Dual Simplex Volume Maximization for Simplex-Structured Matrix Factorization

Maryam Abdolali*
K.N.Toosi University (KNTU)
Tehran, Iran

Giovanni Barbarino† Nicolas Gillis‡
University of Mons
Mons, Belgium

Abstract

Simplex-structured matrix factorization (SSMF) is a generalization of nonnegative matrix factorization, 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.

Keywords: simplex-structured matrix factorization, matrix factorization, minimum volume, sparsity, polarity/duality, hyperspectral imaging

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., [31, 14] 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,

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

*Email: maryam.abdolali@kntu.ac.ir

†Email: giovanni.barbarino@umons.ac.be. GB acknowledges the support by the European Union (ERC consolidator, eLinoR, no 101085607). GB is member of the Research Group GNCS (Gruppo Nazionale per il Calcolo Scientifico) of INdAM (Istituto Nazionale di Alta Matematica).

‡Email: nicolas.gillis@umons.ac.be. NG acknowledges the support by the European Union (ERC consolidator, eLinoR, no 101085607).

where e is the vector of all ones of appropriate dimension. SSMF has several applications in machine learning with two prominent examples including unmixing hyperspectral images where $H(i,j)$ is the proportion/abundance of the ith material within the jth pixel [19, 6, 27], and topic modeling where where $H(i,j)$ is the contribution of the ith topic within the jth document [3, 12, 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 NMF which decomposes a nonnegative matrix, $X \geq 0$, as $X = WH$ where $W \geq 0$ and $H \geq 0$ [22, 16]. In fact, normalizing each column of X to have unit ℓ_1 norm, and assuming w.l.o.g. that the columns of W also have unit ℓ_1 norm, 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.

2.1 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 hull of W to contain more points, and hence obtain equivalent factorizations [17]. It is therefore crucial for SSMF models to include additional constraints or regularizers to obtain identifiable models. There has been three main approaches to achieve this goal: separability, volume minimization and facet-based identification. They are described in the next three sections.

2.2 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

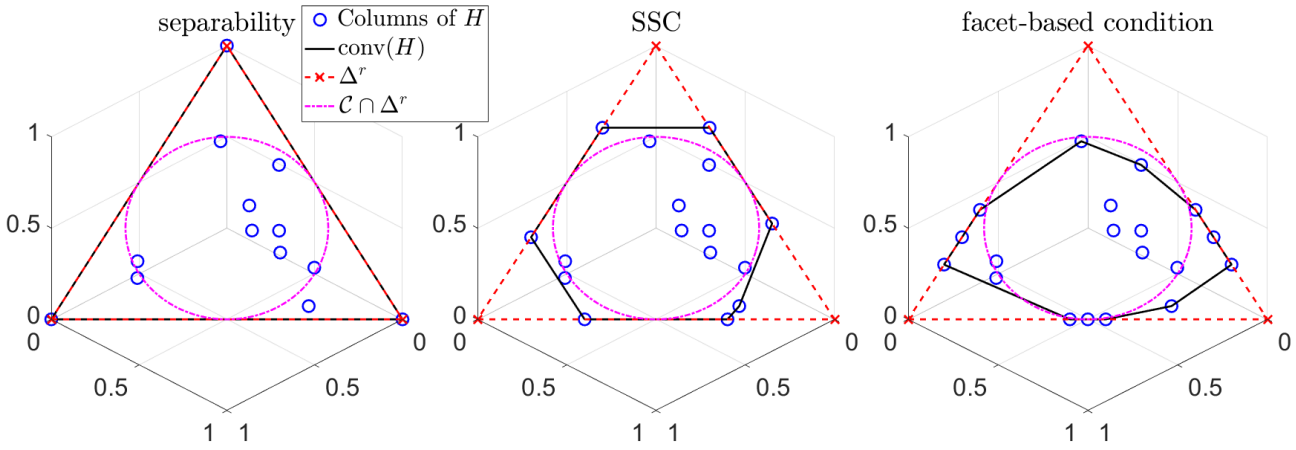


Figure 1: Comparison of separability (left), SSC (middle), and the facet-based condition (right) for the matrix H whose columns lie on Δ^r in the case $r = 3$. On the left, separability requires the columns of H to contain the unit vectors, that is, $H(:, \mathcal{K}) = I_r$ for some $\mathcal{K}$. On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

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 [16, Chapter 7] for a comprehensive survey on these algorithms. However, separability is a strong assumption which might not hold in all real-world scenarios.

2.3 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 and the origin within the column space of W, which is proportional to $\det\left(W^\top W\right)$, is minimized. The first intuitions and empirical evidences came from the hyperspectral imaging literature [10, 28]. Later, minimum-volume SSMF was shown to be identifiable [13, 25] under the so-called sufficiently scattered condition[1] (SSC) introduced in [21]:

Definition 1 (SSC). *The matrix $H \in \mathbb{R}^{r \times n}$ satisfies the SSC if its conic hull, defined by cone$(H) = \left\{y \mid y = Hx, x \geq 0\right\}$, contains the second-order cone $\mathcal{C} = \left\{x \in \mathbb{R}^r_+ \mid e^\top x \geq \sqrt{r-1}||x||_2\right\}$. Moreover, the only real orthogonal matrices $Q \in \mathbb{R}^{r \times r}$ satisfying* cone$(H) \subset$ cone(Q) *are permutation matrices.*

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

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 [11] and [16, Chapter 4] for more details.

Many algorithms have been designed using volume minimization, starting from [10]. A common approach is to minimize the volume of enclosing vertices W by minimizing the determinant of $W^\top W$ [13, 25]:

$$\min_{W,H} \det(W^\top W) \qquad \text{such that} \qquad X = WH \text{ and } H(:,j) \in \Delta^r \text{ for all } j. \tag{1}$$

Under the SSC, solving (1) guarantees to recover the columns of W and H in the SSMF $X = WH$, up to permutation [13, 25]. 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) [9] that attempts to simplify the problem by focusing on the volume of a dimension-reduced transformation of W (via the SVD), say $\bar{W} \in \mathbb{R}^{r \times r-1}$; see Section 3.1 for details. MVES works with the transformed matrix

$$\tilde{W} = [\bar{W}(:,1) - \bar{W}(:,r), \ldots, \bar{W}(:,r-1) - \bar{W}(:,r)] \in \mathbb{R}^{(r-1)\times(r-1)}, \tag{2}$$

and minimizes $|\det(\tilde{W})| = \text{vol}\big(\text{conv}(\bar{W})\big)$. 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 is considered in [30], where the columns of H are restricted to belong 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 (1), are not convex, and hence it is not straightforward to solving 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.

2.4 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 $\{x \mid a^\top x \leq b\}$ for some vector a and scalar b) enclosing a region where the columns of X lie. Two main algorithms in this category are the following:

- Minimum-volume inscribed ellipsoid (MVIE) [26] 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.

- Greedy facet-based polytope identification (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 (FBC) 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.

In the next 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. It is presented in the next section.

3 Proposed model: SSMF based on maximum-volume in the polar

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

If $\text{rank}(H) = r$, which is implied by the SSC condition, 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 2 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 [9] and GFPI [1], see also [27], 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 1.
- In the noiseless case, the preprocessing does not change the geometry and properties of the problem. In noisy settings, it filters noise via dimensionality reduction.
- 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. 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 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., [8, 18].

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 2.1, $\text{conv}(Y) \subseteq \text{conv}(P)$. Let us define the polar of a set.

Definition 2 (Polar)**.** *Given any set $\mathcal{S} \subseteq \mathbb{R}^d$, its polar, denoted $\mathcal{S}^*$, is defined as*

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

Polars have many interesting properties [33]. In particular,

- 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. By extension, given a matrix $P \in \mathbb{R}^{(r-1)\times r}$ whose columns define a simplex containing the origin in its interior, we will refer to its polar matrix as the matrix $\Theta \in \mathbb{R}^{(r-1)\times r}$ whose columns are the vertices of $\text{conv}(P)^*$
- 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 $\begin{bmatrix} Q \\ e^\top/\sqrt{r} \end{bmatrix}$ is an $r \times r$ orthogonal matrix, the polar of $\text{conv}(Q)$ is $\text{conv}(-rQ)$.

