SIAM J. Imaging Sciences
Vol. 7, No. 2, pp. 1420–1450

Successive Nonnegative Projection Algorithm for Robust Nonnegative Blind Source Separation[*]

Nicolas Gillis[†]

Abstract. In this paper, we propose a new fast and robust recursive algorithm for near-separable nonnegative matrix factorization, a particular nonnegative blind source separation problem. This algorithm, which we refer to as the successive nonnegative projection algorithm (SNPA), is closely related to the popular successive projection algorithm (SPA) but takes advantage of the nonnegativity constraint in the decomposition. We prove that SNPA is more robust than SPA and can be applied to a broader class of nonnegative matrices. This is illustrated on some synthetic data sets and on a real-world hyperspectral image.

Key words. nonnegative matrix factorization, nonnegative blind source separation, separability, robustness to noise, hyperspectral unmixing, pure-pixel assumption

AMS subject classifications. 15A23, 65F30, 47A46

DOI. 10.1137/130946782

1. Introduction.

Nonnegative matrix factorization (NMF) has become a widely used tool for analysis of high-dimensional data. NMF decomposes approximately a nonnegative input data matrix $M \in \mathbb{R}_+^{m\times n}$ into the product of two nonnegative matrices $W \in \mathbb{R}_+^{m\times r}$ and $H \in \mathbb{R}_+^{r\times n}$ so that $M \approx WH$. Although NMF is NP-hard in general [33] and ill-posed (see [15] and the references therein), it has been used in many different areas, such as image processing [26], document classification [32], hyperspectral unmixing [30], community detection [34], and computational biology [10]. Recently, Arora et al. [4] introduced a subclass of nonnegative matrices, referred to as separable, for which NMF can be solved efficiently (that is, in polynomial time), even in the presence of noise. This subclass of NMF problems is referred to as near-separable NMF and has been shown to be useful in several applications, such as document classification [5, 3, 25, 11], blind source separation [9], video summarization and image classification [12], and hyperspectral unmixing (see section 1.1).

1.1. Near-separable NMF.

A matrix M is r-separable if there exist an index set $\mathcal{K}$ of cardinality r and a nonnegative matrix $H \in \mathbb{R}_+^{r\times n}$ with $M = M(:,\mathcal{K})H$. Equivalently, M is r-separable if

$$M = W\,[I_r,\, H']\,\Pi,$$

where I_r is the r-by-r identity matrix, H' is a nonnegative matrix, and Π is a permutation. Given a separable matrix, the goal is to identify the r columns of M allowing one to reconstruct it perfectly, that is, to identify the columns of M corresponding to the columns of W. In the

[*] Received by the editors November 26, 2013; accepted for publication (in revised form) April 15, 2014; published electronically June 24, 2014.

http://www.siam.org/journals/siims/7-2/94678.html

[†] Department of Mathematics and Operational Research, Faculté Polytechnique, Université de Mons, 7000 Mons, Belgium (nicolas.gillis@umons.ac.be).

presence of noise, the problem is referred to as near-separable NMF and can be stated as follows.

Near-separable NMF. Given the noisy r-separable matrix $\tilde{M} = WH + N \in \mathbb{R}^{m \times n}$, where N is the noise, $W \in \mathbb{R}^{m \times r}_{+}$, $H = [I_r, H']\Pi$ with $H' \geq 0$, and Π is a permutation, recover approximately the columns of W among the columns of $\tilde{M}$.

An important application of near-separable NMF is blind hyperspectral unmixing in the presence of pure pixels [21, 27]: A hyperspectral image is a set of images taken at different wavelengths. It can be associated with a nonnegative matrix $M \in \mathbb{R}^{m \times n}_{+}$, where m is the number of wavelengths and n the number of pixels. Each column of M is equal to the spectral signature of a given pixel; that is, $M(i,j)$ is the fraction of incident light reflected by the jth pixel at the ith wavelength. Under the linear mixing model, the spectral signature of a pixel is equal to a linear combination of the spectral signatures of the constitutive materials present in the image, referred to as endmembers. The weights in that linear combination are nonnegative and sum to one, and they correspond to the abundances of the endmembers in that pixel. If for each endmember there exists a pixel in the image containing only that endmember, then the pure-pixel assumption is satisfied. This assumption is equivalent to the separability assumption: Each column of W is the spectral signature of an endmember and is equal to a column of M corresponding to a pure pixel; see the survey [6] for more details.

Several provably robust algorithms have been proposed to solve the near-separable NMF problem using, e.g., geometric constructions [4, 3], linear programming [13, 7, 16, 19], or semidefinite programming [28, 20]. In the next section, we briefly describe the successive projection algorithm (SPA), which is closely related to the algorithm we propose in this paper.

1.2. SPA.

SPA is a simple but fast and robust recursive algorithm for solving near-separable NMF; see Algorithm SPA. At each step of the algorithm, the column of the input matrix $\tilde{M}$ with maximum ℓ_2 norm is selected, and then $\tilde{M}$ is updated by projecting each column onto the orthogonal complement of the columns selected so far. It was first introduced in [2], and later proved to be robust in [21].

Theorem 1.1 (see [21, Thm. 3]). *Let $\tilde{M} = WH + N$ be a near-separable matrix (see Assumption 1), where W has full column rank and $\max_i ||N(:,i)||_2 \leq \epsilon$. If $\epsilon \leq \mathcal{O}\big(\frac{\sigma_{\min}(W)}{\sqrt{r}\kappa^2(W)}\big)$, then SPA identifies the columns of W up to error $\mathcal{O}\left(\epsilon\, \kappa^2(W)\right)$; that is, the index set $\mathcal{K}$ identified by SPA satisfies*

$$\max_{1 \leq j \leq r} \min_{k \in \mathcal{K}} \left\| W(:,j) - \tilde{M}(:,k) \right\|_2 \leq \mathcal{O}\left(\epsilon\, \kappa^2(W)\right),$$

where $\kappa(W) = \frac{\sigma_{\max}(W)}{\sigma_{\min}(W)}$ is the condition number of W.

Moreover, SPA can be generalized by replacing the ℓ_2 norm (step 2 of Algorithm SPA) with any strongly convex function with a Lipschitz continuous gradient [21].

SPA is closely related to several hyperspectral unmixing algorithms, such as the automatic target generation process (ATGP) [31] and the successive volume maximization algorithm (SVMAX) [8]. It is also closely related to older techniques from other fields of research, in particular the modified Gram–Schmidt with column pivoting; see, e.g., [21, 27, 14, 17] and the references therein. Although SPA has many advantages (in particular, it is very fast and rather effective in practice), a drawback is that it requires the matrix W to have rank r. In

Algorithm SPA Successive Projection Algorithm [2, 21].

Input: Near-separable matrix $\tilde{M} = WH + N \in \mathbb{R}^{m \times n}$ satisfying Assumption 1, where W has full column rank, the number r of columns to be extracted.

Output: Set of indices $\mathcal{K}$ such that $M(:, \mathcal{K}) \approx W$ (up to permutation).

```
Let $R = \tilde{M}$, $\mathcal{K} = \{\}$, $k = 1$.
while $R \neq 0$ and $k \leq r$ do
    $p = \text{argmax}_j \, ||R_{:j}||_2$. †
    $R = \left( I - \frac{R_{:p} R_{:p}^T}{||R_{:p}||_2^2} \right) R$.
    $\mathcal{K} = \mathcal{K} \cup \{p\}$.
    $k = k + 1$.
end while
```

† In case of a tie, the index j whose corresponding column of the original matrix $\tilde{M}$ maximizes f is selected. In case of another tie, one of these columns is picked randomly.

fact, if M is r-separable with $r < \text{rank}(M)$, then SPA cannot extract enough columns even in noiseless conditions. Moreover, if the matrix W is ill-conditioned, SPA will most likely fail even for very small noise levels (see Theorem 1.1).

1.3. Contribution and outline of the paper.

The main contributions of this paper are the following:

- The introduction of a new fast and robust recursive algorithm for near-separable NMF, referred to as the successive nonnegative projection algorithm (SNPA), which overcomes the drawback of SPA that the matrix W has to have full column rank (section 2).
- The robustness analysis of SNPA (section 3). First, we show that Theorem 1.1 applies to SNPA as well; that is, we show that SNPA is robust to noise when W has full column rank. Second, given a matrix W, we define a new parameter $\beta(W) \geq \sigma_r(W)$ which is in general positive even if W does not have full column rank. We also define $\kappa_\beta(W) = \frac{\max_i ||W(:,i)||_2}{\beta(W)}$ and prove the following theorem.

  Theorem 3.22. Let $\tilde{M}$ be a near-separable matrix satisfying Assumption 1 with $\beta(W) > 0$. If $\epsilon \leq \mathcal{O}\big(\frac{\beta(W)}{\kappa_\beta^3(W)}\big)$, then SNPA with $f(.) = ||.||_2^2$ identifies the columns of W up to error $\mathcal{O}\big(\epsilon \, \kappa_\beta^3(W)\big)$.

  This proves that SNPA applies to a broader class of matrices (W does not need to have full column rank). It also proves that SNPA is more robust than SPA: in fact, even when W has rank r, if $\frac{\sigma_{\min}(W)}{\sqrt{r}\kappa^2(W)} \ll \frac{\beta(W)}{\kappa_\beta^3(W)}$, then SNPA will outperform SPA, as the noise level allowed by SNPA can be much larger.

We illustrate the effectiveness of SNPA on several synthetic data sets and a real-world hyperspectral image in section 4.

1.4. Notation.

The unit simplex is defined as $\Delta^m = \left\{ x \in \mathbb{R}^m \;\middle|\; x \geq 0, \sum_{i=1}^m x_i \leq 1 \right\}$, and the dimension m will be dropped when it is clear from the context. Given a matrix $W \in \mathbb{R}^{m \times r}$, $W(:, j)$, $W_{:j}$, or w_j denotes its jth column. The zero vector is denoted by 0; its dimension will be clear from the context. We also denote $||W||_{1,2} = \max_{x,||x||_1 \leq 1} ||Wx||_2 = \max_i ||W(:,i)||_2$. A matrix $W \in \mathbb{R}^{m \times r}$ is said to have full column rank if $\text{rank}(W) = r$.

2. SNPA.

In this paper, we propose a new family of fast and robust recursive algorithms to solve near-separable NMF problems; see Algorithm SNPA. At each step of the algorithm, the column of the input matrix $\tilde{M}$ maximizing the function f is selected, and then each column of $\tilde{M}$ is projected onto the convex hull of the columns extracted so far and the origin using the semimetric induced by f. (A natural choice for the function f in SNPA is $f(x) = ||x||_2^2$.) Hence the difference with SPA is the way the projection is performed. In this work, we perform the projections at step 5 of SNPA (which are convex optimization problems) using a fast gradient method, which is an optimal first-order method for minimizing convex functions with a Lipschitz continuous gradient [29]; see Appendix A for the implementation details. Although SNPA is computationally more expensive than SPA, it has the same asymptotic complexity, requiring a total of $\mathcal{O}(mnr)$ operations.

Algorithm SNPA Successive Nonnegative Projection Algorithm.

Input: Near-separable matrix $\tilde{M} = WH + N \in \mathbb{R}^{m \times n}$ satisfying Assumption 1 with $\beta(W) > 0$, the number r of columns to be extracted, and a strongly convex function f satisfying Assumption 2.

Output: Set of indices $\mathcal{K}$ such that $\tilde{M}(:,\mathcal{K}) \approx W$ up to permutation.

```
1: Let $R = \tilde{M}$, $\mathcal{K} = \{\}$, $k = 1$.
2: while $R \neq 0$ and $k \leq r$ do
3:     $p = \text{argmax}_j \, f(R_{:j})$. †
4:     $\mathcal{K} = \mathcal{K} \cup \{p\}$.
5:     $R(:,j) = \tilde{M}(:,j) - \tilde{M}(:,\mathcal{K})H^*(:,j)$ for all $j$, where
           $H^*(:,j) = \underset{x \in \Delta}{\text{argmin}} \, f\left(\tilde{M}(:,j) - \tilde{M}(:,\mathcal{K})x\right)$;    see Appendix A.
6:     $k = k + 1$.
7: end while
```

† In case of a tie, the index j whose corresponding column of the original matrix $\tilde{M}$ maximizes f is selected. In case of another tie, one of these columns is picked randomly.

