end{smallmatrix}\right]$ is an $r \times r$ orthogonal matrix, the polar of $\operatorname{conv}(Q)$ is $\operatorname{conv}(-rQ)$.

Given a matrix $P \in \mathbb{R}^{(r-1)\times r}$ whose columns define a simplex containing the origin in its interior, the polar of $\operatorname{conv}(P)$ will also be a simplex, by the properties above, and we can collect the vertices of $\operatorname{conv}(P)^*$ in a matrix called the polar matrix of P.

Definition 3.3 (polar matrix of a simplex). *Let $P \in \mathbb{R}^{(r-1)\times r}$ be such that $\operatorname{conv}(P)$ is full-dimensional and contains the origin in its interior. The polar matrix of P, denoted Θ, is defined columnwise as*

$$\Theta(:, j) := P_{\hat{j}}^{-\top} e,$$

where $P_{\hat{j}}$ is the submatrix of P obtained by deleting its jth column. The vectors $\Theta(:, j)$ are all the vertices of $\operatorname{conv}(P)^*$.

Let us come back to SSMF: given Y, we need to find P such that $\operatorname{conv}(Y) \subseteq \operatorname{conv}(P)$. In the polar, we will have $\operatorname{conv}(P)^* \subseteq \operatorname{conv}(Y)^*$, where the vertices of $\operatorname{conv}(P)^*$ are the facets of $\operatorname{conv}(P)$, given that the origin belongs to the interior of $\operatorname{conv}(Y)$. Hence any matrix $P \in \mathbb{R}^{r-1\times r}$ such that $\operatorname{conv}(P)^* \subseteq \operatorname{conv}(Y)^*$ corresponds to a feasible solution of SSMF. Another well-known property of polars is the following: for a matrix Y,

$$\begin{aligned}\operatorname{conv}(Y)^* &= \{\theta \mid y^\top\theta \leq 1 \text{ for all } y = Yh \text{ such that } h \in \Delta^n\} \\ &= \{\theta \mid (Yh)^\top\theta \leq 1 \text{ for all } h \in \Delta^n\} \\ &= \{\theta \mid h^\top(Y^\top\theta) \leq 1 \text{ for all } h \in \Delta^n\} = \{\theta \mid Y^\top\theta \leq e\},\end{aligned}$$

since $h^\top x \leq 1$ for all $h \in \Delta^n$ if and only if $x \leq e$. In fact, $x \leq e$ implies $h^\top x \leq h^\top e = 1$, while taking $h = e_i$ for each i (the unit vectors) implies $x \leq e$. In the following, we will assume that the origin is in the interior of $\operatorname{conv}(P)$. If now Θ is the polar matrix of P, then $\operatorname{conv}(\Theta)^* = \{x \mid \Theta^\top x \leq e\} = \operatorname{conv}(P)$ since the polar of the polar of a polytope is the polytope itself, and the origin is contained in the interior of $\operatorname{conv}(\Theta) = \operatorname{conv}(P)^*$ since $\operatorname{conv}(P)$ is bounded. Given Θ, P can be recovered by computing the vertices of $\operatorname{conv}(\Theta)^* = \{x \mid \Theta^\top x \leq e\} = \operatorname{conv}(P)$, and vice versa.

The constraint $\operatorname{conv}(\Theta) = \operatorname{conv}(P)^* \subseteq \operatorname{conv}(Y)^*$ can therefore be written as $Y^\top\Theta \leq 1_{n\times r}$ where $1_{n\times r}$ is the matrix of all-ones of size n by r. Any matrix $\Theta \in \mathbb{R}^{r-1\times r}$ satisfying $Y^\top\Theta \leq 1_{n\times r}$ and with the origin in the interior of $\operatorname{conv}(\Theta)$ thus corresponds to a feasible solution to SSMF.

This observation was used in [1] to find a P such that as many data points where located on the facets of $\operatorname{conv}(P)$: Let $A = Y^\top\Theta \leq 1_{n\times r}$, then $A(i, j) = 1$ means that the ith data

point, $Y(:,i)$, are located on the jth facet of $\text{conv}(P)$, given by $\{x \mid \Theta(:,j)^\top x = 1\}$, since $A(i,j) = \Theta(:,j)^\top Y(:,i) = 1$. Hence maximizing the number of ones in A maximizes the number of data points on the facets of $\text{conv}(P)$, which has a unique solution (that is, the SSMF is identifiable) under the facet-based condition [1].

3.3. Maximizing the volume in the polar. In this paper, we do not attempt to maximize the number of data points on the facets of $\text{conv}(P)$, which is a combinatorial problem, which was solved in a greedy fashion using mixed-integer programming via the GFPI algorithm of [1]. Instead, we propose to solve the problem at once, maximizing the volume of $\text{conv}(\Theta)$ in the dual space. The rationale behind this choice is that the larger a set is, the smaller its dual is, since $\mathcal{S}_2^* \subseteq \mathcal{S}_1^*$ implies $\mathcal{S}_1 \subseteq \mathcal{S}_2$, and minimizing the volume in the primal has shown to be a powerful approach; see section 2.2. We therefore propose to solve the following model: Given $Y \in \mathbb{R}^{r-1 \times n}$, solve

$$\max_{\Theta \in \mathbb{R}^{r-1 \times r}} \text{vol}(\text{conv}(\Theta)) \quad \text{such that} \quad Y^\top \Theta \leq 1_{n \times r}. \tag{3.1}$$

Recall that the constraint $Y^\top \Theta \leq 1_{n \times r}$ is equivalent to $\text{conv}(\Theta) \subseteq \text{conv}(Y)^*$. The volume can be computed as follows:

$$\text{vol}(\text{conv}(\Theta)) = \frac{1}{(r-1)!} \left| \det \begin{bmatrix} \Theta \\ e^\top \end{bmatrix} \right|.$$

Link with volume minimization. Solving (3.1) is not equivalent to volume minimization in the primal (2.1). In fact, the problem of maximizing the volume of $\text{conv}(\hat{P})^*$ among all the polar sets of the matrices $\hat{P} \in \mathbb{R}^{r-1 \times r}$ such that $\text{conv}(Y) \subseteq \text{conv}(\hat{P})$ is equivalent to

$$\min_{\hat{P} \in \mathbb{R}^{r-1 \times r}, z \in \mathbb{R}^r} \text{vol}(\text{conv}(\hat{P})) \prod_{i=1}^{r} z_i \text{ such that } \text{conv}(Y) \subseteq \text{conv}(\hat{P}), \quad \begin{bmatrix} \hat{P} \\ e^\top \end{bmatrix} z = e_r, \quad z \geq 0. \tag{3.2}$$

For comparison, the original volume minimization in the primal (2.1) can be written as

$$\min_{\hat{P} \in \mathbb{R}^{r-1 \times r}, z \in \mathbb{R}^r} \text{vol}(\text{conv}(\hat{P})) \text{ such that } \text{conv}(Y) \subseteq \text{conv}(\hat{P}).$$

Notice that $\text{conv}(Y) \subseteq \text{conv}(\hat{P})$ can be rewritten as $Y = \hat{P}H$ where $H(:,j) \in \Delta^r$ for all j. We can thus observe that (3.2) differs from (2.1) and will in general give different results. The key difference is the presence of the vector z, representing the barycentric coordinates of the origin with respect to the simplex whose vertices are the columns of $\hat{P}$. Consider, for example, a simplex $\hat{P}$ with a small z_i, implying that the origin is very close to one of the facets of $\hat{P}$. In turn this yields one of the constraints of $\hat{P}$ to be represented in the dual by a vector $\Theta(:,i)$ whose norm is proportional to $1/z_i$ and consequentially very large. This is the rationale linking the volume of Θ and the vector z.

The models based on volume minimization and volume maximization are trying to solve the same problem, namely finding the vertices (the columns of W) of the convex hull of a set of points (the columns of $X = WH$, $H^\top e = e$, $H \geq 0$) where the vertices are not observed

as in separable NMF, but where only points "sufficiently scattered" in that convex hull are observed. In the noiseless case and under the SSC, both models are provably able to recover the ground-truth W. This is actually one of the main contributions of our paper: showing that maximizing the volume in the dual is able to recover W under the SSC; see in particular Theorem 4.1 and Corollary 4.11.

In the presence of noise, it is unclear which model is more suitable in which situation. In the numerical experiments, we will empirically show that maximum volume performs consistently better; see section 6. However, we do no have a theoretical justification at the moment. This is an important open question in the NMF literature: quantifying the robustness to noise of minimum-volume NMF algorithms [12, 17].

In section 5, we will discuss in detail how we solve (3.1), how it can be adapted in the presence of noise, and how we choose the translation point which is crucial to obtain provable guarantees (section 4). But first, we discuss the identifiability guarantees of solving (3.1).

4. Identifiability.

In this section, we prove identifiability of dual volume maximization under various assumptions, namely under the SSC (section 4.1), separability (section 4.2), and a new condition between the two which we call η-expansion (section 4.3). As we will show, the identifiability depends on the choice of the translation vector v, and we provide in section 4.4 a min-max formulation that optimizes the choice of v (section 4.4). This will be the formulation we solve in section 5 to tackle SSMF.

4.1. SSC.

Let $X = WH$ be a rank-r SSMF. After the preprocessing discussed in section 3.1, we find the corresponding SSMF of $Y = PH$, where now $Y \in \mathbb{R}^{r-1\times n}$ and $P \in \mathbb{R}^{r-1\times r}$, with the same matrix H. Since the SSC in Definition 2.1 is tested on the matrix H, we can suppose from now on that, equivalently, X or Y has an SSC decomposition.

First of all, we prove that if the translation preprocessing of X is operated with respect to the vector v corresponding to the center of W, that is, $v = We/r$, then the matrix Θ polar of P is the unique solution of the maximization problem (3.1). Recall that $P = U^\top[W - ve^\top]$, so $Pe = 0$.

In a nutshell, after a preconditioning with the singular values and left singular vectors of P, we find that the columns of the matrix Y are included in a regular and centered simplex circumscribed to the unit ball in $\mathbb{R}^{r-1}$, while preserving H and thus the SSC property. The unit ball is self-polar, so the SSC forces any possible point of $\mathrm{conv}(Y)^*$ to lie inside the unit ball. The simple observation that any maximum-volume simplex contained in the ball is necessarily regular, and that the regularity is invariant by polar transformation, concludes the proof.

Theorem 4.1. *Let $Y \in \mathbb{R}^{(r-1)\times n}$ with $n \geq r$ such that $Y = PH$ where $P \in \mathbb{R}^{(r-1)\times r}$ has full rank, $Pe = 0$, and H is an $r \times n$ SSC and column stochastic matrix. Then*

$$\max_{\Theta\in\mathbb{R}^{r-1\times r}} \mathrm{vol}(\mathrm{conv}(\Theta)) \quad \textit{such that} \quad Y^\top\Theta \leq 1_{n\times r} \tag{3.1}$$

is uniquely solved by the polar matrix of P.