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

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

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

The constraint $\text{conv}(\Theta) = \text{conv}(P)^* \subseteq \text{conv}(Y)^*$ 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 $\text{conv}(\Theta)$ 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 $\text{conv}(P)$: Let $A = Y^\top\Theta \leq 1_{n\times r}$, then $A(i,j) = 1$ means that the ith data points, $Y(:,i)$, is 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 point 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 smallest 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 2.3. We therefore propose to solve the following model: Given $Y \in \mathbb{R}^{r-1\times n}$, solve

$$\max_{\Theta\in\mathbb{R}^{r-1\times r}} \text{vol}\big(\text{conv}(\Theta)\big) \quad \text{such that} \quad Y^\top\Theta \leq 1_{n\times r}. \tag{3}$$

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

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

Link with volume minimization Solving (3) is not equivalent to volume minimization in the primal (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

$$\min_{\hat{P}\in\mathbb{R}^{r-1\times r}, z\in\mathbb{R}^r} \text{vol}\big(\text{conv}(\hat{P})\big)\prod_{i=1}^r z_i \quad \text{such that} \quad \text{conv}(Y)\subseteq\text{conv}(\hat{P}), \quad \begin{bmatrix}\hat{P}\\ e^\top\end{bmatrix} z = e_r, \quad z\geq 0. \quad (4)$$

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 (4) differs from (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 constraint of $\hat{P}$ to be represented in the dual by a vector θ_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.

We will discuss in details how we handle (3) in Section 5, and how we can adapt it in the presence of noise. But first, we discuss the identifiability guarantees of solving (3).

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 condition in Definition 1 is tested on the matrix H, we can suppose from now on that, equivalently, X or Y has an SSC decomposition.

Fist 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). 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 condition forces any possible point of $\text{conv}(Y)^*$ 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 1. *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, $Pe = 0$, and H an $r \times n$ SSC and column stochastic matrix. Then*

$$\max_{\Theta \in \mathbb{R}^{r-1\times r}} \operatorname{vol}\big(\operatorname{conv}(\Theta)\big) \quad \textit{such that} \quad Y^\top \Theta \leq 1_{n\times r}. \tag{3}$$

is uniquely solved by the polar matrix of P.

Proof. Recall that the problem is equivalent to

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

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 $\begin{bmatrix} Q \\ e^\top/\sqrt{r} \end{bmatrix}$ is an $r \times r$ orthogonal matrix, since $Pe = 0$. Calling $\Psi = \Sigma U^\top \Theta$, the problem transforms into

$$\det(\Sigma)^{-1} \max_{\Psi \in \mathbb{R}^{r-1\times r}} \operatorname{vol}\big(\operatorname{conv}(\Psi)\big) \quad \text{such that} \quad \operatorname{conv}(\Psi) \subseteq \operatorname{conv}(QH)^*. \tag{5}$$

Since H is SSC, $\operatorname{conv}(QH) = Q\operatorname{conv}(H) \supset Q(\Delta^r \cap \mathcal{C})$ 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

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

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

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

Each element of the diagonal in R is bounded by r, so its trace is at most r^2. Since R is positive semi-definite, its determinant is bounded through the arithmetic and geometric means inequality (AM-GM) by $\det(R) \leq [\operatorname{tr}(R)/r]^r \leq r^r$, and the equality is attained if and only if $R = rI$ or equivalently when $\begin{bmatrix} \Psi/r \\ e^\top/\sqrt{r} \end{bmatrix}$ is orthogonal. The matrix $\Psi = -rQ$ thus attains the maximum possible volume and $\operatorname{conv}(-rQ) = \operatorname{conv}(Q)^* \subseteq \operatorname{conv}(QH)^*$, so it is also a solution of problem (5). All other Ψ with the same volume such that $\operatorname{conv}(\Psi) \subseteq \sqrt{r(r-1)}B^{r-1}$ are rotated versions of $-rQ$, that is, $\hat{\Psi} = -rVQ$, where V is orthogonal, but

$$\operatorname{conv}(-rVQ) = \operatorname{conv}(VQ)^* \subseteq \operatorname{conv}(QH)^* \implies \operatorname{conv}(QH) \subseteq \operatorname{conv}(VQ)$$

$$\implies \operatorname{conv}(H) \subseteq \operatorname{conv}\left(\begin{bmatrix} Q^\top & e/\sqrt{r} \end{bmatrix}\begin{bmatrix} V & \\ & 1 \end{bmatrix}\begin{bmatrix} Q \\ e^\top/\sqrt{r} \end{bmatrix}\right) = \operatorname{conv}(\Pi),$$

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 (5) are $-rQ$ and its permuted versions, or also said all possible polar matrices of Q. Tracing back to the original problem, we find that all possible solutions of (3) are the polar matrices of $\Theta^* = (U\Sigma^{-1}\Psi)^* = U\Sigma Q = P$. □

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) 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) 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) 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 2. *Let Y be a $(r-1) \times n$ 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 4.2). If 0 is in the interior of $\mathrm{conv}(Y)$, then*

$$\max_{\Theta \in \mathbb{R}^{r-1 \times r}} \mathrm{vol}\big(\mathrm{conv}(\Theta)\big) \quad \textit{such that} \quad Y^\top \Theta \leq 1_{n \times r}. \tag{3}$$

is uniquely solved by the polar matrix of P.

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

Notice that in the separable case, $v = Xe/r$ is already a good choice for the translation vector. In fact, under Assumption 1, it can be proved that v is in the interior of $\mathrm{conv}(X) = \mathrm{conv}(W)$.

4.3 Between SSC and separability: η-expanded

We have seen that for a 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.

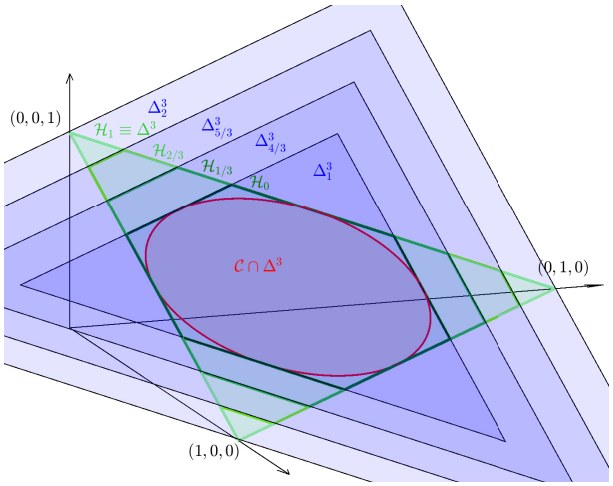
Definition 3. *We say that $H \in \mathbb{R}^{r \times n}$ is η-expanded with $\eta \in [0,1]$ if*

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

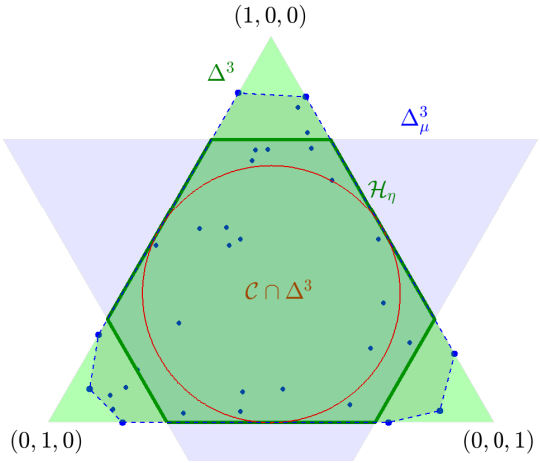
Suppose that $X \in \mathbb{R}^{m \times n}$ has rank $r-1$ and admits a decomposition $X = WH$ where H is column stochastic and η-expanded. The following properties are easily shown:

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

In other words, 0-expansion is close to the SSC, and the property of being η-expanded bridges between SSC and separability. The set $\mathcal{H}_\eta$ can be described also as the intersection of Δ^r and Δ^r_μ obtained by



(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 3 and 2 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 a η-expanded H.

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 we can see in Figure 2. In formulae,

$$\mathcal{H}_\eta = \Delta^r \cap \Delta^r_\mu = \operatorname{conv}(I) \cap \operatorname{conv}\left(\frac{\mu+1}{r} ee^\top - \mu I\right). \tag{6}$$

In case of SSC, Theorem 1 tells us that the only certified good translation vector is $v = We/r$. Instead, in case of separability, Theorem 2 tells us that all vectors inside the interior part 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. To do so, we first need two lemmas that show how the polar duality behaves under translation of the polytopes and how to compute the volume of the polar matrix after such translation. We provide the proofs in Appendix 8.1. From now on, we use $\mathcal{S}^\circ$ to indicate the interior of a set $\mathcal{S}$.