SNPA is also closely related to the fast canonical hull algorithm, referred to as XRAY, from [25]. XRAY is a recursive algorithm for near-separable NMF and projects, at each step, the data points onto the convex cone of the columns extracted so far. The main differences between XRAY and SNPA are the following:

(i) XRAY uses another criterion to select a column of M at each step. This is a crucial difference between SNPA and XRAY. In fact, it was discussed in [25] that in some cases (e.g., when a data point belongs to the cone spanned by two columns of W and these two columns maximize the criterion simultaneously), XRAY may fail to identify a column of W even in noiseless conditions; see the remarks on page 5 of [25]. This will be illustrated in section 4. (Note that there actually exists several variants of XRAY with different but closely related criteria for the selection step; however, they all share this undesirable property.)

(ii) At each step, XRAY projects the data matrix onto the convex cone of the columns extracted so far, while SNPA projects onto their convex hull (with the origin). In this paper, we will assume that the entries of each column of the matrix H sum to at most one (equivalently that the columns of the data matrix belong to the convex hull of the columns of W and the origin); see Assumption 1 (and the ensuing discussion). Hence, performing the projection onto the convex hull allows one to take this prior information into account. However, a variant of SNPA with projections onto the convex cone of the columns extracted so far is also possible, although we have observed[1] that, under Assumption 1, it is less robust. It would be an interesting direction for further research to analyze this variant in detail.[2] (Note that a variant of XRAY with projections onto convex hull does not work because the criterion used by XRAY in the selection step relies on the projections being performed onto the convex cone.)

(iii) XRAY performs the projection step with respect to the ℓ_2 norm, while SNPA performs the projection with respect to the function f.

3. Robustness of SNPA.

In this section, we prove robustness of SNPA for any sufficiently small noise. The proofs are closely related to the robustness analysis of SPA developed in [21].

In section 3.1, we give the assumptions and definitions needed throughout the paper. In section 3.2, we prove that SNPA identifies the columns of W among the columns of M exactly in the noiseless case, which explains the intuition behind SNPA. In section 3.3, we derive our key lemmas, which allow us to show that the robustness analysis of SPA from Theorem 1.1 (which requires W to have full column rank) also applies to SNPA; see Theorems 3.17 and 3.18. In section 3.5, we generalize the analysis to a broader class of matrices for which W is not required to have full column rank; see Theorems 3.21 and 3.22.

In sections 3.6 and 3.7, we briefly discuss some possible improvements of SNPA and the choice of the function f, respectively.

3.1. Assumptions and definitions.

In this section, we describe the assumptions and definitions useful to prove robustness of SNPA.

Without loss of generality, we will assume throughout the paper that the input matrix has the following form.

Assumption 1 (near-separable matrix). The separable matrix $M \in \mathbb{R}^{m\times n}$ can be written as

$$M = WH = W[I_r, H'],$$

where $W \in \mathbb{R}^{m\times r}$, $H \in \mathbb{R}^{r\times n}_+$, and $H(:,j) \in \Delta$ for all j. The near-separable matrix $\tilde{M}$ is given by $\tilde{M} = M + N$, where N is the noise with $||N||_{1,2} \leq \epsilon$.

[1] We performed numerical experiments similar to that of section 4, and the variant of SNPA with projection onto the convex cones was less robust than SNPA while being slightly more robust than XRAY (because of the difference in the selection criterion).

[2] Under Assumption 1, the robustness analysis with projections onto the convex hull is made easier and allows us to derive better error bounds. The reason is that the columns of H^* (see step 5 of Algorithm SNPA) are normalized, while an additional constant would be needed in the analysis (if we would follow exactly the same steps) to bound the norm of these columns if the projection was onto the convex cone.

Any nonnegative near-separable matrix $M = WH = W[I_r, H']\Pi$ (with $W, H' \geq 0$ and Π a permutation) can be put in this form by proper permutation and normalization of its columns. In fact, permuting the columns of M and H so that the first r columns of M correspond to the columns of W, we have $M = W[I_r, H'] = WH$. The permutation does not affect our analysis because SNPA is not sensitive to permutation, while it makes the presentation nicer (the permutation matrix Π can be discarded). For the scaling, we divide each column of M by its ℓ_1 norm (unless it is an all-zero column, in which case we discard it) and divide the corresponding column of H by the same constant (hence we still have $M = WH$). Since $W = M(:, 1{:}r)$, the entries of each column of W sum to one. Since $M(:, j) = WH(:, j)$ for all j, the entries of each column of H must also sum to one: for all j,

$$1 = \sum_i M(i,j) = \sum_i \sum_k W(i,k)H(k,j) = \sum_k H(k,j) \sum_i W(i,k) = \sum_k H(k,j);$$

see also the discussion in [21]. Column normalization makes the presentation nicer: In fact, otherwise the noise that can be tolerated on each column of $\tilde{M} = M + N$ will have to be proportional to the norm of the corresponding column of the matrix H (for example, an all-zero column cannot tolerate any noise because it can be made an extreme ray of the cone spanned by the columns of M for any positive noise level).

Note that Assumption 1 *does not require W to be nonnegative*; hence our result will apply to a broader class than the nonnegative near-separable matrices. It is also interesting to note that data matrices corresponding to hyperspectral images are naturally scaled since the columns of H correspond to abundances and their entries sum to one (see section 1.1 for more details and section 4.2 for some numerical experiments).

We will also assume that, in SNPA, the following hold.

Assumption 2. The function $f : \mathbb{R}^m \to \mathbb{R}_+$ is strongly convex with parameter $\mu > 0$, its gradient is Lipschitz continuous with constant L, and its global minimizer is the all-zero vector with $f(0) = 0$.

A function f is strongly convex with parameter μ if and only if it is convex and for any $x, y \in \text{dom}(f)$ and for all $\delta \in [0, 1]$,

$$f(\delta x + (1-\delta)y) \;\leq\; \delta f(x) + (1-\delta)f(y) - \frac{\mu}{2}\delta(1-\delta)||x-y||_2^2. \tag{3.1}$$

Moreover, its gradient is Lipschitz continuous with constant L if and only if for any $x, y \in \text{dom}(f)$, we have $||\nabla f(x) - \nabla f(y)||_2 \leq L||x - y||_2$. Convex analysis also tells us that if f satisfies Assumption 2, then, for any x, y,

$$f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}||x-y||_2^2 \quad \leq \quad f(y) \quad \leq \quad f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}||x-y||_2^2.$$

In particular, taking $x = 0$, we have, for any $y \in \mathbb{R}^m$,

$$\frac{\mu}{2}||y||_2^2 \quad \leq \quad f(y) \quad \leq \quad \frac{L}{2}||y||_2^2 \tag{3.2}$$

since $f(0) = 0$ and $\nabla f(0) = 0$ (because zero is the global minimizer of f); see, e.g., [23]. Note that this implies $f(x) > 0$ for any $x \neq 0$; hence f induces a semimetric, the distance between two points x and y being defined by $f(x - y)$.

We will use the following notation for the residual computed at step 5 of Algorithm SNPA.

Definition (projection and residual). Given $B \in \mathbb{R}^{m \times s}$ and a function f satisfying Assumption 2, we define the projection $\mathcal{P}_B^f(x)$ of x onto the convex hull of the columns of B and the origin with respect to the semimetric induced by $f(.)$ as follows:

$$\mathcal{P}_B^f(x) : \mathbb{R}^m \to \mathbb{R}^m : x \to \mathcal{P}_B^f(x) = By^*, \quad \text{where } y^* = \operatorname*{argmin}_{y \in \Delta} f(x - By).$$

We also define the residual $\mathcal{R}_B^f$ of the projection $\mathcal{P}_B^f$ as follows:

$$\mathcal{R}_B^f : \mathbb{R}^m \to \mathbb{R}^m : x \to \mathcal{R}_B^f(x) = x - \mathcal{P}_B^f(x).$$

For a matrix $A \in \mathbb{R}^{m \times r}$, we will denote by $\mathcal{P}_B^f(A)$ the matrix whose columns are the projections of the columns of A, that is, $\left(\mathcal{P}_B^f(A)\right)_{:i} = \mathcal{P}_B^f(A_{:i})$ for all i, and $\mathcal{R}_B^f(A) = A - \mathcal{P}_B^f(A)$.

Given a matrix $W \in \mathbb{R}^{m \times r}$, we introduce the following notation:

$$\begin{aligned}
\alpha(W) &= \min_{1 \le j \le r, x \in \Delta} \|W(:,j) - W(:,\mathcal{J})x\|_2, \text{ where } \mathcal{J} = \{1, 2, \dots, r\} \backslash \{j\}, \\
\nu(W) &= \min_i ||w_i||_2, \\
\gamma(W) &= \min_{i \neq j} ||w_i - w_j||_2, \\
\omega(W) &= \min \left\{ \nu(W), \frac{1}{\sqrt{2}} \gamma(W) \right\}, \\
K(W) &= ||W||_{1,2} = \max_i ||w_i||_2, \\
\sigma(W) &= \begin{cases} \sigma_r(W) = \sigma_{\min}(W) & \text{if } m \ge r, \\ 0 & \text{if } m < r. \end{cases}
\end{aligned}$$

The parameter $\alpha(W)$ is the minimum distance between a column of W and the convex hull of the other columns of W and the origin. It is interesting to notice that, under Assumption 1, $\alpha(W) > 0$ is a necessary condition to being able to identify the columns of W among the columns of M (in fact, $\alpha(W) = 0$ means that a column of W belongs to the convex hull of the other columns of W, and the origin hence cannot be distinguished from the other data points). It is also a sufficient condition, as some algorithms are guaranteed to identify the columns of W when $\alpha(W) > 0$ even in the presence of noise [4, 16, 19].

3.2. Recovery in the noiseless case. In this section, we show that, in the noiseless case, SNPA is able to perfectly identify the columns of W among the columns of M. Although this result is implied by our analysis in the noisy case (see section 3.4), it gives the intuition behind the working of SNPA.

Lemma 3.1. *Let $B \in \mathbb{R}^{m \times s}$, $A \in \mathbb{R}^{m \times k}$, and $z \in \Delta^k$, and let f satisfy Assumption 2. Then*

$$f\left(\mathcal{R}_B^f(Az)\right) \le f\left(\mathcal{R}_B^f(A)z\right).$$

Proof. Let us denote $Y(:,j) = \operatorname{argmin}_{y \in \Delta} f(A(:,j) - By)$ for all j, that is, $\mathcal{R}_B^f(A) = A - BY$. We have

$$f\left(\mathcal{R}_B^f(Az)\right) = \min_{y \in \Delta} f\left(Az - By\right) \le f\left(Az - BYz\right) = f\left(\mathcal{R}_B^f(A)z\right).$$

The inequality follows from $Yz \in \Delta$ since $Y(:,j) \in \Delta$ for all j and $z \in \Delta$. ■

Theorem 3.2. *Let $M = W\,[I_r,\, H'] = WH$ be a separable matrix satisfying Assumption 1, where W has full column rank, and let f satisfy Assumption 2. Then* SNPA *applied on matrix M identifies a set of indices $\mathcal{K}$ such that, up to permutation, $M(:,\mathcal{K}) = W$.*

Proof. We prove the result by induction.

First step. Since the columns of M belong to the convex hull of the columns of W and the origin (the entries of each column of H are nonnegative and sum to at most one), and since a strongly convex function is always maximized at a vertex of a polytope, a column of W will be identified at the first step of SNPA (the origin cannot be extracted since, by assumption, it minimizes f). More formally, letting $h \in \Delta^r$, we have

$$\begin{aligned} f\left(Wh\right) &= f\left(\sum_{k=1}^{r} W(:,k)h(k) + \left(1 - \sum_{k=1}^{r} h(k)\right) 0\right) \\ &\leq \sum_{k=1}^{r} h(k) f\left(W(:,k)\right) \\ &\leq \max_k f\left(W(:,k)\right). \end{aligned}$$

The first inequality follows from convexity of f and the fact that $f(0) = 0$. By strong convexity (see (3.1)), the first inequality is always strict unless $h = e_j$ for some j (where e_j is the jth column of the identity matrix). The second inequality follows from $h \in \Delta$ and the fact that $f(x) > 0$ for any $x \neq 0$. Since all columns of M can be written as Wh for some $h \in \Delta^r$, this implies that, at the first step, SNPA extracts the index corresponding to the column of W maximizing f.