Proof. Recall that the problem is equivalent to

$$\max_{\Theta\in\mathbb{R}^{r-1\times r}} \mathrm{vol}(\mathrm{conv}(\Theta)) \quad \text{such that} \quad \mathrm{conv}(\Theta) \subseteq \mathrm{conv}(Y)^*.$$

Let $P = U\Sigma Q$ be the reduced SVD of P where $U \in \mathbb{R}^{r-1\times r-1}$ is orthogonal, $\Sigma \in \mathbb{R}^{r-1\times r-1}$ is diagonal and invertible, and $Q \in \mathbb{R}^{r-1\times r}$ is such that $\left[\begin{smallmatrix} Q \\ e^\top/\sqrt{r}\end{smallmatrix}\right]$ is an $r \times r$ orthogonal matrix, since $Pe = 0$. Calling $\Psi = \Sigma U^\top \Theta$, then $\Theta = U\Sigma^{-1}\Psi$ and $\mathrm{vol}(\mathrm{conv}(\Theta)) = \mathrm{vol}(\mathrm{conv}(U\Sigma^{-1}\Psi)) = \det(\Sigma)^{-1}\mathrm{vol}(\mathrm{conv}(\Psi))$, so the problem transforms into

$$\det(\Sigma)^{-1} \max_{\Psi\in\mathbb{R}^{r-1\times r}} \mathrm{vol}(\mathrm{conv}(\Psi)) \quad \text{such that} \quad \mathrm{conv}(\Psi) \subseteq \mathrm{conv}(QH)^*. \tag{4.1}$$

Since H is SSC, $\mathrm{conv}(QH) = Q\mathrm{conv}(H) \supset Q(\Delta^r \cap \mathcal{C})$, and it is easy to prove that $Q(\Delta^r \cap \mathcal{C}) = \sqrt{\frac{1}{r(r-1)}}B^{r-1}$, where B^{r-1} is the $r-1$ dimensional unit ball. The polar of the unit ball is itself, so

$$\mathrm{conv}(\Psi) \subseteq \mathrm{conv}(QH)^* \subseteq \left(\sqrt{\frac{1}{r(r-1)}}B^{r-1}\right)^* = \sqrt{r(r-1)}B^{r-1},$$

and in particular all the columns of Ψ are bounded in squared norm by $r(r-1)$. Applying the formula for the volume, we find that

$$\mathrm{vol}(\mathrm{conv}(\Psi)) = \frac{1}{(r-1)!}\left|\det\begin{bmatrix}\Psi \\ e^\top\end{bmatrix}\right| = \frac{r^{\frac{r-1}{2}}}{(r-1)!}\sqrt{\det\begin{bmatrix}\frac{1}{\sqrt{r}}\Psi^\top & e\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{r}}\Psi \\ e^\top\end{bmatrix}} = \frac{\sqrt{\det R}}{(r-1)!}.$$

Each element of the diagonal in R is bounded by r, so its trace is at most r^2. Since R is positive semidefinite, its determinant is bounded through the arithmetic and geometric means inequality (AM-GM) by $\det(R) \le [\mathrm{tr}(R)/r]^r \le r^r$, and the equality is attained if and only if $R = rI$ or equivalently when $\left[\begin{smallmatrix} \Psi/r \\ e^\top/\sqrt{r}\end{smallmatrix}\right]$ is orthogonal. The matrix $\Psi = -rQ$ thus attains the maximum possible volume and $\mathrm{conv}(-rQ) = \mathrm{conv}(Q)^* \subseteq \mathrm{conv}(QH)^*$, so it is also a solution of problem (4.1). All other Ψ with the same volume such that $\mathrm{conv}(\Psi) \subseteq \sqrt{r(r-1)}B^{r-1}$ are rotated versions of $-rQ$, that is, $\hat{\Psi} = -rVQ$, where V is orthogonal, but

$$\begin{aligned}&\mathrm{conv}(-rVQ) = \mathrm{conv}(VQ)^* \subseteq \mathrm{conv}(QH)^* \implies \mathrm{conv}(QH) \subseteq \mathrm{conv}(VQ)\\ &\implies \mathrm{conv}(H) \subseteq \mathrm{conv}\left(\begin{bmatrix} Q^\top & e/\sqrt{r}\end{bmatrix}\begin{bmatrix} V & \\ & 1\end{bmatrix}\begin{bmatrix} Q \\ e^\top/\sqrt{r}\end{bmatrix}\right) = \mathrm{conv}(\Pi),\end{aligned}$$

and from SSC, $\Pi = Q^\top VQ + ee^\top/r$ is necessarily a permutation matrix. The simple observation that $\hat{\Psi} = -rQ\Pi$ lets us conclude that the only solutions to problem (4.1) are $-rQ$ and its permuted versions, or, equivalently, all possible polar matrices of Q. Tracing back to the original problem, we find that all possible solutions of (3.1) are the polar matrices of $\mathrm{conv}(\Theta)^* = \mathrm{conv}(U\Sigma^{-1}\Psi)^* = \mathrm{conv}(U\Sigma Q) = \mathrm{conv}(P)$. ∎

4.2. Separability. When the translation is operated with a vector v different from We/r, the SSC property is not enough anymore to guarantee that problem (3.1) is solved by the polar matrix of P. We can thus turn to the stronger separability condition. In this case, whenever v is in the relative interior of $\mathrm{conv}(X) = \mathrm{conv}(W)$, then the problem (3.1) correctly identifies the sought matrix Θ. The idea is very simple: the separability is invariant by the preprocessing of section 3.1, and any feasible Θ in (3.1) must satisfy $\mathrm{conv}(\Theta) \subseteq \mathrm{conv}(Y)^* = \mathrm{conv}(P)^*$ and, in particular, the polar set of $\mathrm{conv}(P)$ has volume larger or equal than $\mathrm{conv}(\Theta)$. The only case of equality is for when the columns of Θ coincide with the vertices of $\mathrm{conv}(P)^*$ in some order. This is enough to prove the following result.

Theorem 4.2. *Let Y be an $(r-1)\times n$ real matrix with $n \geq r$ and a separable decomposition $Y = PH$ with P an $(r-1)\times r$ real matrix, and H an $r\times n$ column stochastic matrix containing the $r\times r$ identity matrix as a submatrix (see section 2.1). If 0 is in the interior of $\operatorname{conv}(Y)$, then*

$$\max_{\Theta\in\mathbb{R}^{r-1\times r}} \operatorname{vol}(\operatorname{conv}(\Theta)) \quad \text{such that} \quad Y^\top\Theta \leq 1_{n\times r} \tag{3.1}$$

is uniquely solved by the polar matrix of P.

Proof. This follows directly from the fact, under the separability assumption, that the solution $\operatorname{conv}(\Theta) = \operatorname{conv}(Y)^*$ is feasible and therefore is the unique solution with maximum volume within $\operatorname{conv}(Y)^*$. ∎

Notice that in the separable case, $v = Xe/n$ is an acceptable choice for the translation vector to obtain identifiability using Theorem 4.2. In fact, under Assumption 3.1, $v = Xe/n$ belong to the interior of $\operatorname{conv}(X) = \operatorname{conv}(W)$.

4.3. Between SSC and separability: η-expanded. We have seen that for an SSC decomposition, we need a precise translation in the preprocessing of X, and instead in the separable case practically any sensible translation yields the correct solution, and we have a perfect candidate for it. To investigate what happens when the problem is not separable, but is more than SSC, we need to introduce a new concept called *expansion* of the data.

Definition 4.3. *We say that $H \in \mathbb{R}^{r\times n}$ is η-expanded with $\eta \in [0,1]$ if*

$$\mathcal{H}_\eta := \Delta_r \cap \left\{ x \in \mathbb{R}^r \;\middle|\; x \leq \left[\eta + (1-\eta)\frac{2}{r}\right] e \right\} \subseteq \operatorname{conv}(H).$$

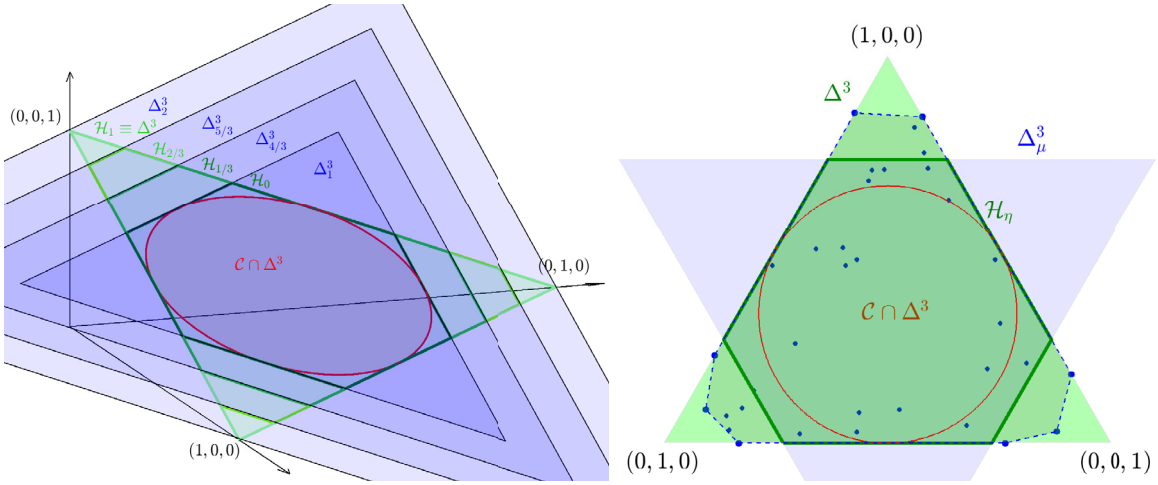
Suppose that $X \in \mathbb{R}^{m\times n}$ has rank $r-1$ and admits a decomposition $X = WH$ where H is column stochastic and η-expanded. The following properties are easily shown:

- $\eta = 1$ if and only if X is separable,
- if $\eta > 0$, then H is SSC,
- $\mathcal{C} \subset \operatorname{cone}(\mathcal{H}_0)$.

In other words, 0-expansion is close to the SSC, and the property of being η-expanded bridges between SSC and separability. In fact, the set $\mathcal{H}_\eta$ is a polytope that contains $\mathcal{C}$ for any $\eta \geq 0$. Notice that H may be SSC without being η-expanded for any $\eta \geq 0$. The set $\mathcal{H}_\eta$ is the intersection of Δ^r and Δ^r_μ obtained by symmetrizing Δ^r with respect to its center e/r and then expanding it by a constant $\mu = (r-2)\eta + 1 \in [1, r-1]$, as shown in Figure 2. In formulae,

$$\mathcal{H}_\eta = \Delta^r \cap \Delta^r_\mu = \operatorname{conv}(I) \cap \operatorname{conv}\left(\frac{\mu+1}{r} ee^\top - \mu I\right). \tag{4.2}$$

In case of SSC, Theorem 4.1 tells us that the only certified good translation vector is $v = We/r$. Instead, in case of separability, Theorem 4.2 tells us that all vectors inside the interior of $\operatorname{conv}(W)$ are good, that is, any vector that can be written as $v = Wq$, where q is strictly positive whose entries sum to one. When H is column stochastic and η-expanded, we can prove that any translation vector that can be written as $v = Wq$, where q is strictly positive, whose entries sum to one, and $0 < q < \frac{r\eta+2(1-\eta)}{2r} e$, yields the correct solution to problem (3.1). To do so, we first need two lemmas that show how the polar duality behaves



(a) Δ^3_μ for $\mu = 1, 4/3, 5/3, 2$ and the associated $\mathcal{H}_\eta = \Delta^3 \cap \Delta^3_\mu$ for $\eta = \mu - 1 = 0, 1/3, 2/3, 1$.

(b) Two-dimensional projection of the columns of a η-expanded and column stochastic H (dots).

Figure 2. *Visual representation in three and two dimensions for the unit simplex Δ^3, the cone $\mathcal{C}$ intersected with Δ^3, the symmetrized and expanded Δ^3_μ, the associated $\mathcal{H}_\eta = \Delta^3 \cap \Delta^3_\mu$, and an η-expanded H.*

under translation of the polytopes and how to compute the volume of the polar matrix after such translation. We provide the proofs in the supplementary material (M165060_01.pdf [local/web 346KB]). From now on, we use $\mathcal{S}^\circ$ to indicate the interior of a set $\mathcal{S}$.

Lemma 4.4. *Suppose the columns of $\Theta \in \mathbb{R}^{(r-1)\times r}$ are the vertices of the polar set of a convex polytope $\mathcal{S}$. For any $w \in \mathcal{S}^\circ$ suppose that $\Theta_w \in \mathbb{R}^{(r-1)\times r}$ are the vertices of the polar set of $\mathcal{S} - w$. If z_w is such that $\Theta_w z_w = 0$ and $e^\top z_w = 1$, then the matrix $\Theta_w \mathrm{diag}(z_w)$ does not depend on w and $\Theta_w = \Theta \mathrm{diag}(e - \Theta^\top w)^{-1}$.*

In particular, given a matrix $A \in \mathbb{R}^{(r-1)\times r}$ with $Ae = 0$, for any $w \in \mathrm{conv}(A)^\circ$ call Θ_w the polar of $A - we^\top$ and suppose $At = v$ and $As = z$ with $t, s \in \Delta^r$ and $v, z \in \mathrm{conv}(A)^\circ$. Then $\Theta_v \mathrm{diag}(t) = \Theta_z \mathrm{diag}(s)$ and $\Theta_v t = \Theta_z s = 0$.