Lemma 1. *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 $\Theta_w \operatorname{Diag}(z_w)$ does not depend on w and $\Theta_w = \Theta \operatorname{Diag}(e - \Theta^\top w)^{-1}$.*

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

Lemma 2. *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,

$$\operatorname{vol}\big(\operatorname{conv}(\Theta N)\big) = |\det(N)| \left|\frac{e^\top N^{-1} w}{e^\top w}\right| \operatorname{vol}\big(\operatorname{conv}(\Theta)\big). \tag{7}$$

Now we can state and prove our result.

Theorem 3. *Suppose that $Y = PH$ with H η-expanded and column stochastic and P full rank. Consider the vector q such that $Pq = 0$ and $e^\top q = 1$. If $0 < q < \frac{r\eta+2(1-\eta)}{2r} e$, then the problem*

$$\max_{\Theta \in \mathbb{R}^{r-1\times r}} \operatorname{vol}\big(\operatorname{conv}(\Theta)\big) \quad \text{such that} \quad Y^\top \Theta \leq 1_{n\times r}. \tag{3}$$

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 $\begin{bmatrix} Q \\ e^\top/\sqrt{r} \end{bmatrix}$ is an $r\times r$ orthogonal matrix, since $P_v e = 0$. Similarly as the proof of Theorem 1, problem (3) is equivalent to

$$\det(\Sigma)^{-1} \max_{\Psi \in \mathbb{R}^{r-1\times r}} \operatorname{vol}\big(\operatorname{conv}(\Psi)\big) \quad \text{such that} \quad \operatorname{conv}(\Psi) \subseteq \operatorname{conv}(Q(I - qe^\top)H)^*. \tag{8}$$

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 $\operatorname{conv}(\Psi) \subseteq (Q(I-qe^\top)\mathcal{H}_\eta)^* = \operatorname{conv}\big((Q(I-qe^\top)\Delta^r)^* \cup (Q(I-qe^\top)\Delta^r_\mu)^*\big)$ where Δ^r_μ is defined in (6) 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

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

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

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

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 $\{v_1, \ldots, v_r\}$:

1. $v_i \in \{1, \alpha_i\}m_i$ for all i

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

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 1 says that $Mq = 0$. As a consequence, using (7), 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 $|S|(\mu+1) \le 2r$ and S is not empty, we find that $V_1 < \operatorname{vol}(\operatorname{conv}(M))$. For $|S|(\mu+1) > 2r$, we have

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

but thanks to 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 $V_1 < \operatorname{vol}(\operatorname{conv}(M))$.

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

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

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 1 we find that $M = -Q\operatorname{Diag}(q)^{-1}$, 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 (7),

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

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

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

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 $V_2 < \operatorname{vol}(\operatorname{conv}(M))$.

The polytope with the biggest volume inside of $(Q(I - qe^\top)\mathcal{H}_\eta)^*$ thus coincides with the polar of $\operatorname{conv}(Q(I-qe^\top))$ that is in particular contained in $\operatorname{conv}(Q(I-qe^\top)H)^*$. The matrices Ψ describing the polar of $Q(I - qe^\top)$ are therefore the unique solutions to (8). Going back to the original problem, we see that it is solved uniquely by $\Theta = U\Sigma^{-1}\Psi$ being the polar of $U\Sigma Q(I - qe^\top) = P_v(I - qe^\top) = P$. $\square$

When X is separable, that is, H is 1-expanded, Theorem 2 says that the only condition needed for the correctness of the solution of the problem 3 is $v = Wq$, where q has sum 1 and $0 < q < e$. In this case, though, Theorem 3 only holds for $0 < q < e/2$. This suggests that the result can be improved.

Conjecture 1. *The thesis of Theorem 3 holds if $q_i > \frac{1-\eta}{r}$ for every i.*

4.4 Min-max approach under the SSC

Under the SSC condition, we have proved that the solution to problem (3) 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), 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 that the solution Θ_w of (3) 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. *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*

$$\mathcal{V}(w) := \sup_{\Theta \in \mathbb{R}^{r-1 \times r}} \operatorname{vol}\big(\operatorname{conv}(\Theta)\big) \quad \textit{such that} \quad (Y - we^\top)^\top \Theta \leq 1_{n \times r}, \tag{9}$$

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

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

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 a r-subset of the s columns of Ψ_w, as we show in Lemma 3 in the Appendix. The maximum is thus achieved by one out of $\binom{s}{r}$ simplices $\Theta_w^{(i)}$, and we can recast the problem as

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

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 1,

$$\mathcal{V}_i(w) := \operatorname{vol}(\operatorname{conv}(\Theta_w^{(i)})) = \operatorname{vol}(\operatorname{conv}(\Theta_\ell^{(i)}) \operatorname{Diag}(e - (\Theta_\ell^{(i)})^\top (w - \ell))^{-1}) = \frac{\operatorname{vol}(\operatorname{conv}(\Theta_\ell))}{\prod_j [e - (\Theta_\ell^{(i)})^\top (w - \ell)]_j}.$$

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 $w_1, w_2 \in \operatorname{conv}(Y)^\circ$,

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

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 (9), so by Lemma 1 and (7),

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

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

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

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

Theorem 4 and Theorem 1 imply the following.

Corollary 1. *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*

$$\inf_{w \in \mathbb{R}^{r-1}} \sup_{\Theta \in \mathbb{R}^{r-1 \times r}} \operatorname{vol}\big(\operatorname{conv}(\Theta)\big) \quad \text{such that} \quad (Y - we^\top)^\top \Theta \leq 1_{n \times r}, \tag{10}$$

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

Corollary 1 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, while 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 2), while the set of v's that lead to identifiability typically contains more than the point Pe/r, as shown in Theorem 3 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

Let us first assume that the translation vector, v, is fixed. In the presence of noise, we propose to consider the following formulation:

$$\max_{Z,\Theta,\Delta} \quad \det(Z)^2 - \lambda\|\Delta\|_F^2 \quad \text{such that} \quad Z = \begin{bmatrix} \Theta \\ e^\top \end{bmatrix} \text{ and } \quad Y^\top\Theta \leq 1_{n\times r} + \Delta. \tag{11}$$

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 $\text{conv}(\Theta)$ 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 [20], we use the block successive upperbound minimization (BSUM) framework [29] 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 k-th 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 j-th row and k-th column from Z. If we fix all columns of Z but the kth, we have

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

For simplicity, let us denote $c = f^{(k)}$ and $x = Z(:,k)$. We want to maximize $f(x) = \det(Z)^2 = \left({f^{(k)}}^\top Z(:,k)\right)^2 = (c^\top x)^2$. 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,

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

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 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 $k = 1, .., r$):

$$\max_{t,\Theta(:,k),\Delta(:,k)} \alpha{f^{(k)}}^\top t - \lambda||\Delta(:,k)||_2^2 \quad \text{such that} \quad t = \begin{bmatrix} \Theta(:,k) \\ 1 \end{bmatrix} \text{ and } Y^\top\Theta(:,k) \leq 1_{(r-1)\times 1} + \Delta(:,k). \tag{12}$$

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. Consequently, we add the constraint to the problem above $\Theta(:,k) = -\sum_{i\neq k}\alpha_i\Theta(:,i)$ with $\alpha_i \geq \epsilon$ for some small $\epsilon > 0$. We will use $\epsilon = 0.01$.

Similar to [20], 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 details 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 conv(W).

2. v_0 is the average of the vertices extracted by SNPA, an effective separable NMF algorithm. This approach is less sensitive to imbalanced distributions within conv(W).

Since the optimal vector $v = We/r$ leads to the smallest volume solution (Theorem 4), we resort to a min-max approach: once (12) is solved and a solution Θ is obtained, 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 $\Theta(:, j)\hat{W}(:, k) = 1$ for $j \neq k$ and then let $\tilde{W}(:, k) = U\hat{W}(:, k) + v$. Then the new translation vector v is chosen as $\tilde{W}e/r$ which minimizes the volume of Θ.

Mitigating sensitivity to initialization Our numerous numerical experiments have shown that solving the optimization problem in (12) is usually not too sensitive to the initialization. However when there exist two of 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 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 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 (11) by alternatively optimizing (12) 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 –we will solve real instances of (11) with $n = 10^4$ in 15 seconds (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 a $\mathcal{O}(r^3)$ complexity. In the noisy case, only a few entries of $\Delta(:, k)$ will be non-zero, namely the entries corresponding to data points outside the hyperplane defined by $\Theta(:, k)$. Further research include the design of a dedicated solver to tackle (12), e.g., using an active-set approach.

Note that in step 8 of Algorithm 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 (12), or a maximum number of 100 iterations.