Induction step. Assume SNPA has extracted some indices $\mathcal{K}$ corresponding to columns of W, that is, $M(:,\mathcal{K}) = W(:,\mathcal{I})$ for some $\mathcal{I}$. We have for any $h \in \Delta^r$ that

$$\begin{aligned} f\left(\mathcal{R}^f_{W(:,\mathcal{I})}(Wh)\right) &\underset{\text{(Lemma 3.1)}}{\leq} f\left(\mathcal{R}^f_{W(:,\mathcal{I})}(W)h\right) \\ &\underset{\text{(Ass. 2)}}{\leq} \sum_{k=1}^{r} h(k) f\left(\mathcal{R}^f_{W(:,\mathcal{I})}(W(:,k))\right) \\ &\underset{(h\in\Delta^r,\, f(x)>0\,\forall x\neq 0)}{\leq} \max_k f\left(\mathcal{R}^f_{W(:,\mathcal{I})}(W(:,k))\right). \end{aligned}$$

Finally, noting that the residual R in SNPA is equal to $\mathcal{R}_{W(:,\mathcal{I})}(M)$ and since

- $\mathcal{R}^f_{W(:,\mathcal{I})}(W(:,k)) = 0$ for all $k \in \mathcal{I}$,
- $\mathcal{R}^f_{W(:,\mathcal{I})}(W(:,k)) \neq 0$ for all $k \notin \mathcal{I}$ because W has full column rank, and
- the second inequality is strict unless $h \neq e_j$ for some j by strong convexity of f,

SNPA identifies a column of W not extracted yet. ■

Note that the proof does not need W to have full column rank but requires only that $\mathcal{R}^f_{W(:,\mathcal{I})}(W(:,k)) \neq 0$ for all $k \notin \mathcal{I}$ for any subset $\mathcal{I}$ of $\{1, 2, \ldots, r\}$. This observation will be exploited in section 3.5 to show the robustness of SNPA when W does not have full column rank.

3.3. Key lemmas. In the following, we derive the key lemmas to prove the robustness of SNPA.

More precisely, we subdivide the columns of W into two subsets as follows: $W = [A, B]$. The columns of the matrix B represent the columns of W which have already been approximately identified by SNPA, while the columns of A are the columns of W yet to be identified. The columns of the matrix $\tilde{B}$ correspond to the columns of matrix $\tilde{M}$ already extracted by SNPA, and we will assume that $||B - \tilde{B}||_{1,2} \leq \bar{\epsilon}$ for some constant $\bar{\epsilon}$. Lemmas 3.3 to 3.11 lead to a lower bound for $\omega\big(\mathcal{R}^f_{\tilde{B}}(A)\big)$ using $\sigma([A, B])$; see Corollary 3.12. Combined with Lemmas 3.13 to 3.15, these lemmas will imply that if W has full column rank, then a column of W not extracted yet (that is, a column of A) is identified approximately at the next step of SNPA; see Theorem 3.16. Finally, using that result inductively leads to the robustness of SNPA; see Theorem 3.17.

Lemma 3.3. *For any $B \in \mathbb{R}^{m\times s}$, $x \in \mathbb{R}^m$, and f satisfying Assumption 2, we have*

$$\left\|\mathcal{R}^f_B(x)\right\|_2 \leq \sqrt{\frac{L}{\mu}}\, \|x\|_2 .$$

Proof. Using (3.2), we have

$$\left\|\mathcal{R}^f_B(x)\right\|_2^2 \leq \frac{2}{\mu} f\left(\mathcal{R}^f_B(x)\right) = \frac{2}{\mu}\min_{y\in\Delta} f(x - By) \leq \frac{L}{\mu}\min_{y\in\Delta} ||x - By||_2^2 \leq \frac{L}{\mu}||x||_2^2$$

since $0 \in \Delta$. ∎

Lemma 3.4. *Let $B \in \mathbb{R}^{m\times s}$ and $B = \tilde{B} + N$ with $||N||_{1,2} \leq \bar{\epsilon}$, and let f satisfy Assumption 2. Then,*

$$\max_j \left\|\mathcal{R}^f_{\tilde{B}}(b_j)\right\|_2 \leq \sqrt{\frac{L}{\mu}}\bar{\epsilon}.$$

Proof. Using (3.2), we have, for all j,

$$\begin{aligned}\left\|\mathcal{R}^f_{\tilde{B}}(b_j)\right\|_2^2 &\leq \frac{2}{\mu} f\left(\mathcal{R}^f_{\tilde{B}}(b_j)\right) = \frac{2}{\mu}\min_{x\in\Delta} f(b_j - \tilde{B}x)\\ &\leq \frac{2}{\mu} f(b_j - \tilde{b}_j) = \frac{2}{\mu} f(n_j)\\ &\leq \frac{L}{\mu}||n_j||_2^2 \leq \frac{L}{\mu}\bar{\epsilon}^2. \qquad \blacksquare\end{aligned}$$

Lemma 3.5. *Let $A \in \mathbb{R}^{m\times k}$, $B \in \mathbb{R}^{m\times s}$, and f satisfy Assumption 2. Then,*

$$\nu\left(\mathcal{R}^f_B(A)\right) \geq \alpha([A, B]).$$

Proof. This follows directly from the definitions of α and $\mathcal{R}^f_B$: in fact,

$$\nu\left(\mathcal{R}^f_B(A)\right) \geq \min_j \min_{y\in\Delta} ||A(:, j) - By||_2 \geq \alpha([A, B]). \qquad \blacksquare$$

Lemma 3.6. *Let Z and $\tilde{Z} \in \mathbb{R}^{m\times r}$ satisfy $||Z - \tilde{Z}||_{1,2} \leq \bar{\epsilon}$. Then,*

$$\alpha(\tilde{Z}) \geq \alpha(Z) - 2\bar{\epsilon}.$$

Proof. Denoting $N = Z - \tilde{Z}$ and $\mathcal{J} = \{1, 2, \dots, k\}\backslash\{j\}$, we have

$$\begin{aligned}
\alpha(\tilde{Z}) &= \min_{1\leq j\leq k, x\in\Delta} ||\tilde{z}_j - \tilde{Z}(:, \mathcal{J})x||_2 \\
&= \min_{1\leq j\leq k, x\in\Delta} ||z_j - n_j - Z(:, \mathcal{J})x + N(:, \mathcal{J})x||_2 \\
&\geq \min_{1\leq j\leq k, x\in\Delta} ||z_j - Z(:, \mathcal{J})x||_2 - ||n_j||_2 - ||N(:, \mathcal{J})x||_2 \\
&\geq \min_{1\leq j\leq k, x\in\Delta} ||z_j - Z(:, \mathcal{J})x||_2 - 2\bar{\epsilon} \\
&= \alpha(Z) - 2\bar{\epsilon}
\end{aligned}$$

since $\max_{x\in\Delta} ||N(:, \mathcal{J})x||_2 \leq \max_{||x||_1\leq 1} ||N(:, \mathcal{J})x||_2 = ||N(:, \mathcal{J})||_{1,2} \leq \bar{\epsilon}$. ■

Corollary 3.7. *Let $A \in \mathbb{R}^{m\times k}$, B, and $\tilde{B} \in \mathbb{R}^{m\times s}$ satisfy $||B - \tilde{B}||_{1,2} \leq \bar{\epsilon}$, and let f satisfy Assumption 2. Then,*

$$\nu\left(\mathcal{R}^f_{\tilde{B}}(A)\right) \geq \Big(\alpha\left([A, B]\right) - \min(s, 2)\bar{\epsilon}\Big).$$

Proof. If $s = 0$, the result follows from Lemma 3.5 (B is an empty matrix). If $s = 1$, then it is easily derived using the same steps as in the proof of Lemma 3.6 (B only has one column). Otherwise, Lemmas 3.5 and 3.6 imply that

$$\nu\left(\mathcal{R}^f_{\tilde{B}}(A)\right) \geq \alpha\left([A, \tilde{B}]\right) \geq \alpha\left([A, B]\right) - 2\bar{\epsilon}. \qquad ■$$

Lemma 3.8. *For any $W \in \mathbb{R}^{m\times r}$, $\alpha(W) \geq \sigma(W)$.*

Proof. We have

$$\begin{aligned}
\alpha(W) &= \min_{1\leq j\leq r} \min_{x\in\Delta} ||W(:, j) - W(:, \mathcal{J})x||_2 \\
&\geq \min_{1\leq j\leq r} \min_{z\in\mathbb{R}^r, z(j)=1} ||Wz||_2 \\
&\geq \min_{z\in\mathbb{R}^r, ||z||_2\geq 1} ||Wz||_2 = \sigma(W). \qquad ■
\end{aligned}$$

Lemma 3.9. *Let $x, y \in \mathbb{R}^m$, $B \in \mathbb{R}^{m\times s}$, and f satisfy Assumption 2. Then,*

$$\left\|\mathcal{R}^f_B(x)\right\|_2 \geq \sigma([B, x]) \quad \text{and} \quad \left\|\mathcal{R}^f_B(x) - \mathcal{R}^f_B(y)\right\|_2 \geq \sqrt{2}\,\sigma([B, x, y]).$$

Proof. Let us denote $z_x = \text{argmin}_{z\in\Delta} f(x - Bz)$ and $z_y = \text{argmin}_{z\in\Delta} f(y - Bz)$; we have

$$\begin{aligned}||\mathcal{R}_B^f(x)||_2 = ||x - Bz_x||_2 \geq \min_{z\in\mathbb{R}^p} ||x + Bz||_2 &= \min_{z\in\mathbb{R}^{p+1}, z(1)=1} ||[x, B]z||_2 \\ &\geq \min_{||z||_2\geq 1} ||[x, B]z||_2 \geq \sigma([x, B])\end{aligned}$$

and

$$\begin{aligned}||\mathcal{R}_B^f(x) - \mathcal{R}_B^f(y)||_2 &= ||(x - Bz_x) - (y - Bz_y)||_2 \\ &\geq \min_{z\in\mathbb{R}^p} ||x - y + Bz||_2 \\ &= \min_{z\in\mathbb{R}^{p+2}, z(1)=1, z(2)=-1} ||[x, y, B]z||_2 \\ &\geq \min_{||z||_2\geq\sqrt{2}} ||[x, y, B]z||_2 = \sqrt{2}\,\sigma([x, y, B]).\end{aligned}$$ ■

Lemma 3.10 (singular value perturbation [22, Cor. 8.6.2]). *Let $\tilde{B} = B + N \in \mathbb{R}^{m\times s}$ with $s \leq m$. Then, for all $1 \leq i \leq s$,*

$$\left|\sigma_i(B) - \sigma_i(\tilde{B})\right| \leq \sigma_{\max}(N) = ||N||_2 \leq \sqrt{s}\,||N||_{1,2}.$$

Lemma 3.11. *Let $x, y \in \mathbb{R}^m$, B, and $\tilde{B} \in \mathbb{R}^{m\times s}$ be such that $||\tilde{B} - B||_{1,2} \leq \bar{\epsilon}$, and let f satisfy Assumption 2. Then,*

$$\left\|\mathcal{R}_{\tilde{B}}^f(x) - \mathcal{R}_{\tilde{B}}^f(y)\right\|_2 \geq \sqrt{2}\,\left(\sigma([B, x, y]) - \sqrt{s}\,\bar{\epsilon}\right).$$

Proof. This follows from Lemmas 3.9 and 3.10. ■

Corollary 3.12. *Let $A \in \mathbb{R}^{m\times k}$, B, and $\tilde{B} \in \mathbb{R}^{m\times s}$ satisfy $||\tilde{B} - B||_{1,2} \leq \bar{\epsilon}$, and let f satisfy Assumption 2. Then,*

$$\omega\left(\mathcal{R}_{\tilde{B}}^f(A)\right) \geq \left(\sigma\left([A, B]\right) - \sqrt{2s}\,\bar{\epsilon}\right).$$