Lemma 4.5. *Given $\Theta \in \mathbb{R}^{(r-1)\times r}$, suppose that $\Theta w = 0$ for a nonzero vector w with $e^\top w \neq 0$. Then for any invertible matrix N,*

$$\mathrm{vol}(\mathrm{conv}(\Theta N)) = |\det(N)| \left| \frac{e^\top N^{-1} w}{e^\top w} \right| \mathrm{vol}(\mathrm{conv}(\Theta)). \tag{4.3}$$

Now we can state and prove our result.

Theorem 4.6. *Suppose that $Y = PH$ with H η-expanded and column stochastic and P full rank. Consider the set of vectors q such that $Pq = 0$ and $e^\top q = 1$. If there is a vector q in this set such that $0 < q < \frac{r\eta + 2(1-\eta)}{2r} e$, then the problem*

$$\max_{\Theta \in \mathbb{R}^{r-1\times r}} \mathrm{vol}(\mathrm{conv}(\Theta)) \quad \textit{such that} \quad Y^\top \Theta \leq 1_{n\times r} \tag{3.1}$$

is solved uniquely by the polar matrix of P.

Proof. Let[2] $v := Pe/r$ and let $P_v := P - ve^\top$ with its SVD being $P_v = U\Sigma Q$. Notice that $Q \in \mathbb{R}^{r-1\times r}$ is such that $\left[\begin{smallmatrix} Q \\ e^\top/\sqrt{r}\end{smallmatrix}\right]$ is an $r\times r$ orthogonal matrix, since $P_v e = 0$. Similarly as the proof of Theorem 4.1, problem (3.1) is equivalent to

$$\det(\Sigma)^{-1} \max_{\Psi\in\mathbb{R}^{r-1\times r}} \operatorname{vol}(\operatorname{conv}(\Psi)) \quad \text{such that} \quad \operatorname{conv}(\Psi) \subseteq \operatorname{conv}(Q(I-qe^\top)H)^*, \tag{4.4}$$

where $\Psi = \Sigma U^\top \Theta$ and

$$Q(I-qe^\top)H = (Q + \Sigma^{-1}U^\top(ve^\top - P)qe^\top)H = \Sigma^{-1}U^\top(U\Sigma Q + ve^\top)H = \Sigma^{-1}U^\top Y.$$

Since $\frac{r\eta+2(1-\eta)}{2r} < \frac{r\eta+2(1-\eta)}{r}$, the vector q is in the interior of $\mathcal{H}_\eta$ and as a consequence $0 = Q(I-qe^\top)q$ is in the interior of $Q(I-qe^\top)\mathcal{H}_\mu$. This enables us to freely utilize the properties of the polar duality and find the necessary condition $\operatorname{conv}(\Psi) \subseteq (Q(I-qe^\top)\mathcal{H}_\eta)^* = \operatorname{conv}((Q(I-qe^\top)\Delta^r)^* \cup (Q(I-qe^\top)\Delta^r_\mu)^*)$, where Δ^r_μ is defined in (4.2) and $\mu = (r-2)\eta + 1 \in [1, r-1]$. The vertices of the polytope $(Q(I-qe^\top)\mathcal{H}_\eta)^*$ are thus (contained in the set of) the vertices of $(Q(I-qe^\top)\Delta^r)^* = \operatorname{conv}(Q - Qqe^\top)^* = \operatorname{conv}(M)$ and of

$$\left(Q(I-qe^\top)\Delta^r_\mu\right)^* = -\frac{1}{\mu}\operatorname{conv}\left(Q(I-qe^\top)\left(I - \frac{\mu+1}{r\mu}ee^\top\right)\right)^* = -\frac{1}{\mu}\operatorname{conv}\left(Q + \frac{1}{\mu}Qqe^\top\right)^*.$$

Notice that $-\frac{1}{\mu}Qq = Q(-\frac{q}{\mu} + \frac{\mu+1}{\mu}\frac{e}{r})$, so due to Lemma 4.4, we get

$$\left(Q(I-qe^\top)\Delta^r_\mu\right)^* = -\frac{1}{\mu}\operatorname{conv}\left(Q - Q\left(-\frac{q}{\mu} + \frac{\mu+1}{\mu}\frac{e}{r}\right)e^\top\right)^* = \operatorname{conv}\left(M\operatorname{diag}\left(\frac{1}{1-\frac{\mu+1}{rq_i}}\right)\right).$$

The vertices of a maximum-volume simplex Ψ in $(Q(I-qe^\top)\mathcal{H}_\eta)^*$ must correspond to r of its vertices, so from the above computation can only be a column m_i of M or $\alpha_i m_i$, where $\alpha_i = 1/(1-\frac{\mu+1}{rq_i})$. If Ψ has among its vertices $\{m_i, \alpha_i m_i, m_j, \alpha_j m_j\}$ with $i \neq j$, then the rank of Ψ is at most $r-2$ and its volume is zero. Therefore, we only need to consider the following sets of r vertices $\{v_1, \dots, v_r\}$:

1. $v_i \in \{1, \alpha_i\}m_i$ for all i,
2. there exists exactly one index i such that both $\alpha_i m_i$ and m_i are among the vertices.

Since M is the polar of $Q - Qqe^\top$, Lemma 4.4 says that $Mq = 0$. As a consequence, using (4.3), the volume of any simplex of the first kind is

$$V_1 := \left|\prod_{i\in S}\alpha_i\right|\left|1 + \sum_{i\in S} q_i\left(\frac{1}{\alpha_i} - 1\right)\right|\operatorname{vol}(\operatorname{conv}(M)) = \left|\prod_{i\in S}\frac{1}{1-\frac{\mu+1}{rq_i}}\right|\left|1 - |S|\frac{\mu+1}{r}\right|\operatorname{vol}(\operatorname{conv}(M)),$$

where $S := \{i \,|\, v_i = \alpha_i m_i\}$, and if S is empty, then V_1 is equal to $\operatorname{vol}(\operatorname{conv}(M))$. By hypothesis, $q_i < \frac{r\eta+2(1-\eta)}{2r} = \frac{\mu+1}{2r}$, so $\alpha_i < 1$. As a consequence, if $|S|(\mu+1) \leq 2r$ and S is not empty, we find that $V_1 < \operatorname{vol}(\operatorname{conv}(M))$. For $|S|(\mu+1) > 2r$, we have

$$V_1 = \prod_{i\in S}\frac{1}{\frac{\mu+1}{rq_i} - 1}\left(|S|\frac{\mu+1}{r} - 1\right)\operatorname{vol}(\operatorname{conv}(M)),$$

[2]We abuse notation here since v is now the translation vector in the reduced space, not in the original one.

but thanks to the Jensen inequality applied to the concave function $f(x) = \ln \frac{1}{1/x-1}$ with weights equal to $1/|S|$ and points $x_i = rq_i/(\mu+1) < 1/2$ we get

$$\prod_{i\in S} \frac{1}{\frac{\mu+1}{rq_i}-1} = \exp\left(\sum_{i\in S} \ln \frac{1}{\frac{\mu+1}{rq_i}-1}\right) \le \exp\left(|S| \ln \frac{1}{\frac{1}{\frac{r}{|S|(\mu+1)}\sum_{i\in S} q_i}-1}\right) \le \left(\frac{1}{|S|\frac{\mu+1}{r}-1}\right)^{|S|},$$

and since $|S| > 2r/(\mu+1) \ge 2$, we find again that $V_1 < \mathrm{vol}(\mathrm{conv}(M))$.

For the polytopes of the second kind, suppose without loss of generality that $v_1 = m_2$, $v_2 = \alpha_2 m_2$ and $v_i = \nu_i m_i$ for $i > 2$, where $\nu_i \in \{1, \alpha_i\}$. Then

$$V_2 := \mathrm{vol}\begin{pmatrix} m_2 & m_2\alpha_2 & m_3\nu_3 & \dots & m_r\nu_r \end{pmatrix} = \frac{|\alpha_2 - 1| \cdot \left|\prod_{k\ge 3} \nu_k\right|}{(r-1)!} |\det(\hat{M})|,$$

where $\hat{M}$ is the top-right $(r-1)\times(r-1)$ submatrix of M. Since M is the polar of $Q(I - qe^\top)$, by Lemma 4.4 we find that $M = -Q\mathrm{diag}(q)^{-1}$, and if $\hat{Q}$ is the submatrix of Q associated to M, then $|\det(\hat{M})| = |\det(\hat{Q})|/\prod_{i>1} q_i$ and by (4.3),

$$\mathrm{vol}(\mathrm{conv}(M)) = \frac{\mathrm{vol}(\mathrm{conv}(Q))}{r\prod_i q_i} = \frac{1}{(r-1)!}\frac{r|\det(\hat{Q})|}{r\prod_i q_i} = \frac{|\det(\hat{M})|}{(r-1)!}\frac{1}{q_1}.$$

If now $S := \{i \,|\, v_i = \alpha_i m_i,\, i > 2\} = \{i \,|\, \nu_i = \alpha_i,\, i > 2\}$, then V_2 reduces to

$$V_2 = \frac{(\mu+1)q_1}{(\mu+1) - rq_2} \prod_{i\in S} \frac{1}{\frac{\mu+1}{rq_i}-1} \mathrm{vol}(\mathrm{conv}(M)),$$

but from the hypothesis $q_i < \frac{\mu+1}{2r}$, so it is immediate to see that

$$\frac{(\mu+1)q_1}{(\mu+1) - rq_2} \prod_{i\in S} \frac{1}{\frac{\mu+1}{rq_i}-1} < \frac{(\mu+1)\frac{\mu+1}{2r}}{(\mu+1) - r\frac{\mu+1}{2r}} = \frac{\mu+1}{r} \le 1,$$

and thus $V_2 < \mathrm{vol}(\mathrm{conv}(M))$.

The polytope with the largest volume inside of $(Q(I - qe^\top)\mathcal{H}_\eta)^*$ thus coincides with the polar of $\mathrm{conv}(Q(I - qe^\top))$ that is in particular contained in $\mathrm{conv}(Q(I - qe^\top)H)^*$. The matrices Ψ describing the polar of $Q(I - qe^\top)$ are therefore the unique solutions to (4.4). Going back to the original problem, we see that it is solved uniquely by $\Theta = U\Sigma^{-1}\Psi$ being the polar of $U\Sigma Q(I - qe^\top) = P_v(I - qe^\top) = P$. ∎

When X is separable, that is, H is 1-expanded, Theorem 4.2 says that the only condition needed for the correctness of the solution of the problem 3.1 is $v = Wq$, where q has sum 1 and $0 < q < e$. In this case, though, Theorem 4.6 only holds for $0 < q < e/2$. This suggests that the result can be improved.

Conjecture 4.7. The thesis of Theorem 4.6 holds if $q_i > \frac{1-\eta}{r}$ for every i.

4.4. Min-max approach under the SSC. Under the SSC, we have proved that the solution to problem (3.1) coincides with the SSC decomposition $Y = PH$ when $Pe = 0$. In the case that $v := Pe/r \neq 0$ one would need to translate Y by v before solving problem (3.1), so that $Y - ve^\top = (P - ve^\top)H$, and the resulting solution Θ would coincide with the polar set of $P - ve^\top$. Since v is not generally known beforehand, we inquire what happens when we translate by a different vector w. We find that the solution Θ_w of (3.1) applied to the matrix $Y - we^\top$ has always a strictly larger volume than the correct solution $\Theta \equiv \Theta_v$, and the volume of Θ_w is actually a convex function in w.