6 Numerical experiments

In this section, we present numerical experiments to show the efficiency of the proposed MV-Dual algorithm under various settings and conditions. All experiments are implemented in Matlab (R2019b), and run on a laptop with 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 the performance of MV-Dual to six state-of-the-art algorithms:

Algorithm 1 Maximum Volume in the Dual (MV-Dual)

Input: 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).

Output: 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 $Y = U^\top \left(X - v_0 e^\top\right) = U^\top X - U^\top v_0 e^\top$,
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 (11) via alternating optimization to update Z_i and Θ_i.

  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

- Successive nonnegative projection algorithm (SNPA) [15] is based on *separability* assumption and presents a robust extension to the successive projection algorithm (SPA) [2, 15] 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 [23]:

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

  This problem is optimized based on a block coordinate descent approach using the fast gradient method. The parameter λ is chosen as in [23]: $\tilde{\lambda} \frac{\|X - W_0 H_0\|_F^2}{\operatorname{logdet}(W_0^\top W_0 + \delta I_r)}$ 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 [23].

- Minimum-Volume Enclosing Simplex (MVES) [9] 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).

- Maximum volume inscribed ellipsoid (MVIE) [26] inscribes a maximum volume ellipsoid in the convex hull of the data points to identify the facets of conv(W).

- Hyperplane-based Craig-simplex-identication (HyperCSI) [24]: HyperCSI 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.
- Greedy facet based polytope identification (GPFI) [1] has the weakest conditions to recover the unique decomposition among the stat-of-the-art methods. This approach sequentially extracts the facets with largest number of points by solving a computationally expensive mixed integer program.

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

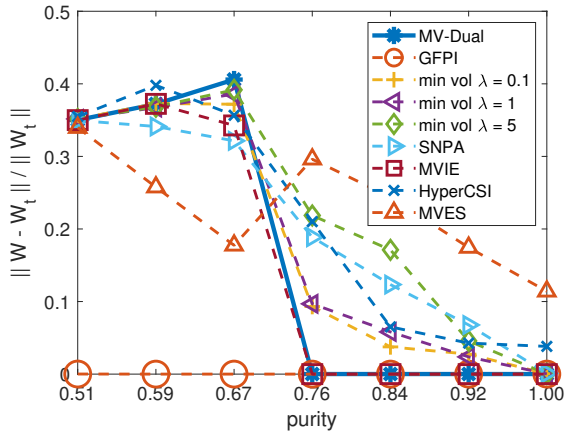
In this section, we compare the SSMF algorithms on noiseless and noisy synthetic data sets.

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 non-zero 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 condition is satisfied for smaller values of purity values [26]. 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:

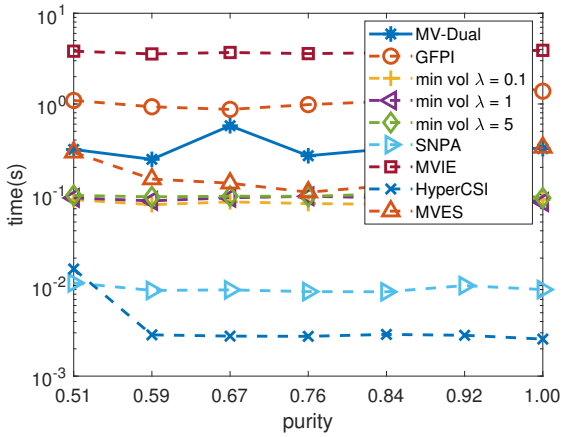
$$\text{variance} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} X(i,j)^2}{10^{SNR/10} \times m \times n}.$$

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 $10, 1, 0.5$ for SNR values of $60, 40, 30$, respectively.

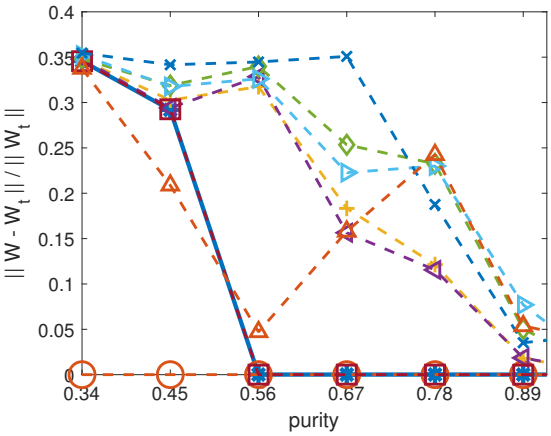
Noiseless data First, we compare the performances for different values of purity parameters p in the noiseless case. Due to the randomness of the data generation process, the reported results are the average over 10 trials. We evaluate the ERR metric for three cases of $r = m = \{3, 4, 5\}$ vs 7 different 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 running times over 10 trails are reported in Figure 3. We observe that:



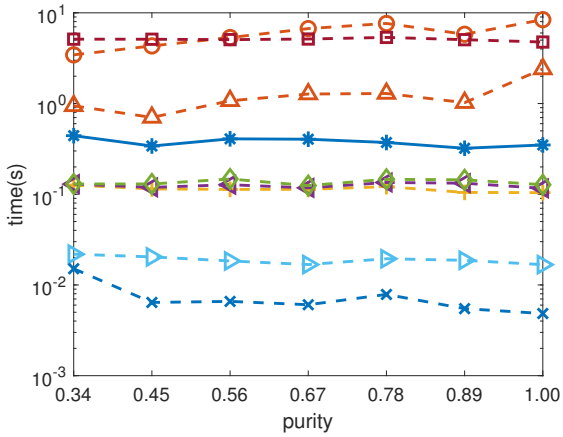
(a) ERR for $r = m = 3$



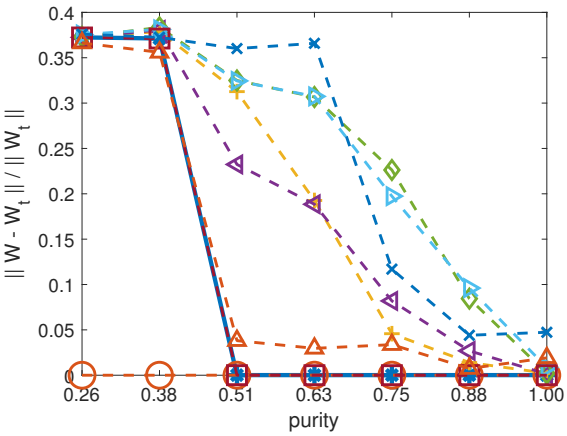
(b) Time(s) for $r = m = 3$



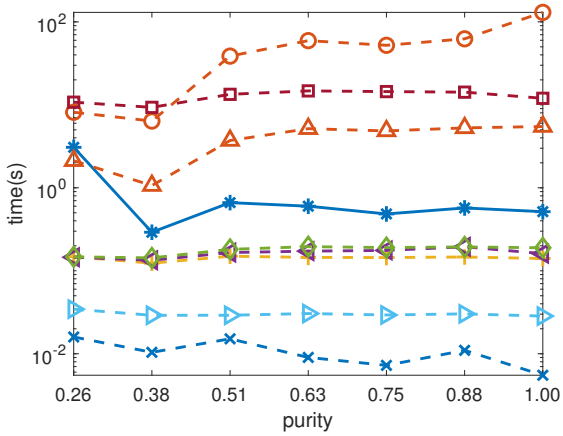
(c) ERR for $r = m = 4$



(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)vs purity over 10 trials for noiseless data and different values of r and m.

- 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 condition 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 exists smaller volume solutions (but with less points on their facets) that contain the data points. This is illustrated for a simple example for $r = 3$ in Fig 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.

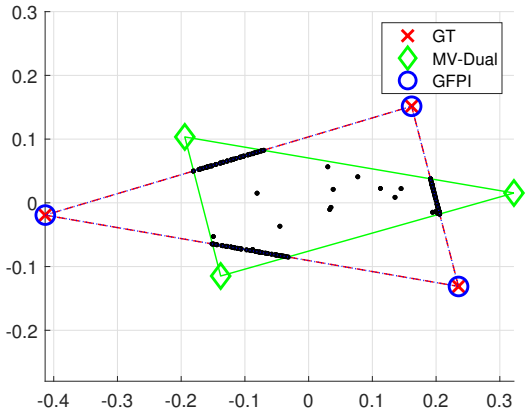


Figure 4: MV-Dual vs GFPI in the case of low-purity. GT stands for ground truth.