Proof. Using Lemma 3.11, we have

$$\frac{1}{\sqrt{2}}\gamma\left(\mathcal{R}_{\tilde{B}}^f(A)\right) = \frac{1}{\sqrt{2}}\min_{i\neq j} ||\mathcal{R}_{\tilde{B}}^f(a_i) - \mathcal{R}_{\tilde{B}}^f(a_j)||_2 \geq \sigma([B, a_i, a_j]) - \sqrt{s}\,\bar{\epsilon} \geq \sigma([A, B]) - \sqrt{s}\,\bar{\epsilon}.$$

Using Corollary 3.7 and Lemma 3.8, we have

$$\nu\left(\mathcal{R}_{\tilde{B}}^f(A)\right) = \min_i \left\|\mathcal{R}_{\tilde{B}}^f(a_i)\right\|_2 \geq \alpha\left([A, B]\right) - \min(s, 2)\bar{\epsilon} \geq \sigma\left([A, B]\right) - \min(s, 2)\bar{\epsilon}.$$

Since $\sqrt{2s} \geq \min(s, 2)$ for any $s \geq 0$, the proof is complete. ■

Lemma 3.13. *Let $B \in \mathbb{R}^{m\times s}$, $A \in \mathbb{R}^{m\times k}$, $n \in \mathbb{R}^m$, and $z \in \Delta^k$, and let f satisfy Assumption 2. Then,*

$$f\left(\mathcal{R}_B^f(Az + n)\right) \leq f\left(\mathcal{R}_B^f(Az) + n\right) \quad \text{and} \quad f\left(\mathcal{R}_B^f(Az + n)\right) \leq f\left(\mathcal{R}_B^f(A)z + n\right).$$

Proof. Let us denote $y^* = \text{argmin}_{y\in\Delta} f(Az - By)$ and $Y(:,j) = \text{argmin}_{y\in\Delta} f(A(:,j) - By)$ for all j, that is, $\mathcal{R}_B^f(A) = A - BY$. We have

$$f\left(\mathcal{R}_B^f(Az+n)\right) = \min_{y\in\Delta} f\left(Az + n - By\right) \leq f\left(Az - By^* + n\right) = f\left(\mathcal{R}_B^f(Az) + n\right)$$

and

$$f\left(\mathcal{R}_B^f(Az+n)\right) = \min_{y\in\Delta} f\left(Az + n - By\right) \leq f\left(Az - BYz + n\right) = f\left(\mathcal{R}_B^f(A)z + n\right),$$

where the inequality follows from $y = Yz \in \Delta$ since $Y(:,j) \in \Delta$ for all j and $z \in \Delta$. ■

Let us also recall two useful lemmas from [21].

Lemma 3.14 (see [21, Lem. 3]). *Let the function f satisfy Assumption 2. Then, for any $||x||_2 \leq K$ and $||n||_2 \leq \epsilon \leq K$, we have*

$$f(x) - \epsilon KL \leq f(x+n) \leq f(x) + \frac{3}{2}\epsilon KL.$$

Lemma 3.15 (see [21, Lem. 2]). *Let $Z = [P, Q]$, where $P \in \mathbb{R}^{m\times k}$ and $Q \in \mathbb{R}^{m\times s}$, and let f satisfy Assumption 2. If $\nu(P) > 2\sqrt{\frac{L}{\mu}}K(Q)$, then, for any $0 \leq \delta \leq \frac{1}{2}$,*

$$f^* = \max_{x\in\Delta} f(Zx) \quad \text{such that } x_i \leq 1 - \delta \quad \text{for } 1 \leq i \leq k$$

satisfies

$$f^* \leq \max_i f(p_i) - \frac{1}{2}\mu\left(1-\delta\right)\delta\,\omega(P)^2.$$

Moreover, the maximum is attained only at point x such that $x_i = 1 - \delta$ for some $1 \leq i \leq k$.

3.4. Robustness of SNPA when W has full column rank.

Theorem 3.16 shows that if SNPA has already extracted some columns of W up to error $\bar{\epsilon}$, then the next extracted column of $\tilde{M}$ will be close to a column of W not yet extracted. This will allow us to prove inductively that SNPA is robust to noise; see Theorem 3.17.

Theorem 3.16. *Let the following hold:*

- *f satisfies Assumption 2, with strong convexity parameter μ, and its gradient has Lipschitz constant L.*
- *$\tilde{M}$ satisfies Assumption 1 with $\tilde{M} = M + N = WH + N$, $W = [A, B]$, $A \in \mathbb{R}^{m\times k}$, $B \in \mathbb{R}^{m\times s}$, $||N||_{1,2} \leq \epsilon$, and $H = [I_r, H'] \in \mathbb{R}_+^{r\times n}$, where $H(:,j) \in \Delta$ for all j.*
- *$\tilde{B} \in \mathbb{R}^{m\times s}$ satisfies*

$$||B - \tilde{B}||_{1,2} \leq \bar{\epsilon} = C\epsilon \quad \text{for some } C \geq 0.$$

- *$W = [A, B]$ is such that $\sigma(W) = \sigma > 0$. We denote $\alpha = \alpha(W)$ and $K = K(W)$.*
- *ϵ is sufficiently small so that*

$$\epsilon < \min\left(\frac{\sigma^2\mu^{3/2}}{144KL^{3/2}}, \frac{\alpha\mu}{4LC}, \frac{\sigma}{2C\sqrt{2s}}\right).$$

Then the index i corresponding to a column $\tilde{m}_i$ of $\tilde{M}$ that maximizes the function $f\big(\mathcal{R}^f_{\tilde{B}}(.)\big)$ satisfies

$$m_i = Wh_i = [A, B]h_i, \quad \textit{where } h_i(\ell) \geq 1 - \delta \textit{ with } 1 \leq \ell \leq k, \tag{3.3}$$

and $\delta = \frac{72\epsilon K L^{3/2}}{\sigma^2 \mu^{3/2}}$, *which implies*

$$||\tilde{m}_i - w_\ell||_2 = ||\tilde{m}_i - a_\ell||_2 \leq \epsilon + 2K\delta = \epsilon \left(1 + 144 \frac{K^2}{\sigma^2} \frac{L^{3/2}}{\mu^{3/2}} \right). \tag{3.4}$$

Proof. First note that $\epsilon \leq \frac{\sigma^2 \mu^{3/2}}{144 K L^{3/2}}$ implies $\delta \leq \frac{1}{2}$. Then, let us show that

$$\nu\left(\mathcal{R}^f_{\tilde{B}}(A)\right) > 2\sqrt{\frac{L}{\mu}} K\left(\mathcal{R}^f_{\tilde{B}}(B)\right)$$

so that Lemma 3.15 will apply to $P = \mathcal{R}^f_{\tilde{B}}(A)$ and $Q = \mathcal{R}^f_{\tilde{B}}(B)$. Since $||B - \tilde{B}||_{1,2} \leq \bar{\epsilon}$, by Lemma 3.4, we have $K\big(\mathcal{R}^f_{\tilde{B}}(B)\big) \leq \sqrt{\frac{L}{\mu}}\bar{\epsilon}$. Therefore,

$$\begin{aligned}
\nu\left(\mathcal{R}^f_{\tilde{B}}(A)\right) &\underset{(\text{Corollary } 3.7)}{\geq} \alpha - 2\bar{\epsilon} \\
&\underset{\left(\bar{\epsilon} = C\epsilon < \frac{\alpha\mu}{4L}, L \geq \mu\right)}{>} 2\frac{L}{\mu}\bar{\epsilon} \\
&\underset{(\text{Lemma } 3.4)}{\geq} 2\sqrt{\frac{L}{\mu}} K\left(\mathcal{R}^f_{\tilde{B}}(B)\right).
\end{aligned}$$

Let us also show that $\omega\big(\mathcal{R}^f_{\tilde{B}}(A)\big) \geq \frac{\sigma}{2}$. By Corollary 3.12 and the assumption that $\bar{\epsilon} = C\epsilon \leq \frac{\sigma}{2\sqrt{2s}}$, we have

$$\omega\left(\mathcal{R}^f_{\tilde{B}}(A)\right) \geq \sigma - \sqrt{2s}\bar{\epsilon} \geq \frac{\sigma}{2}. \tag{3.5}$$

We can now prove (3.3) by contradiction. Assume the extracted index, say the ith, which maximizes $f\big(\mathcal{R}^f_{\tilde{B}}(.)\big)$ among the columns of $\tilde{M}$, satisfies $\tilde{m}_i = m_i + n_i = Wh_i + n_i$ with

$h_i(\ell) < 1-\delta$ for $1 \leq \ell \leq k$. We have

$$
\begin{aligned}
f\left(\mathcal{R}^f_{\tilde{B}}(\tilde{m}_i)\right) &\underset{\text{(Lemma 3.13)}}{\leq} f\left(\mathcal{R}^f_{\tilde{B}}(W)h_i + n_i\right) \\
&\underset{\text{(Lemma 3.14)}}{\leq} f\left(\mathcal{R}^f_{\tilde{B}}(W)h_i\right) + \frac{3}{2}\epsilon K\left(\mathcal{R}^f_{\tilde{B}}(A)\right)L \\
&\underset{\text{(Ass. 2)}}{<} \max_{x\in\Delta^r, x(\ell)\leq 1-\delta\, 1\leq\ell\leq k} f\left(\mathcal{R}^f_{\tilde{B}}(W)x\right) + \frac{3}{2}\epsilon K\sqrt{\frac{L}{\mu}}L \\
&\underset{\text{(Lemma 3.15)}}{\leq} \max_j f\left(\mathcal{R}^f_{\tilde{B}}(a_j)\right) - \frac{1}{2}\mu\delta(1-\delta)\omega\left(\mathcal{R}^f_{\tilde{B}}(A)\right)^2 + \frac{3}{2}\epsilon K\frac{L^{3/2}}{\mu^{1/2}} \\
&\underset{\text{(Lem. 3.13, (3.5))}}{\leq} \max_j f\left(\mathcal{R}^f_{\tilde{B}}(\tilde{a}_j) - n_j\right) - \frac{1}{8}\mu\delta(1-\delta)\sigma^2 + \frac{3}{2}\epsilon K\frac{L^{3/2}}{\mu^{1/2}}, \\
&\underset{\text{(Lemma 3.14)}}{\leq} \max_j f\left(\mathcal{R}^f_{\tilde{B}}(\tilde{a}_j)\right) - \frac{1}{8}\mu\delta(1-\delta)\sigma^2 + \frac{9}{2}\epsilon K\frac{L^{3/2}}{\mu^{1/2}}, \qquad (3.6)
\end{aligned}
$$

where $\tilde{a}_j$ is the perturbed column of M corresponding to w_j (that is, $\tilde{a}_j = w_j + n_j$). The second inequality follows from Lemma 3.14 since, by the convexity of $||.||_2$ and by Lemma 3.3, we have

$$\left\|\mathcal{R}^f_{\tilde{B}}(W)h_i\right\|_2 \leq \max_i \left\|\mathcal{R}^f_{\tilde{B}}(w_i)\right\|_2 \leq \sqrt{\frac{L}{\mu}}K.$$

The third inequality is strict since, by the strong convexity of f, the maximum is attained at a vertex with $x(\ell) = 1-\delta$ for some $1 \leq \ell \leq k$ at optimality, while we assumed $h_i(\ell) < 1-\delta$ for $1 \leq \ell \leq k$. The last inequality follows from Lemma 3.14 since

$$\left\|\mathcal{R}^f_{\tilde{B}}(\tilde{a}_j)\right\|_2 \leq \sqrt{\frac{L}{\mu}}||\tilde{a}_j||_2 \leq \sqrt{\frac{L}{\mu}}(K+\epsilon) \leq 2\sqrt{\frac{L}{\mu}}K.$$

As $\delta \leq \frac{1}{2}$, we have

$$\frac{1}{8}\mu\delta(1-\delta)\sigma^2 \geq \frac{1}{16}\mu\sigma^2\delta = \frac{1}{16}\mu\sigma^2\left(\frac{72\epsilon K L^{3/2}}{\sigma^2\mu^{3/2}}\right) = \frac{9}{2}\epsilon K\frac{L^{3/2}}{\mu^{3/2}}.$$

Combining this inequality with (3.6), we obtain $f\big(\mathcal{R}^f_{\tilde{B}}(\tilde{m}_i)\big) < \max_j f\big(\mathcal{R}^f_{\tilde{B}}(\tilde{a}_j)\big)$, a contradiction since $\tilde{m}_i$ should maximize $f\big(\mathcal{R}^f_{\tilde{B}}(.)\big)$ among the columns of $\tilde{M}$ and the $\tilde{a}_j$'s are among the columns of $\tilde{M}$.