Theorem 4.8. *Let Y be an $(r-1) \times n$ real matrix with $n \geq r$ such that $Y = PH$ with P an $(r-1) \times r$ full rank real matrix and H an $r \times n$ SSC and column stochastic matrix. If*

$$\mathcal{V}(w) := \sup_{\Theta \in \mathbb{R}^{r-1 \times r}} \operatorname{vol}(\operatorname{conv}(\Theta)) \quad \text{such that} \quad (Y - we^\top)^\top \Theta \leq 1_{n \times r}, \tag{4.5}$$

for any vector $w \in \mathbb{R}^{r-1}$, then $\mathcal{V}(w)$ is a convex function with unique minimum at $w = v = Pe/r$.

Proof. If $w \notin \operatorname{conv}(Y)^\circ$, then $\operatorname{conv}(Y)^*$ is unbounded and $\mathcal{V}(w) = \infty$, so from now on we suppose $w \in \operatorname{conv}(Y)^\circ \subseteq \operatorname{conv}(P)^\circ$. We can now rewrite the problem as

$$\mathcal{V}(w) = \sup_{\Theta \in \mathbb{R}^{r-1 \times r}} \operatorname{vol}(\operatorname{conv}(\Theta)) \quad \text{such that} \quad \operatorname{conv}(\Theta) \subseteq \operatorname{conv}(Y - we^\top)^*.$$

The polar matrix Ψ_w of $Y - we^\top$ represents a polytope, so $\mathcal{V}(w)$ will be the volume of a simplex $\operatorname{conv}(\Theta_w)$ whose vertices are an r-subset of the s columns of Ψ_w, as we show in Lemma SM1.1 of the supplementary material (M165060_01.pdf [local/web 346KB]). The maximum is thus achieved by one out of $\binom{s}{r}$ simplices $\Theta_w^{(i)}$, and we can recast the problem as

$$\mathcal{V}(w) = \max_{i=1,\dots,\binom{s}{r}} \operatorname{vol}\left(\operatorname{conv}\left(\Theta_w^{(i)}\right)\right) \quad \text{such that} \quad \Theta_w^{(i)} = \Psi_w I_{s \times r}^{(i)},$$

where $\{I_{s \times r}^{(i)}\}_{i=1,\dots,\binom{s}{r}}$ are all the possible full rank, binary, and column stochastic matrices of size $s \times r$. Since each $\Theta_w^{(i)}$ represents r linear constraints of $\operatorname{conv}(Y - we^\top)^*$, then its polar set $\mathcal{S}_w^{(i)}$ is just the w-translated of a fixed (and possibly unbounded) polytope with r facets containing $\operatorname{conv}(Y)$. If we now fix the vector $\ell = Ye/n \in \operatorname{conv}(Y)^\circ$, then by Lemma 4.4,

$$\begin{aligned}\mathcal{V}_i(w) &:= \operatorname{vol}\left(\operatorname{conv}\left(\Theta_w^{(i)}\right)\right) = \operatorname{vol}\left(\operatorname{conv}\left(\Theta_\ell^{(i)}\right) \operatorname{diag}\left(e - \left(\Theta_\ell^{(i)}\right)^\top (w - \ell)\right)^{-1}\right) \\ &= \frac{\operatorname{vol}\left(\operatorname{conv}(\Theta_\ell)\right)}{\prod_j \left[e - \left(\Theta_\ell^{(i)}\right)^\top (w - \ell)\right]_j}.\end{aligned}$$

Notice now that $-\ln(x)$ and e^x are both convex functions, so we can prove that $\mathcal{V}_i(w)$ is also a convex function. In fact $[e - (\Theta_\ell^{(i)})^\top (w - \ell)]_j > 0$ for any $w \in \operatorname{conv}(Y)^\circ$ and any j, so for any $\lambda \in [0, 1]$ and any couple of points $w_1, w_2 \in \operatorname{conv}(Y)^\circ$,

$$
\begin{aligned}
&\mathcal{V}_i(\lambda w_1 + (1-\lambda)w_2) \\
&\quad = \frac{\operatorname{vol}(\operatorname{conv}(\Theta_\ell))}{\prod_j \left[e - \left(\Theta_\ell^{(i)}\right)^\top (\lambda w_1 + (1-\lambda)w_2 - \ell)\right]_j} \\
&\quad = \frac{\operatorname{vol}(\operatorname{conv}(\Theta_\ell))}{\prod_j \lambda\left[e - \left(\Theta_\ell^{(i)}\right)^\top (w_1 - \ell)\right]_j + (1-\lambda)\left[e - \left(\Theta_\ell^{(i)}\right)^\top (w_2 - \ell)\right]_j} \\
&\quad \le \operatorname{vol}(\operatorname{conv}(\Theta_\ell)) \exp\left(-\lambda \sum_j \ln\left(\left[e - \left(\Theta_\ell^{(i)}\right)^\top (w_1 - \ell)\right]_j\right)\right. \\
&\qquad \left. - (1-\lambda)\sum_j \ln\left(\left[e - \left(\Theta_\ell^{(i)}\right)^\top (w_2 - \ell)\right]_j\right)\right) \\
&\quad \le \operatorname{vol}(\operatorname{conv}(\Theta_\ell)) \left(\lambda \prod_j \frac{1}{\left[e - \left(\Theta_\ell^{(i)}\right)^\top (w_1 - \ell)\right]_j} + (1-\lambda)\prod_j \frac{1}{\left[e - \left(\Theta_\ell^{(i)}\right)^\top (w_2 - \ell)\right]_j}\right) \\
&\quad = \lambda \mathcal{V}_i(w_1) + (1-\lambda)\mathcal{V}_i(w_2).
\end{aligned}
$$

The function $\mathcal{V}(w)$ is now the maximum of convex functions, so it is also convex.

Since $Y - we^\top = (P - we^\top)H$, we have that the polar matrix $\tilde{\Theta}_w$ of $P - we^\top$ satisfies (4.5), so by Lemma 4.4 and (4.3),

$$
\mathcal{V}(w) \ge \operatorname{vol}(\operatorname{conv}(\tilde{\Theta}_w)) = \operatorname{vol}(\operatorname{conv}(\tilde{\Theta}_v \operatorname{diag}(1/rt_i))) = \frac{\operatorname{vol}(\operatorname{conv}(\tilde{\Theta}_v))}{r^r \prod_i t_i},
$$

where $Pt = w$ and $t \in \Delta^r$. A simple application of AM-GM tells us that $\prod_i t_i \le 1/r^r$. We know by Theorem 4.1 that $\mathcal{V}(v) = \operatorname{vol}(\operatorname{conv}(\tilde{\Theta}_v))$, so we conclude that

$$
\mathcal{V}(w) \ge \frac{\operatorname{vol}(\operatorname{conv}(\tilde{\Theta}_v))}{r^r \prod_i t_i} \ge \mathcal{V}(v)
$$

with equality only if $t_i = 1/r$ for every i, i.e., $w = Pe/r = v$. ∎

Remark 4.9. It can be shown that for any translation vector w that is arbitrarily close to Pe/r, there exists an H satisfying the SSC such that the maximum-volume matrix Θ in (4.5) is very different from the polar of P, but its volume $\mathcal{V}(w)$ will be close to that of $\operatorname{conv}(P)^*$.

If H is fixed and satisfies the SSC, then Θ will be equal to $\operatorname{conv}(P)^*$ for translation vectors w belonging to an open neighborhood of Pe/r. In case H is η-expanded for some $\eta > 0$, Theorem 4.6 identifies the above mentioned neighborhood, but if H is only SSC, the neighborhood depends on H and there cannot be uniform bounds.

Remark 4.10. Although the function $\mathcal{V}(w)$ is convex in w, it is hard to evaluate as it requires solving a nonconvex optimization problem in Θ. Hence we cannot use this observation to provide theoretical convergence guarantees. This is a typical issue when dealing with min-max optimization problems (a.k.a. saddle point problems) where the inner problem is not concave; see, e.g., [8] and the references therein. However, in our proposed Algorithm 5.1, the

update of the center of the translation (denoted v in the algorithm; see step 15) is the center of the current solution, which is optimal assuming Θ is fixed; see Theorem 4.8.

Theorems 4.8 and 4.1 imply the following.

Corollary 4.11. *Let Y be an $(r-1)\times n$ real matrix with $n \geq r$ such that $Y = PH$ with P an $(r-1)\times r$ full rank real matrix and H an $r \times n$ SSC and column stochastic matrix. Then*

$$\inf_{w\in\mathbb{R}^{r-1}} \sup_{\Theta\in\mathbb{R}^{r-1\times r}} \operatorname{vol}(\operatorname{conv}(\Theta)) \quad \text{such that} \quad (Y - we^\top)^\top \Theta \leq 1_{n\times r} \tag{4.6}$$

is solved uniquely by $w = Pe/r$ and Θ being the polar matrix of $P - we^\top$.

Corollary 4.11 incites us to update the translation vector v and the solution Θ using a min-max approach: v should be chosen to minimize the volume and Θ to maximize it. This is described in the next section. It is interesting to note that the min-max approach would converge in one iteration under the separability condition (since any w in the convex hull of Y leads to the sought Θ; see Theorem 4.2), while the set of v's that lead to identifiability typically contains more than the point Pe/r, as shown in Theorem 4.6 when H is η-expanded. In practice, we will see that alternating minimization of v and Θ typically converges within a few iterations.

5. Optimization.

In this section, we propose an algorithm to solve our min-max formulation (4.6) that estimates the translation vector, v, together with the corresponding dual solution, Θ; see Algorithm 5.1. This is a natural approach since, in practice, the optimal v, given by the center of the convex hull of W, is unknown, while our min-max formulation (4.6) provides identifiability guarantees; see Corollary 4.11.

To solve (4.6), we resort to a standard alternating optimization strategy: optimize v and Θ alternatively. Let us first assume that the translation vector, v, is known. In the presence of noise, we propose to consider the following formulation:

$$\max_{Z,\Theta,\Delta} \det(Z)^2 - \lambda\|\Delta\|_F^2 \quad \text{such that} \quad Z = \begin{bmatrix} \Theta \\ e^\top \end{bmatrix} \text{ and } \quad Y^\top\Theta \leq 1_{n\times r} + \Delta. \tag{5.1}$$

The matrix Δ belongs to $\mathbb{R}^{(r-1)\times n}$ and represents the noise matrix, while $\lambda > 0$ serves as a regularization parameter. Moreover, we have squared the volume of $\operatorname{conv}(\Theta)$ in the objective to make it smooth (getting rid of the absolute value). In this problem, the objective function is nonconcave; however, all the constraints are linear. Inspired by the work of [21], we use the block successive upperbound minimization framework [30] and iteratively update the columns Θ.

Using the co-factor expansion within Laplace formula, we express $\det(Z)$ as a linear function of the entries in any kth column: $\det(Z) = \sum_{j=1}^{r}(-1)^{j+k}Z(j,k)\det(Z_{-j,-k})$, where $Z_{-j,-k}$ is obtained by removing the jth row and kth column from Z. If we fix all columns of Z but the kth, we have

$$\det(Z) = {f^{(k)}}^\top Z(:,k), \text{ where } f^{(k)}(j) = (-1)^{j+k}\det(Z_{-j,-k}) \quad \text{for } j = 1,\dots,r.$$

For simplicity, let us denote $c = f^{(k)}$ and $x = Z(:,k)$. We want to maximize $f(x) = \det(Z)^2 = ({f^{(k)}}^\top Z(:,k))^2 = (c^\top x)^2$. The function $f(x) = x^\top(cc^\top)x$ is a convex quadratic function that can be lower bounded by its first-order Taylor approximation, that is, for any x_0,

$$f(x) = (c^\top x)^2 \geq f(x_0) + \nabla f(x_0)^\top (x - x_0) = 2[cc^\top x_0]^\top x + (c^\top x_0)^2 = d^\top x + \text{ constants},$$

since $\nabla f(x_0) = 2(cc^\top)x_0$, where $d = 2cc^\top x_0^\top = 2f^{(k)} {f^{(k)}}^\top x_0 = \alpha f^{(k)}$, where x_0 is the previous value of $Z(:,k)$ (from a previous iteration), and $\alpha = 2{f^{(k)}}^\top x_0 = 2\det(Z)$. Hence we have a "minorizer" of $\det(Z)^2$ as a function of $x = Z(:,k)$ around x_0.