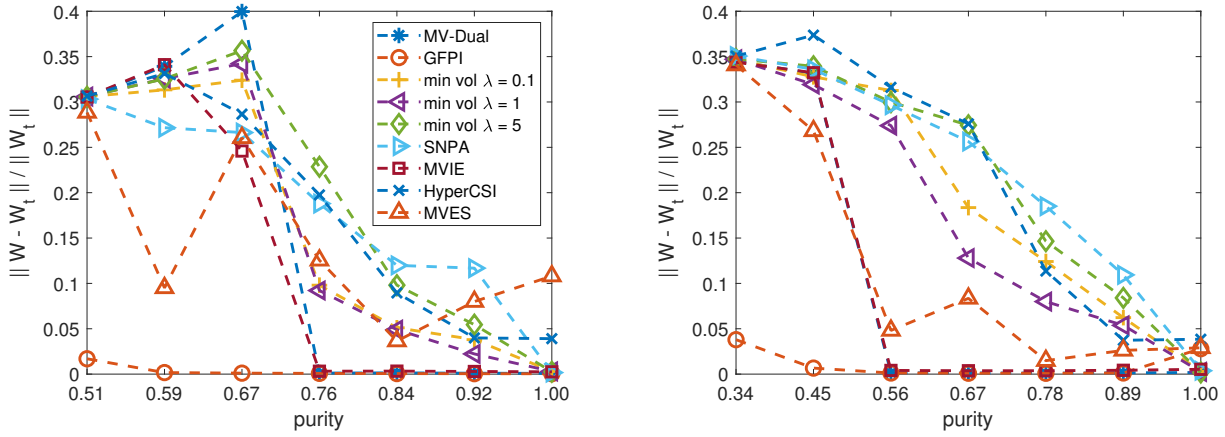
Noisy data We consider three SNRs, in {60, 40, 30}, for $m = r = \{3, 4\}$ and generate synthetic ground-truth factors W_t and H_t identical to the previous noiseless experiment (with $n_1 = 30 \times r$ and $n_2 = 10$). The average ERR metric and running time over 10 trials are reported in Figure 5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

SNR	MVDual	GFPI	min vol $\lambda = 0.1$	min vol $\lambda = 1$	min vol $\lambda = 5$	SNPA	MVIE	HyperCSI	MVES
30	0.56±0.11	7.76±3.51	0.12±0.01	0.13±0.01	0.14±0.02	0.01±0.001	5.28±0.23	0.01±0.004	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±0.00	4.96±0.12	0.005±0.004	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±0.00	3.78±0.12	0.001±0.00	0.26±0.07

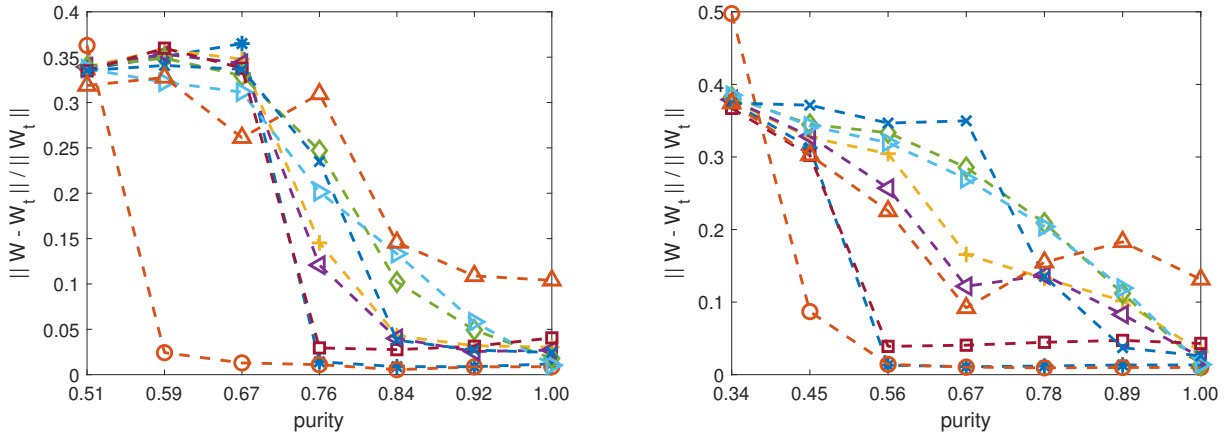
Table 1: Average run times in seconds of SSMF algorithms on noisy synthetic data for $r = 3$.

We observe that:

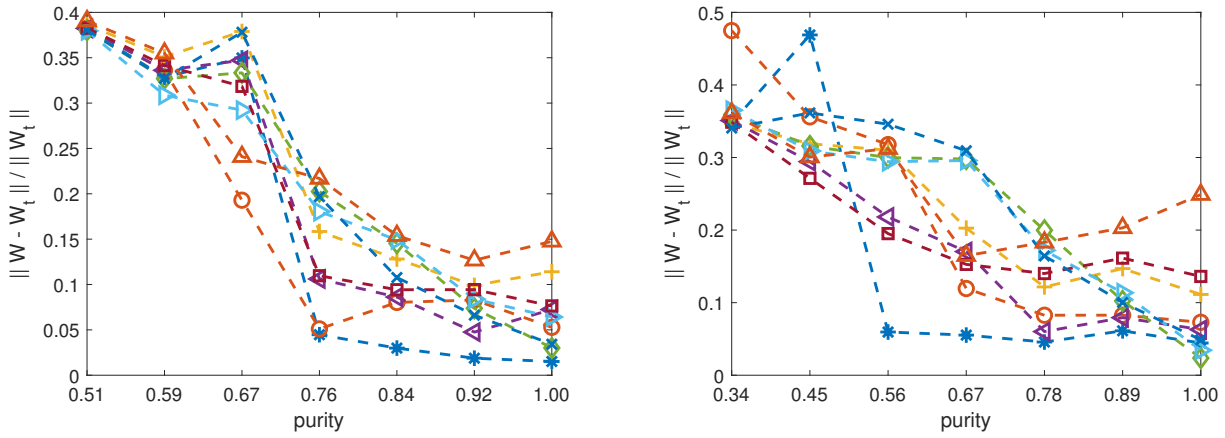
- GFPI is the most effective algorithm when the noise level is low, but it is the slowest.



(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 vs purity over 10 trials for noisy data and different values of r, m and SNR levels.

SNR	MVDual	GFPI	min vol $\lambda = 0.1$	min vol $\lambda = 1$	min vol $\lambda = 5$	SNPA	MVIE	HyperCSI	MVES
30	1.36±0.99	143.24±76.91	0.14±0.01	0.15±0.01	0.18±0.02	0.02±0.002	6.46±0.29	0.01±0.004	0.61±0.07
40	0.97±0.82	64.52±36.69	0.15± 0.01	0.17± 0.03	0.20± 0.04	0.02± 0.003	7.13± 0.40	0.01±0.007	0.75±0.08
60	0.56±0.05	22.79±8.87	0.16±0.01	0.19±0.03	0.21±0.04	0.02±0.01	7.58±0.37	0.01±0.01	1.22±0.25

Table 2: Average run times in seconds of SSMF algorithms on noisy synthetic data for $r = 4$.

- As the noise level increases, the performances 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.
- 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 Appendix 8.2, 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.

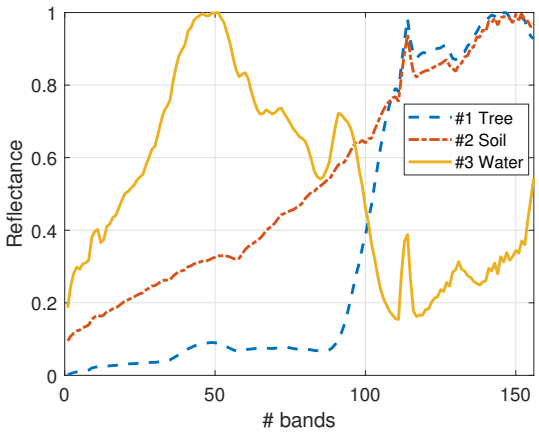
6.2 Unmixing hyperspectral data

We apply SSMF algorithms for the unmixing problem on two real-world hyperspectral images: Samson and Jasper Ridge [32]. The goal is to identify 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:

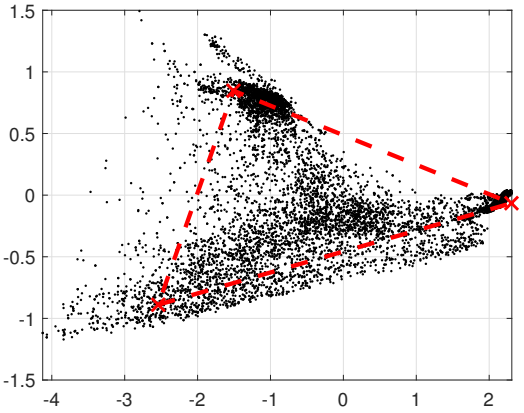
- Mean Removed Spectral Angle (MRSA) between two vectors $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^m$ is defined as
$$\text{MRSA}(x, y) = \frac{100}{\pi} \cos^{-1}\left(\frac{(x - \bar{x}e)^\top (y - \bar{y}e)}{\|x - \bar{x}e\|_2 \|y - \bar{y}e\|_2}\right),$$
where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x(i)$. We will report the average MRSA between the columns of W (permuted to minimize that quantity) and W_t.
- Relative Reconstruction Error (RE): measures how well the data matrix is reconstructed using W and H.
$$\text{RE} = \frac{\|X - WH\|_F}{\|X\|_F}.$$

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" [32]. The solution obtained by MV-Dual is illustrated in Figure 6. We compare the performance of MV-Dual to other SSMF algorithms in Table 3. We set $\lambda = 0.002$ in this experiment.