To prove (3.4), we use (3.3) and observe that

$$m_i = (1-\delta')w_\ell + \sum_{k\neq\ell}\beta_k w_k \qquad \text{for some } \ell \text{ and } 1-\delta' \geq 1-\delta$$

so that $\sum_{k\neq\ell}\beta_k \leq \delta' \leq \delta$. Therefore,

$$\|m_i - w_\ell\|_2 = \left\|-\delta' w_\ell + \sum_{k\neq\ell}\beta_k w_k\right\|_2 \leq 2\delta'\max_j ||w_j||_2 \leq 2\delta' K \leq 2K\delta,$$

which gives

$$||\tilde{m}_i - w_\ell||_2 \leq ||(\tilde{m}_i - m_i) + (m_i - w_\ell)||_2 \leq \epsilon + 2K\delta$$

for some $1 \leq \ell \leq k$. ■

We can now prove the robustness of SNPA when W has full column rank.

Theorem 3.17. *Let $\tilde{M} = WH + N \in \mathbb{R}^{m \times n}$ satisfy Assumption 1 with $m \geq r$, and let f satisfy Assumption 2 with strong convexity parameter μ and its gradient with Lipschitz constant L. Let us denote $K = K(W)$ and $\sigma = \sigma(W)$, and let $||N||_{1,2} \leq \epsilon$ with*

$$\epsilon < \min\left(\frac{\alpha\mu}{4L}, \frac{\sigma}{2\sqrt{2r}}\right)\left(1 + 144\frac{K^2}{\sigma^2}\frac{L^{3/2}}{\mu^{3/2}}\right)^{-1}. \tag{3.7}$$

Let also $\mathcal{K}$ be the index set of cardinality r extracted by SNPA. Then there exists a permutation π of $\{1, 2, \ldots, r\}$ such that

$$\max_{1 \leq j \leq r} ||\tilde{m}_{\mathcal{K}(j)} - w_{\pi(j)}||_2 \leq \bar{\epsilon} = \epsilon\left(1 + 144\frac{K^2}{\sigma^2}\frac{L^{3/2}}{\mu^{3/2}}\right).$$

Proof. The result follows using Theorem 3.16 inductively with

$$C = \left(1 + 144\frac{K^2}{\sigma^2}\frac{L^{3/2}}{\mu^{3/2}}\right).$$

The matrix B in Theorem 3.16 corresponds to the columns of W extracted so far by SNPA (note that at the first step B is the empty matrix), while the columns of A correspond to the columns of W not yet extracted. By Theorem 3.16, if the columns of B are at a distance at most $\bar{\epsilon} = C\epsilon$ from some columns of W, then the next extracted column will be a column of the matrix A and be at a distance at most $\bar{\epsilon}$ from another column of W. The results therefore follows by induction.

Note that $\epsilon < \frac{\sigma}{2C\sqrt{2r}}$ implies $\epsilon < \frac{\sigma^2\mu^{3/2}}{144KL^{3/2}}$ since $\sigma \leq K$ and

$$\epsilon < \frac{\sigma}{C} = \frac{\sigma}{1 + 144\frac{K^2}{\sigma^2}\frac{L^{3/2}}{\mu^{3/2}}} \leq \frac{\sigma^3\mu^{3/2}}{144K^2L^{3/2}} \leq \frac{\sigma^2\mu^{3/2}}{144KL^{3/2}}. \qquad ■$$

Finally, SNPA with $f(.) = ||.||_2^2$ satisfies the same error bound as SPA when W has full column rank.

Theorem 3.18. *Let $\tilde{M}$ be a near-separable matrix satisfying Assumption 1 where W has full column rank. If $\epsilon \leq \mathcal{O}\big(\frac{\sigma_{\min}(W)}{\sqrt{r}\kappa^2(W)}\big)$, then SNPA with $f(.) = ||.||_2^2$ identifies all the columns of W up to error $\mathcal{O}\left(\epsilon\,\kappa^2(W)\right)$.*

Proof. This follows directly from $K(W) \leq \sigma_{\max}(W)$, $\alpha(W) \geq \sigma(W)$ (Lemma 3.8), Theorem 3.16, and the fact that $\mu = L = 2$ for $f(x) = ||x||_2^2$. ■

3.5. Generalization to column-rank-deficient W.

SNPA can be applied to a broader class of near-separable matrices: In fact, the assumption that W must have full column rank in Theorem 3.18 is *not necessary* for SNPA to be robust to noise. We now define the parameter $\beta(W)$, which will replace $\sigma(W)$ in the robustness analysis of SNPA.

Definition (parameter β). Given $W \in \mathbb{R}^{m\times r}$ and f satisfying Assumption 2, we define

$$\nu_\beta(W) = \min_j \left\| \mathcal{R}^f_{W(:,\mathcal{J})}(w_j) \right\|_2 \qquad \text{with } \mathcal{J} = \{1,2,\dots,r\}\backslash\{j\},$$

$$\gamma_\beta(W) = \min_{i\neq j} \left\| \mathcal{R}^f_{W(:,\mathcal{J})}(w_i) - \mathcal{R}^f_{W(:,\mathcal{J})}(w_j) \right\|_2 \qquad \text{with } \mathcal{J} \subseteq \{1,2,\dots,r\}\backslash\{i,j\},$$

and

$$\beta(W) = \min\left(\nu_\beta(W), \frac{1}{\sqrt{2}}\gamma_\beta(W) \right).$$

The quantity $\beta(W)$ is the minimum between

- the norms of the residuals of the projections of the columns of W onto the convex hull of the other columns of W, and
- the distances between these residuals.

For example, if the columns of W are the vertices of a triangle in the plane ($W \in \mathbb{R}^{2\times 3}$), SPA can extract only two columns (the residual will be equal to zero after two steps because $\text{rank}(W) = 2$), while, in most cases, $\beta(W) > 0$ and SNPA is able to identify correctly the three vertices even in the presence of noise. In particular, $\beta(W)$ is larger than $\sigma(W)$ and hence is positive for matrices with full column rank.

Lemma 3.19. *For any $W \in \mathbb{R}^{m\times r}$, $\beta(W) \geq \sigma(W)$.*

Proof. This follows directly from Lemma 3.9. ■

However, the condition $\beta(W) > 0$ is not necessarily satisfied for any set of vertices w_i's, and hence SNPA is not robust to noise for any matrix W with $\alpha(W) > 0$. For example, in the case of a triangle in the plane, $\beta(W) = 0$ if and only if the residuals of the projections of two columns of W onto the segment joining the origin and the last column of W are equal to one another (this requires that they be on the same side and at the same distant from that segment). It is the case, for example, for the following matrix:

$$W = \begin{pmatrix} 4 & 1 & 3 \\ 0 & 1 & 1 \end{pmatrix}$$

with $\beta(W) = 0$ while $\alpha(W) > 0$, and any data point on the segment $[W(:,2), W(:,3)]$ could be extracted at the second step of SNPA. In fact, we have

$$\mathcal{R}^f_{W(:,1)}\Big(W(:,2)\Big) = \mathcal{R}^f_{W(:,1)}\Big(W(:,3)\Big) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

We can link $\beta(W)$ and $\alpha(W)$ as follows.

Lemma 3.20. *For any $W \in \mathbb{R}^{m\times r}$, $\alpha(W) \geq \sqrt{\frac{\mu}{L}}\beta(W)$.*

Proof. Denoting $\mathcal{J} = \{1,2,\dots,r\}\backslash\{j\}$, we have

$$\begin{aligned}\alpha(W) &= \min_{1\le j\le r}\min_{x\in\Delta}||w_j - W(:,\mathcal{J})x||_2\\ &\ge \sqrt{\frac{2}{L}}\min_{1\le j\le r}\min_{x\in\Delta} f\left(w_j - W(:,\mathcal{J})x\right)\\ &= \sqrt{\frac{2}{L}}\min_{1\le j\le r} f\left(\mathcal{R}^f_{W(:,\mathcal{J})}(w_j)\right)\\ &\ge \sqrt{\frac{\mu}{L}}\min_{1\le j\le r}\left\|\mathcal{R}^f_{W(:,\mathcal{J})}(w_j)\right\|_2 \ge \sqrt{\frac{\mu}{L}}\beta(W). \qquad \blacksquare\end{aligned}$$

In the following, we get rid of $\sigma(W)$ in the robustness analysis of SNPA to replace it with $\beta(W)$. In Theorems 3.16 and 3.17, $\sigma(W)$ was used to lower bound $\omega\big(\mathcal{R}^f_{\tilde{B}}(A)\big)$; see, in particular, (3.5). Hence we need to lower bound $\omega\big(\mathcal{R}^f_{\tilde{B}}(A)\big)$ using $\beta(W)$. The remaining of the proof follows exactly the same steps as the proofs of Theorem 3.16 and 3.17, and we do not repeat them here.

Theorem 3.21. *Let $\tilde{M} = WH + N \in \mathbb{R}^{m\times n}$ be a near-separable matrix satisfying Assumption 1, and let f satisfy Assumption 2 with strong convexity parameter μ and its gradient with Lipschitz constant L. Let us denote $K = K(W)$ and $\beta = \beta(W)$, and let $||N||_{1,2} \le \epsilon$ with*

$$\epsilon < \min\left(\frac{\beta^2\mu^{3/2}}{144KL^{3/2}}, \frac{\alpha\mu}{4LC}, \frac{\beta^2\mu}{128KLC}\right), \tag{3.8}$$

with $C = \left(1 + 144\frac{K^2}{\beta^2}\frac{L^{3/2}}{\mu^{3/2}}\right)$. Let also $\mathcal{K}$ be the index set of cardinality r extracted by SNPA. Then there exists a permutation π of $\{1,2,\dots,r\}$ such that

$$\max_{1\le j\le r}||\tilde{m}_{\mathcal{K}(j)} - w_{\pi(j)}||_2 \le \bar{\epsilon} = C\epsilon.$$

Proof. Using the first four assumptions of Theorem 3.16 (that is, all assumptions of Theorem 3.16 but the upper bound on ϵ) and denoting $\beta = \beta(W)$, we show in Appendix B that

$$\omega\left(\mathcal{R}^f_{\tilde{B}}(A)\right) \ge \beta - 2\sqrt{\frac{6KL\bar{\epsilon}}{\mu}}.$$

Therefore, if

$$2\sqrt{\frac{6KL\bar{\epsilon}}{\mu}} \le \frac{\beta}{2} \qquad \iff \qquad \bar{\epsilon} = C\epsilon \le \frac{\beta^2\mu}{96KL},$$

then $\omega\big(\mathcal{R}^f_{\tilde{B}}(A)\big) \ge \frac{\beta}{2}$. Hence $\sigma(W)$ can be replaced with $\beta(W)$ in (3.5), (3.6), and the following derivations, and Theorem 3.16 applies under the same conditions, except for $\delta = \frac{72\epsilon KL^{3/2}}{\beta^2\mu^{3/2}}$ and the condition on ϵ given by (3.8). $\blacksquare$

Let us denote $1 \le \kappa_\beta(W) = \frac{K(W)}{\beta(W)} \le \kappa(W)$.