Per this inequality, the iterative maximization of $\det(Z)^2$ involves sequentially updating columns of Z and optimizing the lower-bound expression for each column of Z until convergence. In each iteration, individual columns of Z (and Θ) are updated by considering every other column as fixed and solving a quadratic programming problem of the form (for $k = 1, \ldots, r$):

(5.2)
$$\max_{t,\Theta(:,k),\Delta(:,k)} \alpha {f^{(k)}}^\top t - \lambda ||\Delta(:,k)||_2^2 \text{ s.t. } t = \begin{bmatrix} \Theta(:,k) \\ 1 \end{bmatrix} \text{ and } Y^\top \Theta(:,k) \leq 1_{(r-1)\times 1} + \Delta(:,k).$$

However, this optimization problem alone is insufficient to guarantee the boundedness of the corresponding simplex in the primal space. The columns in Θ define a bounded simplex in $\mathbb{R}^{r-1}$ if and only if the positive hull of Θ spans $\mathbb{R}^r$, or equivalently if 0 is in the interior of its convex hull. This means that there exists $\alpha \in \mathbb{R}^r$ such that $\Theta\alpha = 0$, $\alpha > 0$, and $e^\top \alpha = 1$. Consequently, dividing $\Theta\alpha = 0$ by α_k and rearranging the terms, we obtain $\Theta(:,k) = -\sum_{i \neq k} \alpha_i' \Theta(:,i)$ with $\alpha_i' = \alpha_i/\alpha_k \geq \epsilon > 0$, which we add as a constraint to the problem above. We will use $\epsilon = 0.01$ in the experiments.

Similar to [21], we use a numerical trick to define the vector $f^{(k)}$ as the columns of Z^{-1}. This is based on Carner's rule and helps to avoid round-off errors.

Initializing and updating the translation vector v. As explained in detail in section 4, the choice of the translation vector v in the preprocessing step, $Y = U^\top(X - ve^\top)$, is crucial for the identifiability of SSMF via volume maximization in the dual. The best choice for v is We/r but it is unknown a priori. To initialize v, we resort to two strategies:

1. $v_0 = Xe/n$, which is the sample average. This solution could be a bad approximation of We/r when the samples are not well scattered within $\text{conv}(W)$.
2. v_0 is the average of the vertices extracted by the successive nonnegative projection algorithm (SNPA), an effective separable NMF algorithm. This approach is less sensitive to imbalanced distributions within $\text{conv}(W)$.

The first step of our proposed algorithm (Algorithm 5.1) is to initialize v with v_0, and then resort to an alternating strategy. For v fixed, we solve (5.2) and a solution Θ is obtained. The corresponding matrix W can be estimated via the vertices of the dual of Θ, by solving a system of linear equations: to estimate the kth column of W, solve $\Theta(:,j)^\top \hat{W}(:,k) = 1$ for $j \neq k$ and then let $\tilde{W}(:,k) = U\hat{W}(:,k) + v$, where v is the current translation vector. Then, for Θ fixed, we update v using $v \leftarrow \tilde{W}e/r$: this will reduce the value of the objective in (4.6); see Theorem 4.8.

Mitigating sensitivity to initialization. Our numerous numerical experiments have shown that solving the optimization problem in (5.2) is usually not too sensitive to the initialization. However, when there exist two or more candidate simplices with close volumes, the algorithm might converge to suboptimal solutions. To reduce sensitivity to initialization, the optimization algorithm is executed multiple times concurrently, each time with distinct random

Algorithm 5.1. Maximum volume in the dual (MV-Dual).

Require: Data matrix $X \in \mathbb{R}^{m \times n}$, a factorization rank r, the regularization parameter $\lambda > 0$, the number of random initializations n_init (default = 5).

Ensure: A matrix W such that $X \approx WH$ where H is column stochastic.

% Step 1. *Initialization of* v *and* Y

Initialize v_0 with the sample mean $v_0 = Xe/n$ or with $X(:, \mathcal{K})e/|\mathcal{K}|$ where $\mathcal{K}$ is obtained via SNPA.

Let $Y = U^\top (X - v_0 e^\top) = U^\top X - U^\top v_0 e^\top$, where the columns of U are the first $(r-1)$ singular vectors of $X - v_0 e^\top$.

% Step 2. *Initialize the set of solutions*

Initialize the set of n_init solutions as $\mathcal{S} = \{Z_i\}_{i=1}^{n\_init}$ where $Z_i = \begin{bmatrix} \Theta_i \\ e^\top \end{bmatrix} \in \mathbb{R}^{r \times r}$ and the entries of $\Theta_i \in \mathbb{R}^{r-1 \times r}$ are sampled from $\mathcal{N}(0, 1)$.

$p = 1$.

while not converged: $p = 1$ or $\frac{\|v_p - v_{p-1}\|_2}{\|v_{p-1}\|_2} > 0.01$ **do**

  % Step 3.a. *Update* Θ *and* W

  for each candidate matrix Z_i in $\mathcal{S}$ (can be parallelized) **do**

    Solve (5.1) via alternating optimization to update Z_i and Θ_i.

    % Note: The problem (5.1) *also involves the variable* Δ *that models the noise, but it is not explicitly needed to estimate* Θ *and* W*, and hence we discard it from the description of the algorithm.*

  end for

  Compute the volume of each of candidate solutions in $\mathcal{S}$ and select the one with the largest volume, which we denote Θ.

  Recover $\hat{W}$ by computing the dual of conv (Θ).

  Project back to the original space: $W = U\hat{W} + v_{p-1}$.

  % Step 3.b. *Update* v *and* Y

  Let $v_p \leftarrow We/r$, and let $Y \leftarrow U^\top X - U^\top v_p e^\top$.

  $p = p + 1$.

end while

initializations. The selected Θ is the one that results in the largest volume. We will use five random initializations for this purpose in our numerical experiments.

Algorithm 5.1 summarizes our proposed algorithm for SSMF, which we refer to as MV-Dual.

Computational cost. The preprocessing requires the computation of the truncated SVD, in $\mathcal{O}(mnr^2)$ operations. The main cost is to solve (5.1) by alternatively optimizing (5.2) which is a quadratic program in $\mathcal{O}(n)$ variables and constraints. Such problems require $\mathcal{O}(n^3)$ operations in the worst case. However, we have observed that it is typically solved significantly faster by the solver; rather in linear time in n, and we will solve real instances of (5.1) with

$n = 10^4$ in 15 s (Table 4). The reason is that this problem has a particular structure. The variables $Z(:,k)$ and $\Theta(:,k)$ are r-dimensional, while the n-dimensional variable, $\Delta(:,k)$, only appears with the identity matrix in the constraints. In the noiseless case, $\Delta(:,k) = 0$ and hence it could be removed from the formulation leading to an $\mathcal{O}(r^3)$ complexity. In the noisy case, only a few entries of $\Delta(:,k)$ will be nonzero, namely the entries corresponding to data points outside the hyperplane defined by $\Theta(:,k)$. Further research includes the design of a dedicated solver to tackle (5.2), e.g., using an active-set approach.

Note that in step 8 of Algorithm 5.1, we use the stopping criterion $\frac{\|Z_\ell - Z_{\ell-1}\|_F}{\|Z_{\ell-1}\|_F} \leq 10^{-3}$ where Z_ℓ is obtained after updating each column of $Z_{\ell-1}$ using (5.2), or a maximum number of 100 iterations.

6. Numerical experiments.

In this section, we present numerical experiments to show the behavior of the proposed MV-Dual algorithm under various settings and conditions compared to the state of the art. All experiments are implemented in MATLAB (R2019b) and run on a laptop with an Intel Core i7-9750H @2.60 GHz CPU and 16 GB RAM. The code, data, and all experiments are available from https://github.com/mabdolali/MaxVol_Dual/.

SSMF algorithms. We compare MV-Dual to six state-of-the-art algorithms:

- SNPA [16] is based on the *separability* assumption and presents a robust extension to the successive projection algorithm (SPA) [2, 16] by taking advantage of the nonnegativity constraint in the decomposition.
- Simplex volume minimization (Min-Vol) fits a simplex with minimum volume to the data points using the following optimization problem [24]:

$$\min_{W,H} \|X - WH\|_F^2 + \lambda \text{logdet}(W^\top W + \delta I_r) \quad \text{s.t. } H(:,j) \in \Delta^r \text{ for all } j.$$

  This problem is optimized based on a block coordinate descent approach using the fast gradient method. The parameter λ is chosen as in [24]: $\tilde{\lambda} \frac{\|X - W_0 H_0\|_F^2}{\text{logdet}(W_0^\top W_0 + \delta I_r)}$ where (W_0, H_0) is obtained by SNPA and $\tilde{\lambda} \in \{0.1, 1, 5\}$ where 0.1 is the default value in [24].
- MVES [10] searches for an enclosing simplex with minimum volume and converts the problem into a determinant maximization problem by focusing on the inverse of $\tilde{W}$ defined in (2.2).
- Maximum-volume inscribed ellipsoid (MVIE) [27] inscribes a maximum-volume ellipsoid in the convex hull of the data points to identify the facets of $\text{conv}(W)$.
- Hyperplane-based Craig-simplex-identication (HyperCSI) [25] is a fast algorithm based on SPA but does not rely on separability assumption. HyperCSI extracts the *purest* samples using SPA and uses these samples to estimate the enclosing facets of the simplex.
- GPFI [1] has the weakest conditions to recover the unique decomposition among the state-of-the-art methods. This approach sequentially extracts the facets with the largest number of points by solving a computationally expensive mixed integer program.

We have kept the original implementations of the authors. Note that MVES, MVIE, HyperCSI, and GPFI also use a preprocessing using the truncated SVD to denoise the data, as MV-Dual does.

To assess the quality of a solution, W, we measure the relative distance between the column of W and the columns of the ground-truth W_t:

$$ERR = \min_{\pi,\text{a permutation}} \frac{\|W_t - W_\pi\|_F}{\|W_t\|_F},$$

where W_π is obtained by permuting the columns of W.

6.1. Synthetic data. In this section, we compare the SSMF algorithms on noiseless and noisy synthetic data sets.

Data generation. We generate synthetic data following [1]. Two categories of samples are generated: n_1 samples are produced exactly on the r facets, and n_2 samples are produced within the simplex, for a total number of $n = n_1 + n_2$ samples. The entries of the ground-truth matrix W_t are uniformly distributed in the interval $[0, 1]$ and the nonzero columns of H_t are generated using the Dirichlet distribution with all parameters equal to $1/d$ where d is the dimension of the simplex where samples are generated. We define the purity parameter $p \in (\frac{1}{r-1}, 1]$ as $p(H_t) = \min_{1\leq k\leq} \|H_t(k,:)\|_\infty$, which quantifies how well the ground-truth data is spread within $\text{conv}(W_t)$. (The lower bound $\frac{1}{r-1}$ comes from the fact that n_1 columns of H are on facets of Δ^r, that is, have at least one entry equal to zero.) Given a purity level p, columns of H are resampled as long as they contain an entry larger than p. Note that the separability assumption is satisfied when $p(H_t) = 1$, hence the columns of W_t appear as columns among the samples in X. The SSC is satisfied for smaller values of purity values [27]. For the noisy setting, we add independent and identically distributed mean-zero Gaussian noise to the data, with variance chosen according to the following formula for a given signal-to-noise (SNR) ratio:

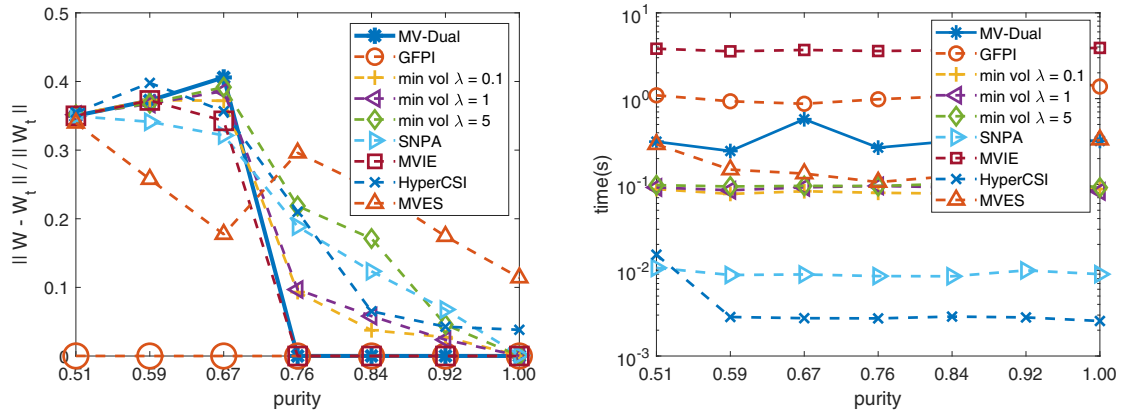
$$\text{variance} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} X(i,j)^2}{10^{SNR/10} \times m \times n}.$$

Parameter setting. For noiseless cases, we can set λ to any high number. We used $\lambda = 100$ for all the noiseless experiments. We set λ to $10, 1, 0.5, 0.03, 0.01$ for SNR values of $60, 40, 30, 20, 10$, respectively.