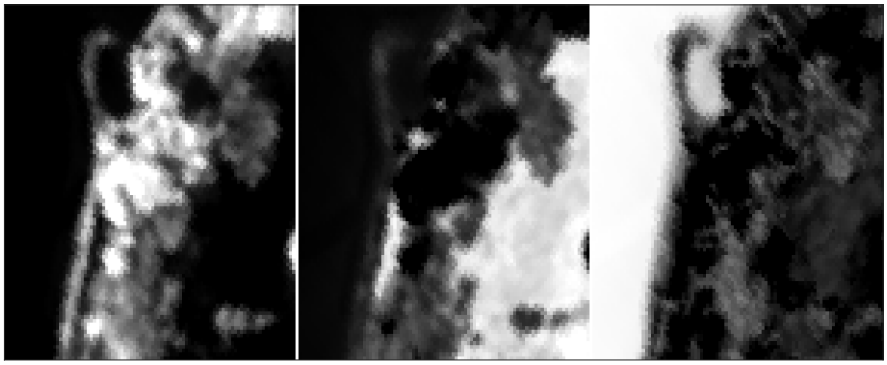
MV-Dual has the best MRSA, slightly better than Min-Vol, and has a larger computational time. This is expected as Min-Vol uses specialized first-order algorithm for the optimization, whereas MV-Dual uses the generic *quadprog* method of Matlab within each iteration of the optimization procedure. Moreover, MV-Dual has a higher relative error: this is expected since, as opposed to Min-Vol, it does not directly minimize this quantity.



(a) Spectral signatures of the estimated endmembers.



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



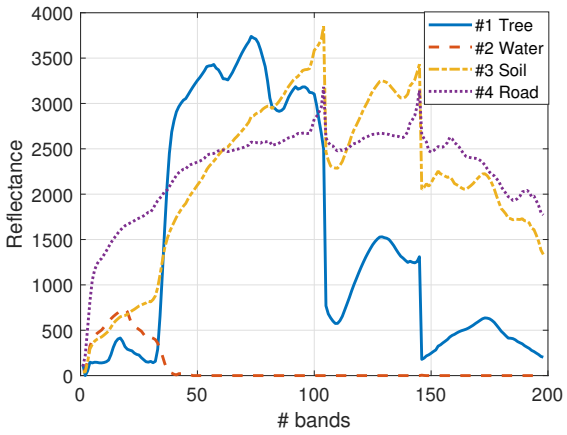
(c) Abundance maps estimated by MV-Dual. From left to right: soil, tree and water.

Figure 6: MV-Dual applied on the Samson hyperspectral image.

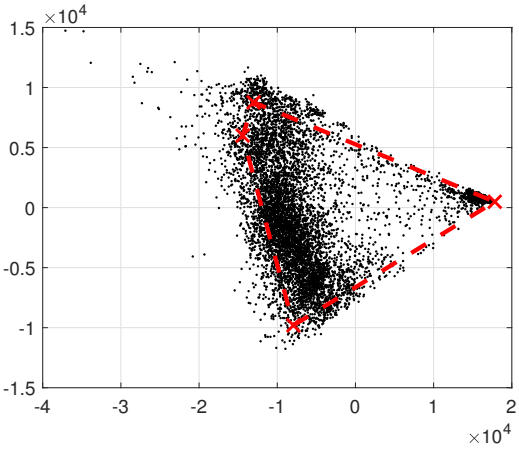
Table 3: Comparing the performances of MV-Dual with state-of-the-art SSMF algorithms on Samson data set. Numbers marked with * indicate that the corresponding algorithms did not converge within 100 seconds.

	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

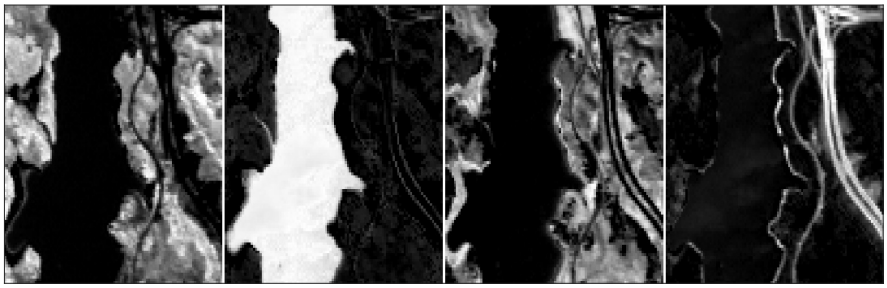
Jasper-ridge data set The Jasper Ridge data set consists of 100×100 pixels with 224 spectral bands, with four endmembers ($r = 4$) in this image: "road", "soil", "water" and "tree" [32]. Similar to the Samson data set, we plot the extracted endmembers, projected fitted convex hull and abundance maps obtained by MV-Dual in Figure 7. We set $\lambda = 0.0015$ in this experiment. The detailed numerical comparison with other algorithms is reported in Table 4.



(a) Spectral signatures of the estimated endmembers.



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



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

Figure 7: 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 Jasper-Ridge data set. Numbers marked with * indicate that the corresponding algorithms did not converge within 100 seconds.

	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

The conclusions are similar as for the previous data set: MV-Dual has the best performance in terms of MRSA, here significantly smaller than Min-Vol, while the relative error is worse, but very close, to that of Min-Vol, and the computational is larger but reasonable.

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 sufficiently scattered condition (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 include

- The implementation of dedicated and faster algorithms, with convergence guarantees, to solve our min-max formulation (10).
- 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 (11) by more robust norms, e.g., the component-wise ℓ_1 norm.
- Adapt the theory and model in the rank-deficient case, that is, when Assumption 1 is not satisfied: $\text{conv}(W)$ is not a simplex but a polytope in dimension d with more than $d+1$ vertices.

References

[1] Abdolali, M., Gillis, N.: Simplex-structured matrix factorization: Sparsity-based identifiability and provably correct algorithms. SIAM Journal on Mathematics of Data Science **3**(2), 593–623 (2021)

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

[3] Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., Zhu, M.: A practical algorithm for topic modeling with provable guarantees. In: International Conference on Machine Learning, pp. 280–288 (2013)

[4] Arora, S., Ge, R., Kannan, R., Moitra, A.: Computing a nonnegative matrix factorization–provably. In: Proceedings of the forty-fourth annual ACM symposium on Theory of Computing, pp. 145–162 (2012)

[5] Bakshi, A., Bhattacharyya, C., Kannan, R., Woodruff, D.P., Zhou, S.: Learning a latent simplex in input-sparsity time. In: International Conference on Learning Representations (ICLR) (2021)

[6] Bioucas-Dias, J.M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., Chanussot, J.: Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **5**(2), 354–379 (2012)

[7] Boardman, J.W., Kruse, F.A., Green, R.O.: Mapping target signatures via partial unmixing of AVIRIS data. In: Proc. Summary JPL Airborne Earth Science Workshop, Pasadena, CA, pp. 23–26 (1995)

[8] Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? Journal of the ACM (JACM) **58**(3), 1–37 (2011)

[9] Chan, T.H., Chi, C.Y., Huang, Y.M., Ma, W.K.: A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing. IEEE Trans. Signal Process. **57**(11), 4418–4432 (2009)

[10] Craig, M.D.: Minimum-volume transforms for remotely sensed data. IEEE Trans. Geosci. Remote Sens. **32**(3), 542–552 (1994)