Theorem 3.22. *Let $\tilde{M}$ be a near-separable matrix (see Assumption 1) where W satisfies $\beta(W) > 0$. If $\epsilon \leq \mathcal{O}\big(\frac{\beta(W)}{\kappa_\beta^3(W)}\big)$, then SNPA with $f(.) = ||.||_2^2$ identifies all the columns of W up to error $\mathcal{O}\big(\epsilon\, \kappa_\beta^3(W)\big)$.*

Proof. This follows from Theorem 3.21, Lemma 3.20, and $\mu = L = 2$ for $f(.) = ||.||_2^2$. ■

Note that the bounds are cubic in $\kappa_\beta(W)$ while they were quadratic in $\kappa(W)$. The reason is that the lower bound on $\omega\big(\mathcal{R}^f_{\tilde{B}}(A)\big)$ based on $\beta(W)$ is of the form $\beta(W) - \mathcal{O}(\sqrt{\bar{\epsilon}})$, while the one based on $\sigma(W)$ was of the form $\sigma(W) - \mathcal{O}(\sqrt{r\bar{\epsilon}})$. However, the dependence on $\sqrt{r}$ has disappeared.

Remark 1. Because SNPA extracts the columns of W in a specific order, the parameter $\beta(W)$ could be replaced by the following larger parameter. Assume without loss of generality that the columns of W are ordered in such a way that the kth column of W is extracted at the kth step of SNPA. Then, $\beta(W)$ in the robustness analysis of SNPA can be replaced with the larger

$$\beta'(W) = \left(\min_{i<k} \left\| \mathcal{R}^f_{W(:,1:i)}(w_k) \right\|_2, \frac{1}{\sqrt{2}} \min_{1 \leq i \leq r-2, i<k\neq l} \left\| \mathcal{R}^f_{W(:,1:i)}(w_k) - \mathcal{R}^f_{W(:,1:i)}(w_l) \right\|_2 \right).$$

It is also interesting to notice that $\beta(W)$ could be equal to zero for some function f while being positive for others. Hence, ideally, the function f should be chosen such that $\beta(W)$ is maximized. Note, however, that W is unknown and $\beta(W)$ is expensive to compute (although $\beta'(W)$ could be used instead), so the problem seems rather challenging. This is a topic for further research.

3.6. Improvements using postprocessing, preconditioning, and outlier detection.

It is possible to improve the performance of SNPA, the same way it was done for SPA:

- A first possibility is to use *postprocessing* [3, Alg. 4]. Let $\mathcal{K}$ be the set of indices extracted by SNPA, and denote $\mathcal{K}(k)$ the index extracted at step k. For $k = 1, 2, \dots r$, the postprocessing
  - projects each column of the data matrix onto the convex hull of $M(:, \mathcal{K} \backslash \{\mathcal{K}(k)\})$,
  - identifies the column of the corresponding projected matrix maximizing $f(.)$ (say the kth), and
  - updates $\mathcal{K} = \mathcal{K} \backslash \{\mathcal{K}(k)\} \cup \{k'\}$.

  This allows one to improve the bound on the error of Theorem 1.1 from $\mathcal{O}\left(\epsilon\, \kappa^2(W)\right)$ to $\mathcal{O}\left(\epsilon\, \kappa(W)\right)$.
- A second possibility is to *precondition* the input near-separable matrix [20], making the condition number of W constant, while multiplying the error by a factor of at most $\sigma_{\min}^{-1}(W)$. This allows one to improve the bound on the noise of Theorem 1.1 from $\epsilon \leq \mathcal{O}\big(\frac{\sigma_{\min}(W)}{\sqrt{r}\kappa^2(W)}\big)$ to $\epsilon \leq \mathcal{O}\big(\frac{\sigma_{\min}(W)}{r\sqrt{r}}\big)$ and the bound on the error from $\mathcal{O}\left(\epsilon\, \kappa^2(W)\right)$ to $\mathcal{O}\left(\epsilon\, \kappa(W)\right)$.
- A third possibility for improvement is to deal with outliers. They will be identified by SNPA along with the columns of W and can be discarded in a second step by computing the optimal weights needed to reconstruct all columns of the input matrix with the extracted columns [12, 21, 19].

We do not focus in this paper on these improvements, as they are straightforward applications of existing techniques. Our focus in section 4 is rather to show the better performance of SNPA compared to the original SPA.

3.7. Choice of the function f. According to our theoretical analysis (see Theorems 3.17 and 3.21), the best possible case for the function f is to have $\mu = L$, in which case $f(x) = ||x||_2^2$. However, our analysis considers a worst-case scenario and, in some cases, it might be beneficial to use other functions f. For example, if the noise is sparse (that is, only a few entries of the data matrix are perturbed), it was shown that it is better to use ℓ_p norms with $1 < p < 2$ (that is, use $f(x) = ||x||_p^p$); see the discussion in [21, sect. 4] and [1] for more numerical experiments. More generally, it would be particularly interesting to analyze good choices for the function f depending on the noise model. Note also that the assumptions on the function f can be relaxed to the condition that the gradient of f be continuously differentiable and that f be locally strongly convex, and hence our result applies, for example, to ℓ_p norms for $1 < p < +\infty$; see [21, Rem. 3].

4. Numerical experiments. In this section, we compare the following algorithms:

- **SPA**: the successive projection algorithm; see Algorithm SPA.
- **SNPA**: the successive nonnegative projection algorithm; see Algorithm SNPA with $f(x) = ||x||_2^2$.
- **XRAY**: the recursive algorithm similar to SNPA [25]. It extracts columns recursively and projects the data points onto the convex cone generated by the columns extracted so far. (We use in this paper the variant referred to as *max*.)

Our goal is to do the following:

1. Illustrate our theoretical result, namely that SNPA applies to a broader class of nonnegative matrices and is more robust than SPA; see section 4.1.
2. Show that SNPA can be used successfully on real-world hyperspectral images; see section 4.2, where the popular Urban data set is used for comparison.

We also show that SNPA is more robust to noise than XRAY (in some cases significantly).

The MATLAB code is available from https://sites.google.com/site/nicolasgillis/. All tests are performed using MATLAB on a laptop Intel CORE i5-3210M CPU @2.5GHz 2.5GHz 6Go RAM.

Remark 2 (comparison with standard NMF algorithms). We do not compare near-separable NMF algorithms to standard NMF algorithms, whose goal is to solve

$$\min_{U \geq 0, V \geq 0} ||M - UV||_F^2. \tag{4.1}$$

In fact, we believe these two classes of algorithms, although closely related, are difficult to compare. In particular, the solution of (4.1) is in general nonunique,[3] even for a separable matrix $M = W[I_r, H']\Pi$. This is the case, for example, if the support (the set of nonzero entries) of a column of W contains the support of another column [15, Rem. 7]; in particular, most hyperspectral images have a dense endmember matrix W and hence have a nonunique NMF decomposition. (More generally, it will be the case if and only if the NMF of W is

[3] A solution (U', V') is considered to be different from (U, V) if it cannot be obtained by permutation and scaling of the columns of U and rows of V; see [15] for more details on nonuniqueness issues for NMF.

nonunique.) Therefore, without regularization, standard NMF algorithms will in general fail to identify the matrix W from a near-separable matrix $\tilde{M} = W[I_r, H']\Pi + N$. However, it is likely they will generate a solution with smaller error $||\tilde{M} - UV||_F^2$ than near-separable NMF algorithms because they have more degrees of freedom.

It is interesting to note that near-separable NMF algorithms can be used as good initialization strategies for standard NMF algorithms (which usually require some initial guess for U and V); see the discussion in [17] and the references therein.

4.1. Synthetic data sets.

For well-conditioned matrices, $\beta(W)$ and $\sigma(W)$ are close to one another (in fact, $\beta(I_r) = \sigma(I_r) = 1$), in which case we observed that SPA and SNPA provide very similar results (in most cases, they extract the same index set). Therefore, in this section, our focus will be on the following:

(i) Near-separable matrices for which W does not have full column rank ($\text{rank}(W) < r$) to illustrate the fact that SNPA applies to a broader class of nonnegative matrices. We will refer to this case as the "rank-deficient" case; see section 4.1.1.

(ii) Ill-conditioned matrices to illustrate the fact that SNPA is more tolerant to noise than SPA; in fact, our analysis suggests it is the case when $\sigma(W) \ll \beta(W)$. We will refer to this case as the "ill-conditioned" case; see section 4.1.2.

In both the rank-deficient and ill-conditioned cases, we take $r = 20$ and generate the matrices H and N (to obtain near-separable matrices $\tilde{M} = WH + N = W[I_r, H'] + N$) in two different ways (as in [21]):

1. *Dirichlet.* $H = [I_r, I_r, H']$ so that each column of W is repeated twice, while H' has 200 columns (hence $n = 240$) generated at random following a Dirichlet distribution with parameters drawn uniformly at random in the interval $[0, 1]$. The repetition of the columns of the identity matrix allows one to check whether algorithms are sensitive to duplicated columns of W (some near-separable NMF algorithms are; see, e.g., [13, 12, 7]). Each entry of the noise N is drawn following a normal distribution $\mathcal{N}(0, 1)$ and then multiplied by the parameter δ.
2. *Middle points.* $H = [I_r, H']$, where each column of H' has exactly two nonzero entries equal to 0.5 (hence $n = r + \binom{r}{2} = 210$). The columns of W are not perturbed (that is, $N(:, 1:r) = 0$), while the remaining ones are perturbed towards the outside of the convex hull of the columns of W (that is, $N(:, j) = \delta(M(:, j) - \bar{w})$ for all j), where $\bar{w} = \frac{1}{r}\sum_{k=1}^{r} w_k$ and δ is a parameter.

4.1.1. Rank-deficient case.

To generate near-separable matrices with $\text{rank}(W) < r$, we take $m = 10$ (since $r = 20$, these matrices cannot have full column rank) while each entry of the matrix $W \in [0, 1]^{m \times r}$ is drawn uniformly at random in the interval [0,1]. Moreover, we check that each column of W is not too close to the convex cone generated by the other columns of W by making sure that for all j,

$$\min_{x \geq 0} ||W(:, j) - W(:, \mathcal{J})x||_2 \geq 0.01||W(:, j)||_2 \quad \text{with } \mathcal{J} = \{1, 2, \dots, r\} \backslash \{j\}.$$

If this condition is not met (which occurs very rarely), we generate another matrix until it is. Hence, we have that $\text{rank}(W) = \text{rank}(M) = 10$ while $M = WH$ is 20-separable with $\alpha(W) > 0$. (By a slight abuse of language, we refer to this case as the "rank-deficient case,"

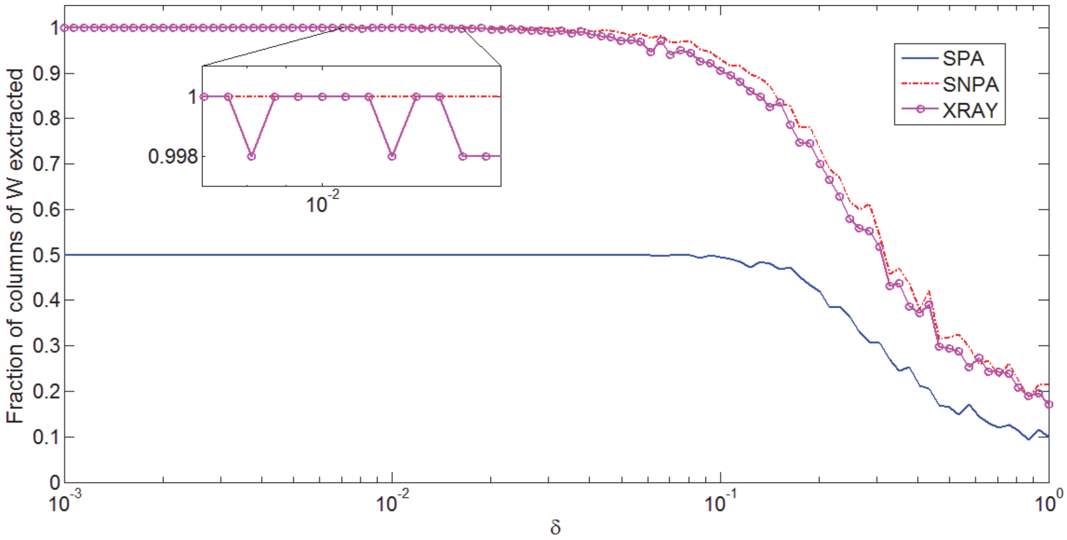

Figure 1. *Comparison of the different near-separable NMF algorithms on rank-deficient data sets ("Dirichlet" type).*