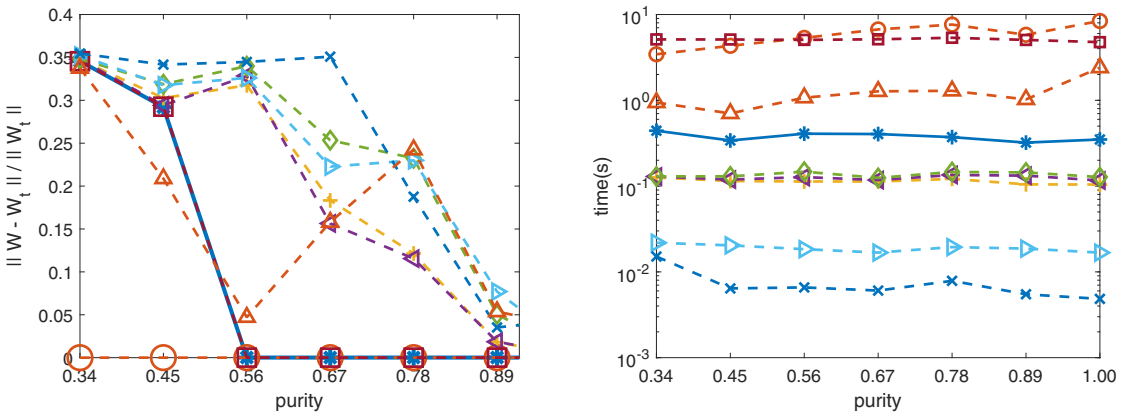
Noiseless data. First, we compare the performances for different values of purity parameters p in the noiseless case. Due to the randomness of the data generation process, the reported results are the average over 10 trials. We evaluate the ERR metric for three cases of $r = m = \{3, 4, 5\}$ versus seven different purity values $p \in [\frac{1}{r-1} + 0.01, 1]$. For the data generation, we set $n_1 = 30 \times r$ (30 samples on each facet) and $n_2 = 10$ (10 samples within the simplex) for a total of $n = 30 \times r + 10$ samples. The average ERR and run times over 10 trials are reported in Figure 3. We observe the following:

- MV-Dual performs as well as MVIE and has significantly lower computational time.
- GFPI achieves perfect recovery of ground-truth factors for all purity levels in all cases. However, the run time of GFPI is significantly larger as it relies on solving mixed integer programs.
- Min-vol performs better than SNPA for purities less than one, but does not recover the ground-truth factors even when the SSC is satisfied.

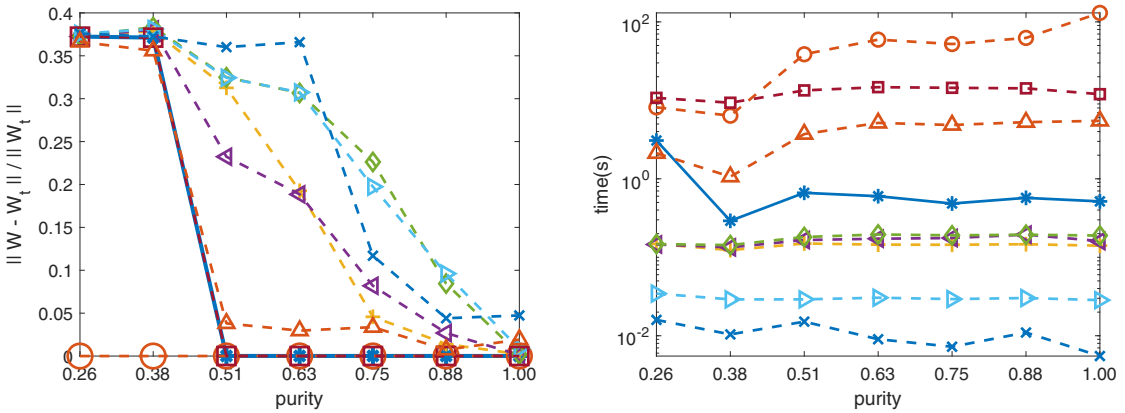
For low values of the purity, only GFPI performs perfectly. The reason is that the data does not satisfy the SSC, and there exist smaller volume solutions (but with less points on



(a) ERR for $r = m = 3$ (b) Time(s) for $r = m = 3$



(c) ERR for $r = m = 4$ (d) Time(s) for $r = m = 4$



(e) ERR for $r = m = 5$ (f) Time(s) for $r = m = 5$

Figure 3. *Average ERR metric and running time (in seconds) versus purity over* 10 *trials for noiseless data and different values of* r *and* m.

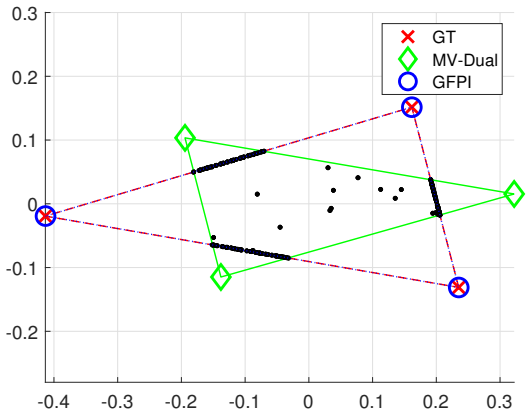


Figure 4. *MV-Dual versus GFPI in the case of low purity. GT stands for ground truth.*

their facets) that contain the data points. This is illustrated for a simple example for $r = 3$ in Figure 4, where the facet-based criterion used in GFPI finds the correct endmembers, whereas the volume-based MV-Dual selects the enclosing simplex with smaller volume.

Noisy data. We consider five SNRs, $\{60, 40, 30, 20, 10\}$, for $m = r = \{3, 4\}$ and generate synthetic ground-truth factors W_t and H_t identical to the previous noiseless experiment (with $n_1 = 30 \times r$ and $n_2 = 10$). The average ERR metrics over 10 trials are reported in Figures 5 and 6 and the average run times in Tables 1 ($r = 3$) and 2 ($r = 4$).

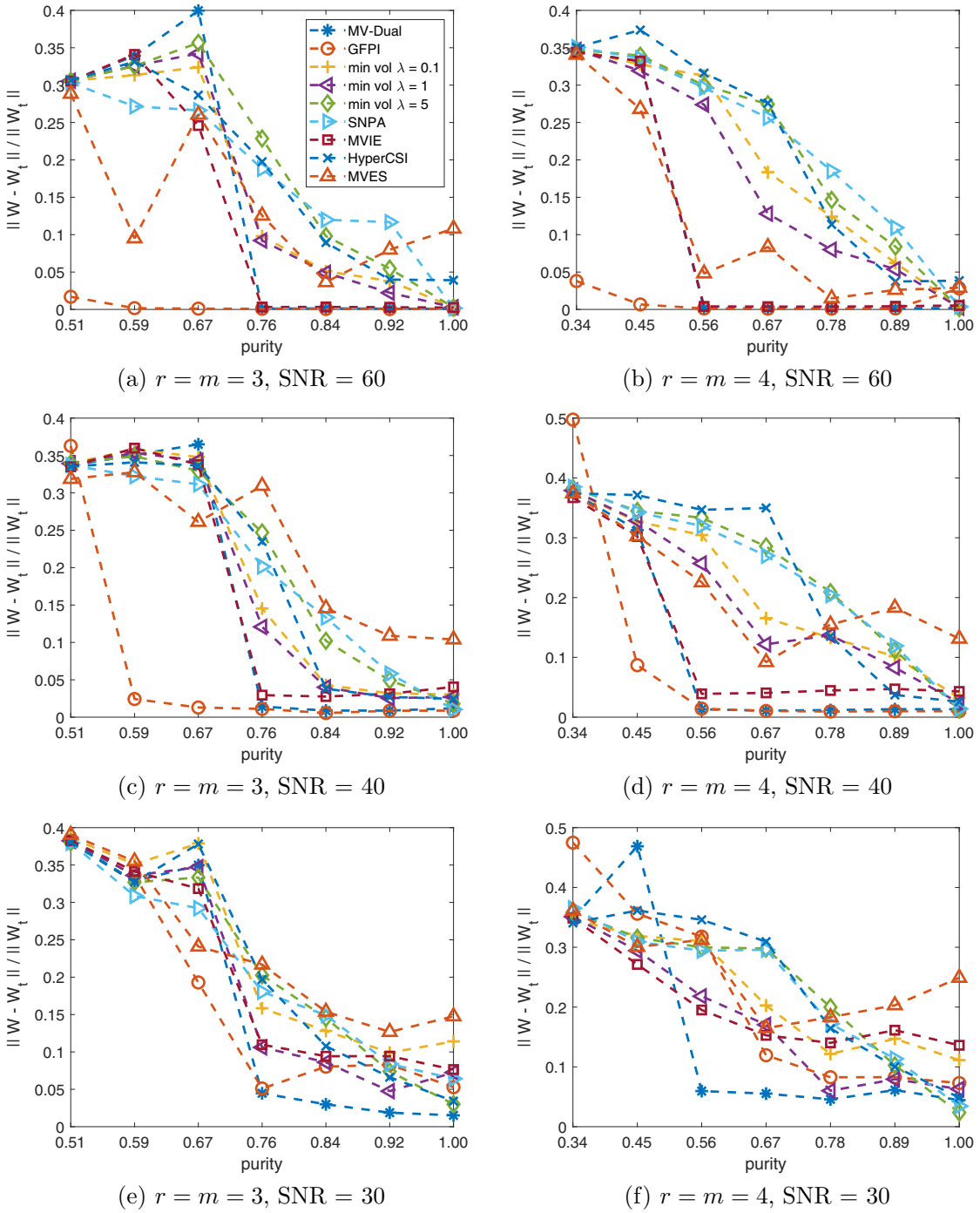
We observe as follows:

- GFPI is the most effective algorithm when the noise level is low, but it is the slowest.
- As the noise level increases, the performance of MVIE and GFPI gets worse. This indicates that MVIE and GFPI are more sensitive to noise. In fact, for high noise level and high purity, MV-Dual performs the best. Although we do not have a theoretical justification for this observation, our intuition is that GFPI learns the columns of Θ in a greedy manner: any mistake at any step will propagate at the next ones. On the other hand, MV-Dual learns all columns of Θ simultaneously, and hence there is no such propagation of the error. Hence in a setting where the SSC is satisfied (that is, for sufficiently high purity levels), MV-Dual performs better in higher noise regimes.
- MV-Dual is the second best algorithm in low noise regimes, and the most effective algorithm as the noise level increases. Moreover, MV-Dual is significantly faster than both MVIE and GFPI.

In the supplementary material (M165060_01.pdf [local/web 346KB]), we discuss the convergence of MV-Dual and sensitivity to the parameter λ. In a nutshell, the conclusions are as follows:

- MV-Dual requires a few updates of the translation vector v to converge, on average less than 10.
- MV-Dual is not too sensitive to the choice of λ.