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

[12] Fu, X., Huang, K., Sidiropoulos, N.D., Shi, Q., Hong, M.: Anchor-free correlated topic modeling. IEEE transactions on pattern analysis and machine intelligence **41**(5), 1056–1071 (2018)

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

[14] Fu, X., Vervliet, N., De Lathauwer, L., Huang, K., Gillis, N.: Computing large-scale matrix and tensor decomposition with structured factors: A unified nonconvex optimization perspective. EEE Signal Process. Mag. **37**(5), 78–94 (2020)

[15] Gillis, N.: Successive nonnegative projection algorithm for robust nonnegative blind source separation. SIAM Journal on Imaging Sciences **7**(2), 1420–1450 (2014)

[16] Gillis, N.: Nonnegative matrix factorization. SIAM, Philadelphia (2020)

[17] Gillis, N., Kumar, A.: Exact and heuristic algorithms for semi-nonnegative matrix factorization. SIAM Journal on Matrix Analysis and Applications **36**(4), 1404–1424 (2015)

[18] Gillis, N., Vavasis, S.A.: On the complexity of robust PCA and ℓ_1-norm low-rank matrix approximation. Mathematics of Operations Research **43**(4), 1072–1084 (2018)

[19] Heinz, D.C., Chein-I-Chang: Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. IEEE Trans. Geosci. Remote Sens. **39**(3), 529–545 (2001)

[20] Huang, K., Fu, X.: Detecting overlapping and correlated communities without pure nodes: Identifiability and algorithm. In: International Conference on Machine Learning, pp. 2859–2868 (2019)

[21] Huang, K., Sidiropoulos, N.D., Swami, A.: Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. IEEE Trans. Signal Process. **62**(1), 211–224 (2013)

[22] Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**, 788–791 (1999)

[23] Leplat, V., Ang, A.M., Gillis, N.: Minimum-volume rank-deficient nonnegative matrix factorizations. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3402–3406 (2019)

[24] Lin, C.H., Chi, C.Y., Wang, Y.H., Chan, T.H.: A fast hyperplane-based minimum-volume enclosing simplex algorithm for blind hyperspectral unmixing. IEEE Trans. Signal Process. **64**(8), 1946–1961 (2015)

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

[26] Lin, C.H., Wu, R., Ma, W.K., Chi, C.Y., Wang, Y.: Maximum volume inscribed ellipsoid: A new simplex-structured matrix factorization framework via facet enumeration and convex optimization. SIAM Journal on Imaging Sciences **11**(2), 1651–1679 (2018)

[27] Ma, W.K., Bioucas-Dias, J.M., Chan, T.H., Gillis, N., Gader, P., Plaza, A.J., Ambikapathi, A., Chi, C.Y.: A signal processing perspective on hyperspectral unmixing: Insights from remote sensing. EEE Signal Process. Mag. **31**(1), 67–81 (2013)

[28] Miao, L., Qi, H.: Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. IEEE Trans. Geosci. Remote Sens. **45**(3), 765–777 (2007)

[29] Razaviyayn, M., Hong, M., Luo, Z.Q.: A unified convergence analysis of block successive minimization methods for nonsmooth optimization. SIAM Journal on Optimization **23**(2), 1126–1153 (2013)

[30] Tatli, G., Erdogan, A.T.: Polytopic matrix factorization: Determinant maximization based criterion and identifiability. IEEE Trans. Signal Process. **69**, 5431–5447 (2021)

[31] Udell, M., Horn, C., Zadeh, R., Boyd, S., et al.: Generalized low rank models. Foundations and Trends® in Machine Learning **9**(1), 1–118 (2016)

[32] Zhu, F.: Hyperspectral unmixing: ground truth labeling, datasets, benchmark performances and survey. arXiv preprint arXiv:1708.05125 (2017)

[33] Ziegler, G.M.: Lectures on polytopes, vol. 152. Springer Science & Business Media (2012)

8 Appendix

8.1 Proofs of Lemma 1 and Lemma 2

Proof of Lemma 1. For any column θ_i of Θ, call $\mathcal{S}_i$ the corresponding face of $\mathcal{S}_i$, i.e. $\mathcal{S}_i := \{x \in \mathcal{S} : \theta_i^\top x = 1\}$, whose affine span is the affine subspace $\theta_i/\|\theta_i\|^2 + \theta_i^\perp$. When we translate by w, the column $\theta_i^{(w)}$ of Θ_w corresponding to the face $\mathcal{S}_i - w$ will satisfy $(\theta_i^{(w)})^\top(x-w) = 1$ for every $x \in \theta_i/\|\theta_i\|^2 + \theta_i^\perp$, and in particular

$$\theta_i^{(w)} || \, \theta_i, \quad (\theta_i^{(w)})^\top(\theta_i/\|\theta_i\|^2 - w) = 1 \implies \theta_i^{(w)} = \frac{1}{1-\theta_i^\top w}\theta_i \implies \Theta_w = \Theta \operatorname{Diag}(e - \Theta^\top w)^{-1}.$$

Notice that if $\Theta z = 0$ and $e^\top z = 1$, then $0 = \Theta_w \operatorname{Diag}(e - \Theta^\top w)z$ and $e^\top \operatorname{Diag}(e - \Theta^\top w)z = e^\top z - z^\top \Theta^\top w = 1$, so $z_w = \operatorname{Diag}(e - \Theta^\top w)z$ and $\Theta_w \operatorname{Diag}(z_w) = \Theta \operatorname{Diag}(z)$.

Let now Θ_v be the polar of $A_v := A - ve^\top$, where $A \in \mathbb{R}^{(r-1)\times r}$, $Ae = 0$ and $At = v \in \operatorname{conv}(A)^\circ$ with $t \in \Delta^r$. The face of $\operatorname{conv}(A_v)$ associated to $\theta_i^{(v)}$ is generated by all the columns of A_v except for the i-th one. As a consequence, $A_v^\top \Theta_v$ has all entries equal to one except for the diagonal, and in particular it is symmetric. As a consequence, $A_v^\top \Theta_v t = \Theta_v^\top A_v t = 0$ but since $A_v^\top$ is column full rank, we find that $\Theta_v t = 0$, so $z_v \equiv t$ and by the previous result $\Theta_v \operatorname{Diag}(z_v) = \Theta_v \operatorname{Diag}(t) = \Theta_{At} \operatorname{Diag}(t)$ does not depend on t.

□

Proof of Lemma 2. Using the definition of volume and the multiplicativity of the determinant,

$$\begin{aligned}\operatorname{vol}(\Theta N) &= \frac{1}{(n-1)!}\left|\det\begin{pmatrix}\Theta N \\ e^T\end{pmatrix}\right| = \frac{1}{(n-1)!}\left|\det\begin{pmatrix}\Theta \\ e^T N^{-1}\end{pmatrix}\right| |\det(N)| \\ &= \frac{1}{(n-1)!}\left|\det\begin{pmatrix}\Theta \\ e^T\end{pmatrix}\right| \left|\det\left(I + \frac{1}{e^T w} w e^T (N^{-1} - I)\right)\right| |\det(N)|.\end{aligned}$$

The eigenvalues of $I + \frac{1}{e^T w} w e^T (N^{-1} - I)$ are all equal to 1, except, possibly, for the eigenvalue $1 + \frac{1}{e^T w} e^T (N^{-1} - I) w = \frac{e^T N^{-1} w}{e^T w}$, thus completing the proof.

□

Here we show that given a bounded convex polytope $\mathcal{S} \subseteq \mathbb{R}^{r-1}$ with at least r vertices, the vertices of a maximum volume simplex contained in it coincide with r of the vertices of $\mathcal{S}$. This is useful in the proof of Theorem 4.

Lemma 3. *Given a bounded convex polytope $\mathcal{S} \subseteq \mathbb{R}^{r-1}$ with at least r vertices, the problem*

$$\max_{\Theta \in \mathbb{R}^{r-1 \times r}} \operatorname{vol}\big(\operatorname{conv}(\Theta)\big) \quad \textit{such that} \quad \operatorname{conv}(\Theta) \subseteq \operatorname{conv}(\mathcal{S})$$

is solved by a matrix Θ whose columns coincide with r vertices of $\mathcal{S}$.