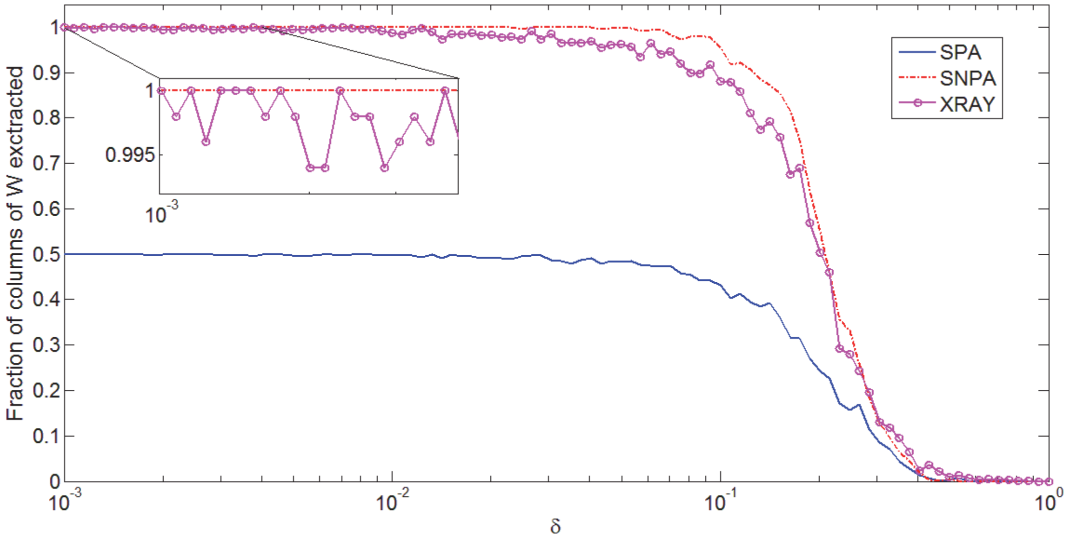

Figure 2. *Comparison of the different near-separable NMF algorithms on rank-deficient data sets ("Middle points" type).*

although the matrix M actually has full (row) rank—to be more precise, we should refer to it as the "column-rank-deficient case.") It is interesting to point out that, on average, $\beta'(W) \approx 0.2$ (see Remark 1 for the definition of $\beta'(W)$, which can replace $\beta(W)$ in the analysis of SNPA) while $\kappa_{\beta'}(W) = \frac{\max_j ||W(:,j)||_2}{\beta'(W)} \approx 10$.

For 100 different values of the noise parameter δ (using `logspace(-3,0,100)`), we generate 25 matrices of each type: Figure 1 (resp., Figure 2) displays the fraction of columns of W correctly identified by the different algorithms for the experiment "Dirichlet" (resp., "Middle

Table 1

Robustness for the rank-deficient "Dirichlet" experiment, that is, the largest value of δ for which all (resp., 95% of the) columns of W are correctly identified, and average running time in seconds of the different near-separable NMF algorithms.

	SPA	SNPA	XRAY
Robustness (100%)	0	$\mathbf{1.7^{*}10^{-2}}$	$7.6^{*}10^{-3}$
Robustness (95%)	0	$\mathbf{8.9^{*}10^{-2}}$	$6.1^{*}10^{-2}$
Time (seconds)	< 0.01	7.67	1.16

Table 2

Robustness and average running time for the rank-deficient "Middle points" experiment.

	SPA	SNPA	XRAY
Robustness (100%)	0	$\mathbf{2.3^{*}10^{-2}}$	10^{-3}
Robustness (95%)	0	$\mathbf{10^{-1}}$	$5.5^{*}10^{-2}$
Time (seconds)	< 0.01	7.83	1.05

points").

SPA is significantly faster than XRAY and SNPA since the projection at each step simply amounts to a matrix-vector product, while, for SNPA and XRAY, the projection requires solving n linearly constrained least squares problems in r variables. XRAY is faster than SNPA because the projections are simpler to compute. (Also, a different algorithm was used: XRAY requires solving least squares problems in the nonnegative orthant, which is solved with an efficient coordinate descent method from [18].)

Tables 1 and 2 give the robustness and the average running time for both experiments.

For both experiments ("Dirichlet" and "Middle points"), SPA cannot extract more than 10 columns. In fact, the input matrix has only 10 rows; hence the residual computed by SPA is equal to zero after 10 steps. SNPA is able to identify correctly the 20 columns of W, and it does so for larger noise levels than XRAY. In particular, for the "Middle points" experiment, XRAY performs rather poorly (in terms of robustness) because it does not deal very well with data points on the faces of the convex hull of the columns of W (in this case, on the middle of the segment joining two vertices); see the discussion in [25]. For example, for $\delta = 0.1$, SNPA identifies more than 95% of the columns of W, while XRAY identifies less than 90%.

4.1.2. Ill-conditioned case. In this section, we perform an experiment very similar to the third and fourth experiments in [21] to assess the robustness to noise of the different algorithms on ill-conditioned matrices. We take $m = 20$, while each entry of the matrix $W \in [0,1]^{m\times r}$ is drawn uniformly at random in the interval [0,1] (as in the rank-deficient case). Then the compact singular value decomposition (U, S, V^T) of W is computed (using the function `svds(M,r)` of MATLAB), and W is replaced with $U\Sigma V^T$, where Σ is a diagonal matrix with $\Sigma(i,i) = \alpha^{i-1}$ $(1 \leq i \leq r)$, where $\alpha^{r-1} = 1000$ so that $\kappa(W) = 1000$. Finally, to obtain a nonnegative matrix W, we replace W with $\max(W, 0)$ (this step is necessary because XRAY applies only to nonnegative input matrices). Note that this changes the conditioning, with the average of $\kappa(W)$ being equal to 5000. It is interesting to point out that for these matrices $\beta'(W)$ is usually much smaller than $\sigma(W)$: The order of magnitudes are

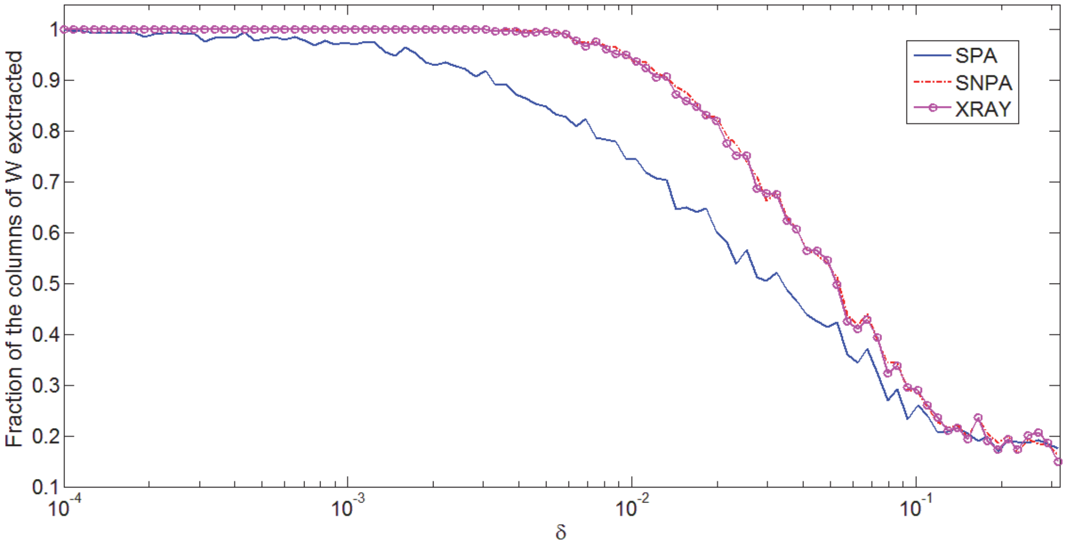

Figure 3. *Comparison of the different near-separable NMF algorithms on ill-conditioned data sets ("Dirichlet" type).*

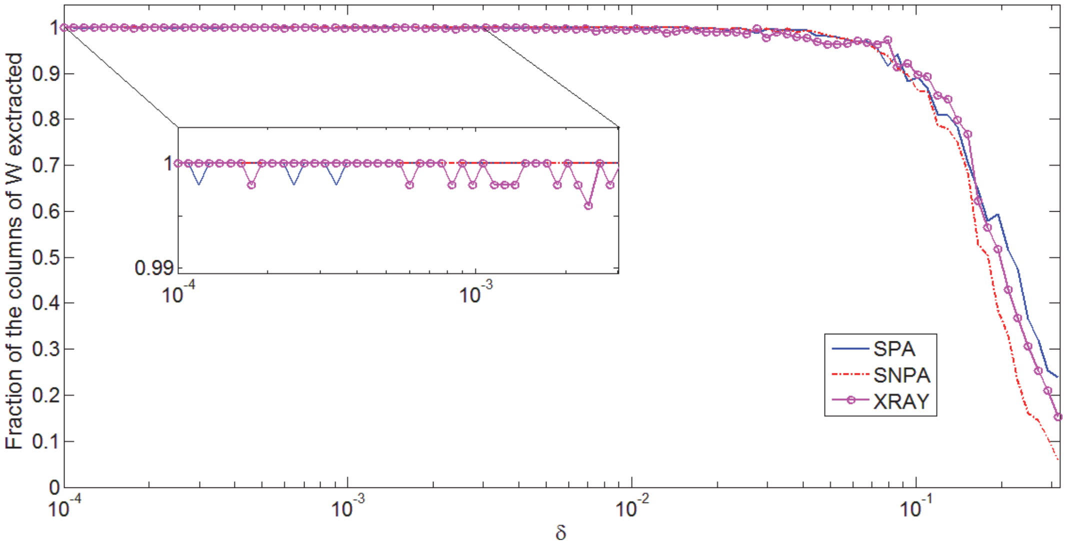

Figure 4. *Comparison of the different near-separable NMF algorithms on ill-conditioned data sets ("Middle points" type).*

$\beta'(W) \approx 10^{-2} \ll \sigma_{\min(W)} \approx 10^{-3}$, while $\kappa_{\beta'}(W) \approx 10 \ll \kappa(W) \approx 10^3$.

For 100 different values of the noise parameter δ (using `logspace(-4,-0.5,100)`), we generate 25 matrices of each type: Figure 3 (resp., Figure 4) displays the fraction of columns of W correctly identified by the different algorithms for the experiment "Dirichlet" (resp., "Middle points").

Tables 3 and 4 give the robustness and the average running time for both experiments.

For the same reasons as before, SPA is significantly faster than XRAY, which is faster

Table 3

Robustness and average running time for the ill-conditioned "Dirichlet" experiment.

	SPA	SNPA	XRAY
Robustness (100%)	10^{-4}	$\mathbf{3.1^*10^{-3}}$	$\mathbf{3.1^*10}^{-3}$
Robustness (95%)	1.44^*10^{-3}	$\mathbf{9.45^*10^{-3}}$	$\mathbf{9.45^*10^{-3}}$
Time (seconds)	< 0.01	7.43	1.08

Table 4

Robustness and average running time for the ill-conditioned "Middle points" experiment.

	SPA	SNPA	XRAY
Robustness (100%)	1.1^*10^{-4}	$\mathbf{1.6^*10^{-2}}$	1.6^*10^{-4}
Robustness (95%)	7.4^*10^{-2}	7.3^*10^{-2}	$\mathbf{8.2^*10^{-2}}$
Time (seconds)	< 0.01	9.22	1.32

than SNPA.

For both experiments ("Dirichlet" and "Middle points"), SNPA outperforms SPA in terms of robustness, as expected by our theoretical findings. In fact, SNPA is about 10 (resp., 100) times more robust than SPA for the experiment "Dirichlet" (resp., "Middle points"); that is, it identifies correctly all columns of W for the noise parameter δ 10 (resp., 100) times larger; see Tables 3 and 4. Moreover, for the "Dirichlet" experiment, SNPA identifies significantly more columns of W, even for larger noise levels (which fall outside the scope of our analysis); for example, for $\delta = 0.01$, SNPA identifies correctly about 95% of the columns of W, while SPA identifies about 75%.