6.2. Unmixing hyperspectral data. We apply SSMF algorithms for the unmixing problem on two real-world hyperspectral images: Samson and Jasper Ridge [33]. The goal is to identify

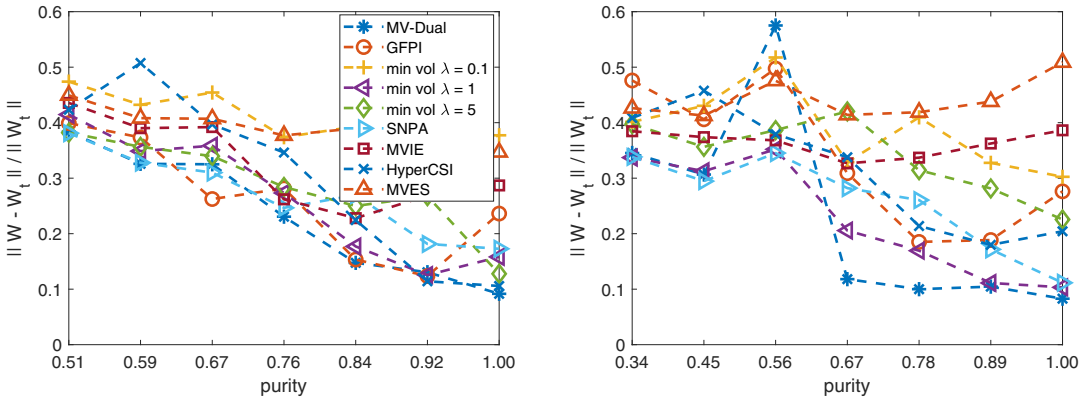


(a) $r = m = 3$, SNR $= 60$ (b) $r = m = 4$, SNR $= 60$

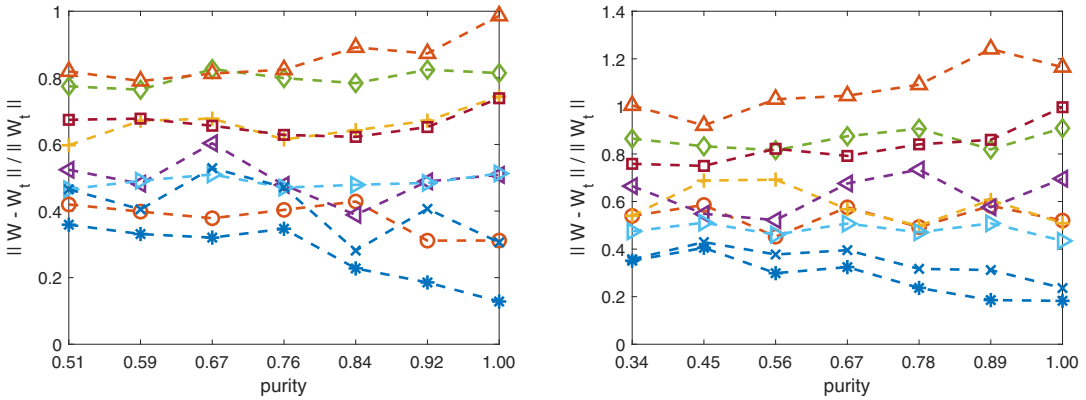
(c) $r = m = 3$, SNR $= 40$ (d) $r = m = 4$, SNR $= 40$

(e) $r = m = 3$, SNR $= 30$ (f) $r = m = 4$, SNR $= 30$

Figure 5. *Average ERR metric versus purity over* 10 *trials for noisy data and different values of* r, m *and SNR levels* 60, 40, 30.



(a) $r = m = 3$, SNR $= 20$ (b) $r = m = 4$, SNR $= 20$



(a) $r = m = 3$, SNR $= 10$ (b) $r = m = 4$, SNR $= 10$

Figure 6. *Average ERR metric versus purity over* 10 *trials for noisy data and different values of* r, m *and SNR levels* 20 *and* 10.

the so-called pure pixels (a.k.a. endmembers) which are the columns of W, while the weight matrix H contains the abundances of these pure pixels in the pixels of the image. To compare the performance, we use two metrics usually used in this literature:

- The mean removed spectral angle (MRSA) between two vectors $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^m$ is defined as

$$\mathrm{MRSA}(x, y) = \frac{100}{\pi} \cos^{-1}\left(\frac{(x - \bar{x}e)^\top (y - \bar{y}e)}{\|x - \bar{x}e\|_2 \|y - \bar{y}e\|_2}\right),$$

  where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x(i)$. The MRSA belongs to the interval [0,100] and corresponds an angle measure between $x - \bar{x}$ and $y - \bar{y}$. The MRSA equals 50 if $x - \bar{x}$ and $y - \bar{y}$ are orthogonal, and it equals 0 (resp., 100) if $x - \bar{x} = \alpha(y - \bar{y})$ for $\alpha > 0$ (resp., $\alpha < 0$). We will report the average MRSA between the columns of W (permuted to minimize that quantity) and W_t, and hence the smaller the average MRSA, the better.

Table 1

Average run times in seconds of SSMF algorithms on noisy synthetic data for $r = 3$.

SNR	MVDual	GFPI	min vol $\lambda = 0.1$	min vol $\lambda = 1$	min vol $\lambda = 5$	SNPA	MVIE	HyperCSI	MVES
10	0.92±0.19	3.39±0.48	0.08±0.01	0.12±0.01	0.15±0.02	0.01±0	3.69±0.21	0.01±0	0.13±0.02
20	0.78±0.06	4.69±0.54	0.08±0.01	0.08±0.01	0.09±0.01	0.01±0	3.93±0.06	0.01±0	0.16±0.02
30	0.56±0.11	7.76±3.51	0.12±0.01	0.13±0.01	0.14±0.02	0.01±0	5.28±0.23	0.01±0	0.30±0.04
40	0.45±0.06	4.18±1.12	0.10±0.01	0.11±0.01	0.13±0.01	0.01±0	4.96±0.12	0.01±0	0.30±0.05
60	0.42±0.06	1.47±0.45	0.07±0.01	0.08±0.01	0.09±0.01	0.01±0	3.78±0.12	0.01±0	0.26±0.07

Table 2

Average run times in seconds of SSMF algorithms on noisy synthetic data for $r = 4$.

SNR	MVDual	GFPI	min vol $\lambda = 0.1$	min vol $\lambda = 1$	min vol $\lambda = 5$	SNPA	MVIE	Hyper CSI	MVES
10	2.80±0.50	136±12	0.11±0.005	0.15±0.01	0.17±0.01	0.01±0	1.87±0.14	0.01±0	0.38±0.03
20	1.86±0.71	153±23	0.13±0.01	0.15±0.01	0.13±0.01	0.01±0	5.07±0.20	0.01±0	0.43±0.43
30	1.36±0.99	143±77	0.14±0.01	0.15±0.01	0.18±0.02	0.02±0	6.46±0.29	0.01±0	0.61±0.07
40	0.97±0.82	64.5±36.7	0.15±0.01	0.17±0.03	0.20±0.04	0.02±0	7.13±0.40	0.01±0	0.75±0.08
60	0.56±0.05	22.8±8.87	0.16±0.01	0.19±0.03	0.21±0.04	0.02±0	7.58±0.37	0.01±0	1.22±0.25

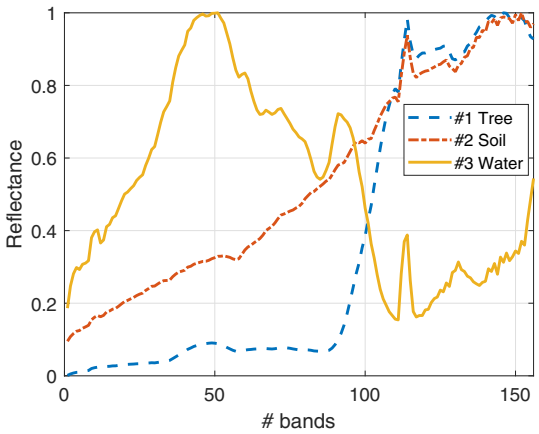
- The relative reconstruction error measures how well the data matrix is reconstructed using W and H:

$$\mathrm{RE} = \frac{\|X - WH\|_F}{\|X\|_F}.$$

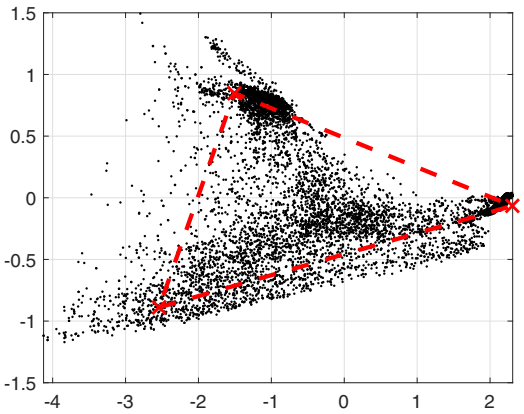
Samson data set. The Samson image has 95×95 pixels, each with 156 spectral bands, and contains three endmembers ($r = 3$): "soil," "water," and "tree" [33]. The solution obtained by MV-Dual is illustrated in Figure 7, which shows that MV-Dual is able to decompose the Samson image into its main three components. We compare the performance of MV-Dual to other SSMF algorithms in Table 3. We set $\lambda = 0.2$ in this experiment.

MV-Dual has the best MRSA, slightly better than Min-Vol, but has a larger computational time. This is expected as Min-Vol uses a specialized first-order algorithm for the optimization, whereas MV-Dual uses the generic *quadprog* method of MATLAB within each iteration of the optimization procedure. Moreover, MV-Dual has a higher relative error: this is expected since, as opposed to Min-Vol, it does not directly minimize this quantity. Moreover, although the relative error is a useful metric, it does not necessarily mean that the decomposition is meaningful for the application at hand. For example, a plain NMF algorithm would obtain a smaller error even faster (for this data set, 2.51% in 0.2 s), while the estimated W would be extremely poor as the NMF is never unique in this application (as W is dense).

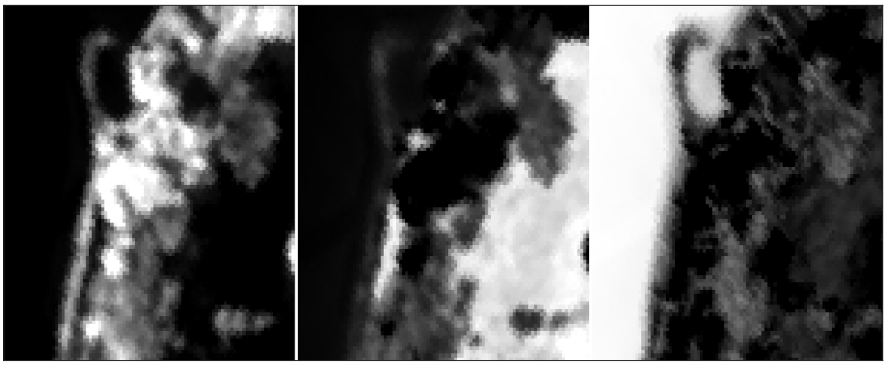
Jasper Ridge data set. The Jasper Ridge data set consists of 100×100 pixels with 224 spectral bands, with four endmembers ($r = 4$) in this image: "road," "soil," "water," and "tree" [33]. Similar to the Samson data set, we plot the extracted endmembers, projected fitted convex hull, and abundance maps obtained by MV-Dual in Figure 8, where we see that



(a) Spectral signatures of the estimated endmembers (columns of W) by MV-Dual.



(b) Two-dimensional projection of the data points (dots), and of the MV-Dual polytope.



(c) Abundance maps (rows of H) estimated by MV-Dual. From left to right: soil, tree and water.