Proof. By definition, $\operatorname{vol}(\operatorname{conv}(\Theta)) = \frac{1}{(r-1)!} \left| \det \begin{pmatrix} \Theta \\ e^T \end{pmatrix} \right|$. If we single out a column θ of Θ and fix all other entries, we can see that $\det \begin{pmatrix} \Theta \\ e^T \end{pmatrix} = c + q^\top \theta$ for a scalar $c \in \mathbb{R}$ and a vector $q \in \mathbb{R}^{r-1}$. The above maximization problem is equivalent to maximize $\operatorname{vol}(\operatorname{conv}(\Theta))^2$, that is proportional to $(c + q^\top \theta)^2 = \theta^\top q q^\top \theta + 2 c q^\top \theta + c^2$, i.e. a convex function in $\theta \in \mathcal{S}$. A global maximum of a convex function on a convex bounded polytope can always be found at one of its vertices. As a consequence, given any feasible Θ we can find a new feasible $\tilde{\Theta}$ with greater or equal volume and with vertices corresponding to a subset of the vertices of $\mathcal{S}$ by optimizing sequentially over the columns of Θ. Since $\mathcal{S}$ has a finite number of vertices, then one of the maximum volume simplices contained inside $\mathcal{S}$ must coincide with a simplex formed by r of its vertices. □

8.2 Convergence and sensitivity of MV-Dual

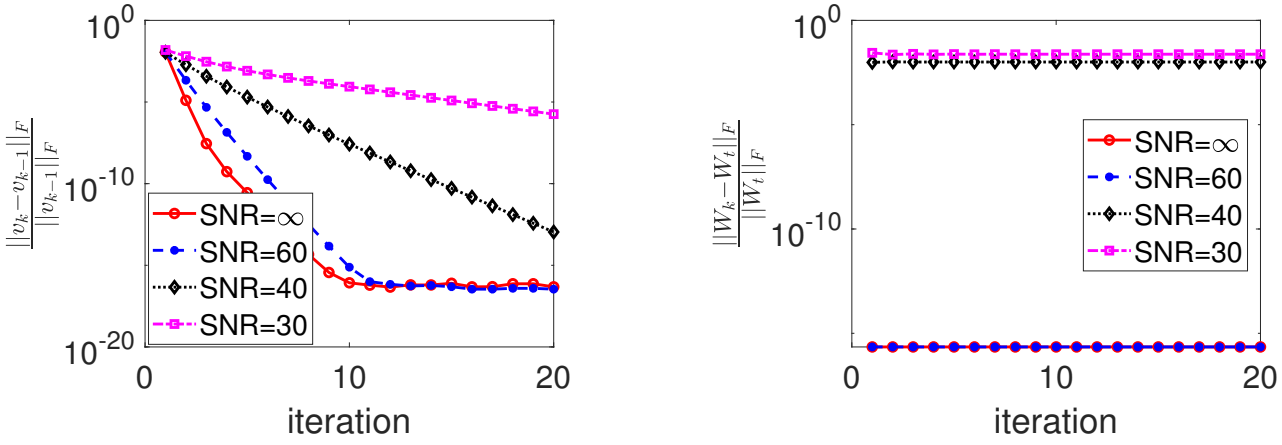
Convergence analysis Let us provide some insights on the convergence of MV-Dual. We choose $m = r = 3$, and $p = 0.76$ which is at the phase transition and hence leads to more difficult instances (see Figure 5 in the paper). We explore two scenarios to generate the samples:

1. Balanced case: samples are evenly distributed across the three facets, as in the paper. We let $n_1 = 90$ and $n_2 = 10$.

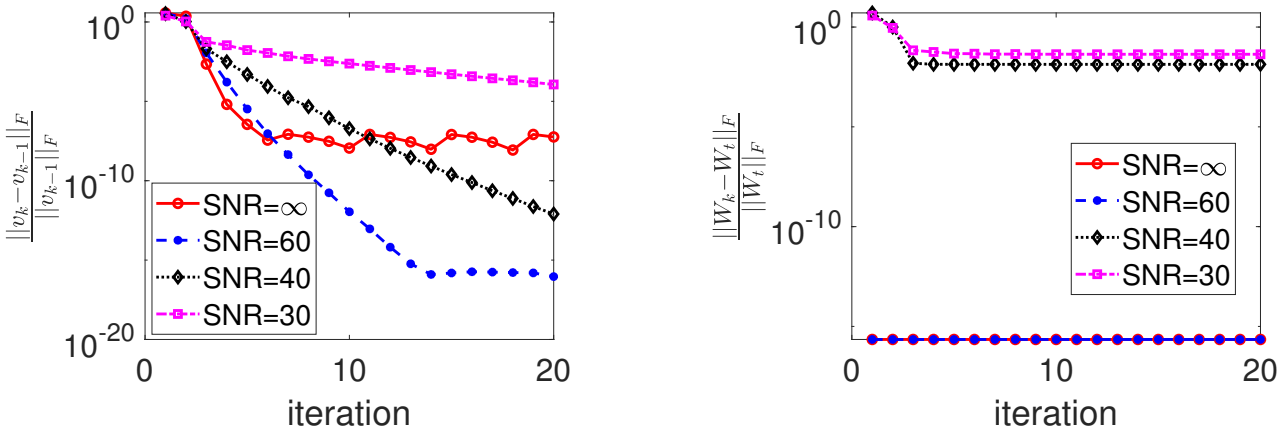
2. Imbalanced case: samples are unevenly distributed among facets. In this case, we set $n = 640$, with the number of samples on the three facets being $\{500, 100, 30\}$, and 10 samples chosen within the simplex.

At iteration k, let v_k indicate the obtained translation vector v and W_k be the estimated endmembers. Figure 8 shows the evolution of $\frac{\|v_k - v_{k-1}\|_2}{\|v_{k-1}\|_2}$ and $\frac{\|W_k - W_t\|_F}{\|W_t\|_F}$ where W_t is the ground truth, for different iterations. MV-Dual converges fast for all noise levels, as less than 5 iterations are needed for W_k to converge in MV-Dual. The explanation is that the solution v_k does not need to attain the minimum for W_k to correctly identify W_t; see Section 4.

Sensitivity to λ Figure 9 displays the average of ERR metric over 10 trials for various values of λ, for $m = r = 3$ and SNR=30. We observe that the performance of MV-Dual is stable w.r.t. the choice of $\lambda \in [0.4, 100]$, and the only noticeable sensitivity is in the 'transition' phase, with purity below 0.76, because there are two different simplices with small volumes containing the data points.



(a) Balanced data



(b) Imbalanced data

Figure 8: Convergence analysis of MV-dual ($m = r = 3$).

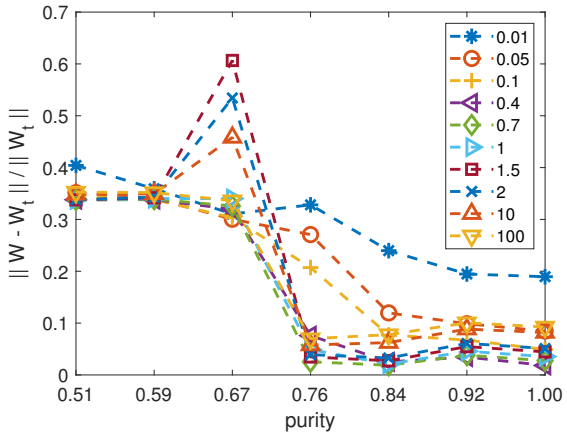


Figure 9: Performance of MV-Dual with respect to various values of λ.

MV-Dual vs GFPI We have seen that volume-based approaches, such as MV-Dual, outperform sample-counting-based ones, such as GFPI, in higher noise regimes, and that MV-Dual has a more stable performance. This phenomenon is illustrated on another example on Figure 10, where $m = r = 3$, SNR $= 30$ and there are 50 samples on each facet, with an additional 50 samples spread within the simplex. The worse performance of GFPI is noticeable as the noise has perturbed the orientation of the facets, leading to a worse estimation of W_t.

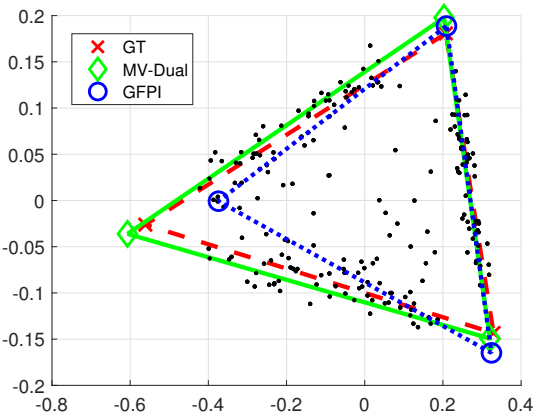


Figure 10: MV-Dual vs GFPI. GT stands for ground truth.