Figure 7. *MV-Dual applied on the Samson hyperspectral image.*

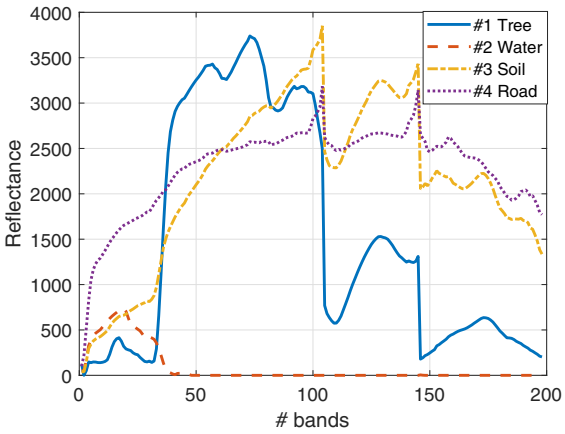
Table 3

*Comparing the performance of MV-Dual with state-of-the-art SSMF algorithms on Samson data set. Numbers marked with * indicate that the corresponding algorithms did not converge within* 100 *seconds. The best result is highlighted in bold.*

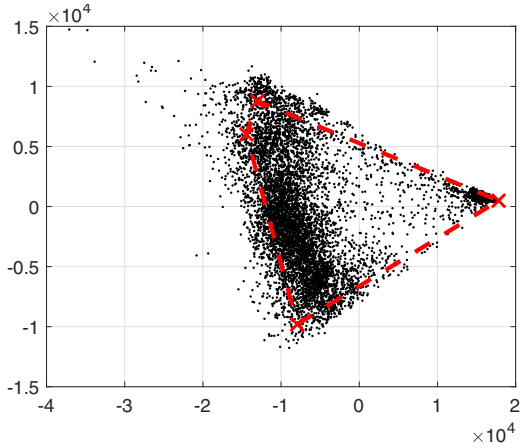
	SNPA	Min-Vol	HyperCSI	GFPI	MV-Dual
MRSA	2.78	2.58	12.91	2.97	**2.50**
$\frac{\|\|X-WH\|\|_F}{\|\|X\|\|_F}$	4.00%	**2.69%**	5.35%	4.02%	5.81%
Time (s)	**0.37**	1.30	0.90	100*	15.78

MV-Dual is able to decompose the image into its four main components. We set $\lambda = 0.15$ in this experiment. The detailed numerical comparison with other algorithms is reported in Table 4.

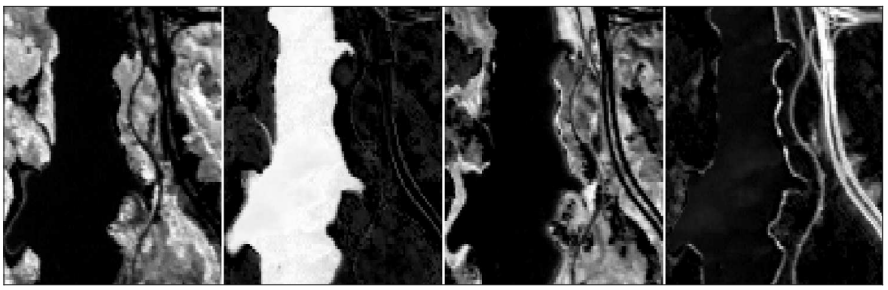
MV-Dual has the best performance in terms of MRSA, significantly smaller than Min-Vol. Its relative error is worse than Min-Vol, but very close (6.21% versus 6.09%). Again, the computational of MV-Dual is larger but reasonable.



(a) Spectral signatures of the estimated endmembers (columns of W).



(b) Two-dimensional projection of the data points (dots), and of the MV-Dual polytope.



(c) Abundance maps (rows of H) estimated. From left to right: road, tree, soil, water.

Figure 8. *MV-Dual applied on the Jasper Ridge hyperspectral image.*

Table 4

*Comparing the performances of MV-Dual with the state-of-the-art SSMF algorithms on the Jasper Ridge data set. Numbers marked with * indicate that the corresponding algorithms did not converge within* 100 *seconds. The best result is highlighted in bold.*

	SNPA	Min-Vol	HyperCSI	GFPI	MV-Dual
MRSA	22.27	6.03	17.04	4.82	**3.74**
$\frac{\|\|X-WH\|\|_F}{\|\|X\|\|_F}$	8.42%	**6.09%**	11.43%	6.47%	6.21%
Time (s)	**0.60**	1.45	0.88	100*	43.51

7. Conclusion.

SSMF is the problem of finding a set of points whose convex hull contains a given set of data points. To make the problem meaningful and identifiable, several approaches have been proposed, the two most popular ones being to (1) minimize the volume of the sought convex hull and (2) identify the facets of that convex hull by leveraging the fact that they should contain as many data points as possible (leading to sparse representations). In this paper, we have proposed a new approach to tackle SSMF by maximizing the volume of the polar of that convex hull. We showed that this approach also leads to identifiability under the same assumption as the minimum-volume approaches, namely, the SSC. However, the two

models are not equivalent, and our proposed maximum-volume approach is able to obtain more consistent solutions on synthetic data experiments, especially in high noise regimes, while having a low computational cost. We also showed that it provides competitive results to unmix real-world hyperspecral images.

Further work includes the following:

- The implementation of dedicated and faster algorithms, with convergence guarantees, to solve our min-max formulation (4.6).
- A strategy to tune λ automatically. In the paper, we used a fixed value of λ, but it would be possible to tune it, e.g., based on the relative error of the current solution.
- The design of more robust models, e.g., replacing the ℓ_2 norm–based SVD preprocessing and the minimization of the Frobenius norm of Δ in (5.1) by more robust norms, e.g., the componentwise ℓ_1 norm.
- Adapt the theory and model in the rank-deficient case, that is, when Assumption 3.1 is not satisfied: $\text{conv}(W)$ is not a simplex but a polytope in dimension d with more than $d+1$ vertices.

Acknowledgment. We are grateful to the anonymous reviewers, who carefully read the manuscript; their feedback allowed us to improve it significantly.

REFERENCES

[1] M. Abdolali and N. Gillis, *Simplex-structured matrix factorization: Sparsity-based identifiability and provably correct algorithms*, SIAM J. Math. Data Sci., 3 (2021), pp. 593–623.

[2] M. C. U. Araújo, T. C. B. Saldanha, R. K. H. Galvao, T. Yoneyama, H. C. Chame, and V. Visani, *The successive projections algorithm for variable selection in spectroscopic multicomponent analysis*, Chemom. Intell. Lab. Syst., 57 (2001), pp. 65–73.

[3] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu, *A practical algorithm for topic modeling with provable guarantees*, in Proceedings of the International Conference on Machine Learning, 2013, pp. 280–288.

[4] S. Arora, R. Ge, R. Kannan, and A. Moitra, *Computing a nonnegative matrix factorization–provably*, in Proceedings of the 44th Annual ACM Symposium on Theory of Computing, 2012, pp. 145–162.

[5] A. Bakshi, C. Bhattacharyya, R. Kannan, D. P. Woodruff, and S. Zhou, *Learning a latent simplex in input-sparsity time*, in Proceedings of the International Conference on Learning Representations, 2021.

[6] 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., 5 (2012), pp. 354–379.

[7] J. W. Boardman, F. A. Kruse, and R. O. Green, *Mapping target signatures via partial unmixing of AVIRIS data*, in Proceedings of the JPL Airborne Earth Science Workshop, 1995, pp. 23–26.

[8] M. Boroun, E. Yazdandoost Hamedani, and A. Jalilzadeh, *Projection-free methods for solving nonconvex-concave saddle point problems*, Adv. Neural Inf. Process. Syst., 36 (2024).

[9] E. J. Candès, X. Li, Y. Ma, and J. Wright, *Robust principal component analysis?*, J. ACM, 58 (2011), pp. 1–37.

[10] T.-H. Chan, C.-Y. Chi, Y.-M. Huang, and W.-K. Ma, *A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing*, IEEE Trans. Signal Process., 57 (2009), pp. 4418–4432.

[11] M. D. Craig, *Minimum-volume transforms for remotely sensed data*, IEEE Trans. Geosci. Remote Sens., 32 (1994), pp. 542–552.

[12] 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., 36 (2019), pp. 59–80.

[13] X. Fu, K. Huang, N. D. Sidiropoulos, Q. Shi, and M. Hong, *Anchor-free correlated topic modeling*, IEEE Trans. Pattern Anal. Mach. Intell., 41 (2018), pp. 1056–1071.
[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., 63 (2015), pp. 2306–2320.
[15] X. Fu, N. Vervliet, L. De Lathauwer, K. Huang, and N. Gillis, *Computing large-scale matrix and tensor decomposition with structured factors: A unified nonconvex optimization perspective*, IEEE Signal Process. Mag., 37 (2020), pp. 78–94.
[16] N. Gillis, *Successive nonnegative projection algorithm for robust nonnegative blind source separation*, SIAM J. Imaging Sci., 7 (2014), pp. 1420–1450.
[17] N. Gillis, *Nonnegative Matrix Factorization*, SIAM, Philadelphia, 2020.
[18] N. Gillis and A. Kumar, *Exact and heuristic algorithms for semi-nonnegative matrix factorization*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 1404–1424.
[19] N. Gillis and S. A. Vavasis, *On the complexity of robust PCA and ℓ_1-norm low-rank matrix approximation*, Math. Oper. Res., 43 (2018), pp. 1072–1084.
[20] D. C. Heinz and C.-I. Chang, *Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery*, IEEE Trans. Geosci. Remote Sens., 39 (2001), pp. 529–545.
[21] K. Huang and X. Fu, *Detecting overlapping and correlated communities without pure nodes: Identifiability and algorithm*, in Proceedings of the International Conference on Machine Learning, 2019, pp. 2859–2868.
[22] K. Huang, N. D. Sidiropoulos, and A. Swami, *Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition*, IEEE Trans. Signal Process., 62 (2013), pp. 211–224.
[23] D. D. Lee and H. S. Seung, *Learning the parts of objects by non-negative matrix factorization*, Nature, 401 (1999), pp. 788–791.
[24] V. Leplat, A. M. Ang, and N. Gillis, *Minimum-volume rank-deficient nonnegative matrix factorizations*, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 3402–3406.
[25] C.-H. Lin, C.-Y. Chi, Y.-H. Wang, and T.-H. Chan, *A fast hyperplane-based minimum-volume enclosing simplex algorithm for blind hyperspectral unmixing*, IEEE Trans. Signal Process., 64 (2015), pp. 1946–1961.
[26] 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., 53 (2015), pp. 5530–5546.
[27] C.-H. Lin, R. Wu, W.-K. Ma, C.-Y. Chi, and Y. Wang, *Maximum volume inscribed ellipsoid: A new simplex-structured matrix factorization framework via facet enumeration and convex optimization*, SIAM J. Imaging Sci., 11 (2018), pp. 1651–1679.
[28] W.-K. Ma, J. M. Bioucas-Dias, T.-H. Chan, N. Gillis, P. Gader, A. J. Plaza, A. Ambikapathi, and C.-Y. Chi, *A signal processing perspective on hyperspectral unmixing: Insights from remote sensing*, IEEE Signal Process. Mag., 31 (2013), pp. 67–81.
[29] L. Miao and H. Qi, *Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization*, IEEE Trans. Geosci. Remote Sens., 45 (2007), pp. 765–777.
[30] M. Razaviyayn, M. Hong, and Z.-Q. Luo, *A unified convergence analysis of block successive minimization methods for nonsmooth optimization*, SIAM J. Optim., 23 (2013), pp. 1126–1153.
[31] G. Tatli, and A. T. Erdogan, *Polytopic matrix factorization: Determinant maximization based criterion and identifiability*, IEEE Trans. Signal Process., 69 (2021), pp. 5431–5447.
[32] M. Udell, C. Horn, R. Zadeh, and S. Boyd, *Generalized low rank models*, Found. Trends Mach. Learn., 9 (2016), pp. 1–118.
[33] F. Zhu, *Hyperspectral Unmixing: Ground Truth Labeling, Datasets, Benchmark Performances and Survey*, preprint, arXiv:1708.05125, 2017.
[34] G. M. Ziegler, *Lectures on Polytopes*, Grad. Texts in Math. 152, Springer, New York, 2012.