For the "Dirichlet" experiment, SNPA is as robust as XRAY, while, for the "Middle points" experiment, it is significantly more robust (for the same reason as in the rank-deficient case), as it extracts correctly all columns of W for δ 100 times larger. However, the three algorithms overall perform similarly on the "Middle points" experiment in the sense that the fraction of columns of W correctly identified do not differ by more than about 5% for all $\delta \leq 0.1$.

4.2. Real-world hyperspectral image. In this section, we analyze the Urban data set[4] with $m = 162$ and $n = 307 \times 307 = 94249$. It is mainly constituted of grass, trees, dirt, road, and different roof and metallic surfaces; see Figure 5.

We run the near-separable NMF algorithms to extract $r = 8$ endmembers. As mentioned in section 3.1, Assumption 1 is naturally satisfied by hyperspectral images (up to permutation), and hence no normalization of the data is necessary. On this data set, SPA took less than half a second to run, SNPA about one minute, and XRAY about half a minute.

Figure 6 displays the extracted spectral signatures, and Figure 7 displays the corresponding abundance maps, that is, the rows of

$$H^* = \underset{H \geq 0}{\operatorname{argmin}} ||M - M(:\mathcal{K})H||_F,$$

where $\mathcal{K}$ is the set of extracted indices by a given algorithm. SPA and SNPA extract six common indices (out of the eight, the first one being different is the fourth).

[4] Available from http://www.agc.army.mil/.

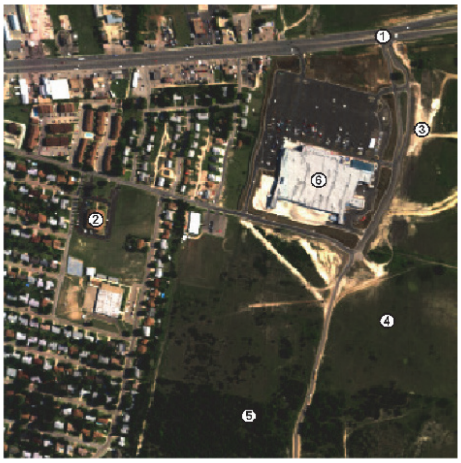

Figure 5. *Urban data set taken from an aircraft (Army Geospatial Center) with road surfaces (1), roofs* 1 *(2), dirt (3), grass (4), trees (5), and roofs* 2 *(6).*

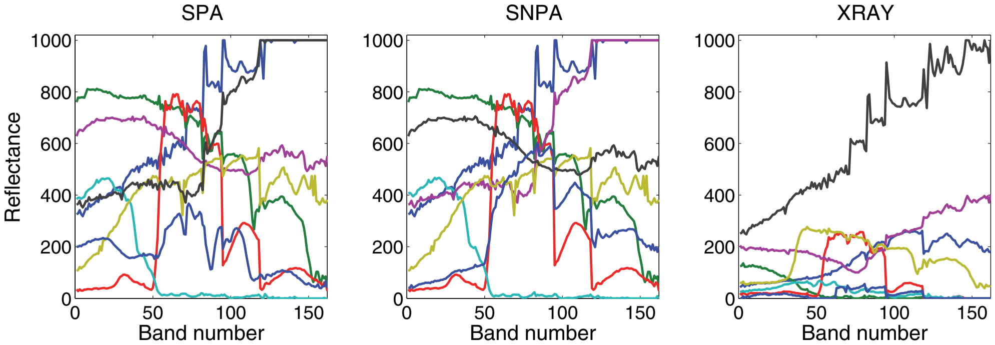

Figure 6. *Spectral signatures of the extracted endmembers.*

SNPA performs better than both SPA and XRAY, as it is the only algorithm able to distinguish the grass and trees (3rd and 8th extracted endmembers) while identifying the road surfaces (1st), the dirt (6th), and the roof tops (7th). In particular, the relative error in percent, that is,

$$100 * \frac{\min_{H\geq 0} ||M - M(:,\mathcal{K})H||_F}{||M||_F} \in [0, 100],$$

for SPA is 9.45, for XRAY 6.82, and for SNPA 5.64. In other words, SNPA is able to identify eight columns of M which can reconstruct M better. (Note that, as opposed to SPA and SNPA, which look for endmembers with large norms, XRAY focuses on extracting extreme rays of the convex cone generated by the columns of M. Hence it is likely for XRAY to identify columns with smaller norms. This explains the different scaling of the extracted endmembers in Figure 6.)

Further research includes the comparison of SNPA with other endmember extraction algorithms and its incorporation into more sophisticated techniques, e.g., where preprocessing is used to remove outliers and noise or where pure-pixel search algorithms (that is, near-separable

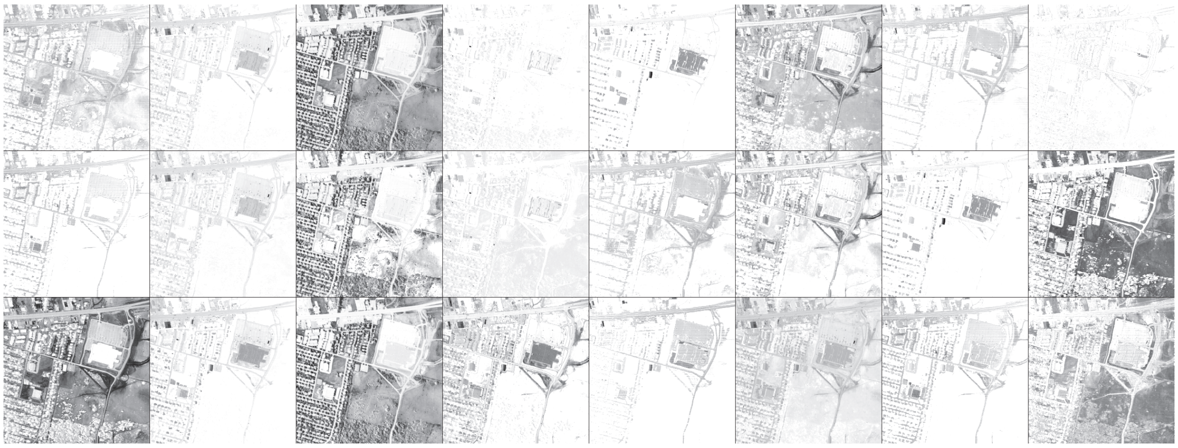

Figure 7. *Abundances maps corresponding to the extracted indices. From top to bottom:* SPA, SNPA, *and* XRAY.

NMF algorithms) are used as an initialization for more sophisticated (iterative) methods not relying on the pure-pixel assumption; see, e.g., [8], where SPA is used.

4.3. Document data sets. Because, as for SPA, SNPA requires one to normalize the input near-separable matrices not satisfying Assumption 1 (that is, near-separable matrices for which the columns of H do not belong to Δ), it may introduce distortion in the data set [25]. In particular, the normalization amplifies the noise of the columns of M with small norms (see the discussion in [19]).

In document data sets, the columns of the matrix H are usually not assumed to belong to the unit simplex, and hence normalization is necessary for applying SNPA. Therefore, XRAY should be preferred, and it has been observed that, for document data sets, SNPA and SPA perform similarly while XRAY performs better [24]; see also [25].

5. Conclusion and further research. In this paper, we have proposed a new fast and robust recursive algorithm for near-separable NMF, which we referred to as the successive nonnegative projection algorithm (SNPA). Although computationally more expensive than the successive projection algorithm (SPA), SNPA can be used to solve large-scale problems, running in $\mathcal{O}(mnr)$ operations while being more robust and applicable to a broader class of nonnegative matrices. In particular, SNPA seems to be a good alternative to SPA for real-world hyperspectral images.

There exist several algorithms robust for any near-separable matrix requiring only that $\alpha(W) > 0$ [4, 16, 19], which are therefore more general than SNPA, which requires $\beta(W) > 0$. In fact, under Assumption 1, $\alpha(W) > 0$ is a necessary condition for being able to identify the columns of W among the columns of $\tilde{M}$. However, these algorithms are computationally much more expensive (n linear programs in $\mathcal{O}(n)$ variables or a single linear program in $\mathcal{O}(n^2)$ variables have to be solved). Therefore, it would be an interesting direction for further research to develop, if possible, faster (recursive?) algorithms provably robust for any near-separable matrix $\tilde{M} = W[I_r, H'] + N$ with $\alpha(W) > 0$.

Appendix A. Fast gradient method for least squares on the simplex. Algorithm FGM is a fast gradient method for solving

$$\min_{x \in \Delta^r} f(Ax - y), \tag{A.1}$$

where $y \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times r}$. To achieve an accuracy of ϵ in the objective function, the algorithm requires $\mathcal{O}\big(\frac{1}{\sqrt{\epsilon}}\big)$ iterations. In other words, the objective function converges to the optimal value at rate $\mathcal{O}\left(\frac{1}{k^2}\right)$, where k is the iteration number.

Algorithm FGM Fast Gradient Method for Solving (A.1); see [29, p. 90].

Input: A point $y \in \mathbb{R}^m$, a matrix $A \in \mathbb{R}^{m \times r}$, a function f whose gradient is Lipschitz continuous with constant L_f, and an initial guess $x \in \Delta^r$.

Output: An approximate solution $x \approx \operatorname{argmin}_{z \in \Delta} f(Az - y)$ so that $Ax \approx \mathcal{P}_A^f(y)$.

```
$\alpha_0 \in (0,1)$; $z = x$; $L = L_f \sigma_{\max}(A)^2$ .
for $k = 1$ : maxiter do
    $x^\dagger = x$.   % Keep the previous iterate in memory.
    $x = \mathcal{P}_\Delta\Big(z - \frac{1}{L}\nabla f(Az - y)\Big)$. % $\mathcal{P}_\Delta$ is the projection on $\Delta$; see Appendix A.1.
    $z = x + \beta_k\left(x - x^\dagger\right)$,   where $\beta_k = \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2 + \alpha_{k+1}}$ with $\alpha_{k+1} \geq 0$ s.t. $\alpha_{k+1}^2 = (1 - \alpha_{k+1})\alpha_k^2$.
end for
```

Remark 3. Note that the function $g(x) = f(Ax - b)$ is not necessarily strongly convex, even if f is. This would require A to have full column rank, which we we do not assume here. If it were the case, then even faster methods could be used, although the convergence of Algorithm FGM would become linear [29].

Remark 4 (stopping condition). For the numerical experiments in section 4, we used maxiter $= 500$ and combined it with a stopping condition based on the evolution of the iterates; see the online code for more details.

A.1. Projection on the unit simplex Δ.

In Algorithm FGM, the projection onto the unit simplex Δ needs to be computed; that is, given $y \in \mathbb{R}^r$, we have to compute

$$\mathcal{P}_\Delta(y) = x^* \;=\; \operatorname*{argmin}_x \frac{1}{2}||x - y||^2 \text{ such that } x \in \Delta.$$

Let us construct the Lagrangian dual corresponding to the sum-to-one constraint (since the problem above has a Slater point, there is no duality gap):

$$\max_{\mu \geq 0} \min_{x \geq 0} \frac{1}{2}||x - y||^2 - \mu(1 - e^T x),$$

where e is the all-one vector and $\mu \geq 0$ the Lagrangian multiplier. For μ fixed, the optimal solution in x is given by

$$x^* = \max(0, y - \mu e).$$

If the sum-to-one constraint is not active, that is, $\sum_i x_i^* < 1$, we must have $\mu = 0$ and hence $x^* = \max(0, y)$ (hence this happens if and only if $\max(0, y) \in \Delta$). Otherwise the value of μ

can be computed by solving the system $\sum_i x_i^* = 1$ and $x^* = \max(0, y - \mu e)$, equivalent to finding μ satisfying $\sum_{i=1}^n \max(0, y_i - \mu) = 1$ (μ can be found easily after having sorted the entries of y).

Appendix B. Lower bound for $\omega\big(\mathcal{R}^f_{\tilde{B}}(A)\big)$ depending on $\beta([A,B])$.

In this appendix, we derive a lower bound on $\omega\big(\mathcal{R}^f_{\tilde{B}}(A)\big)$ based on $\beta([A,B])$.

Lemma B.1. *Let $x \in \mathbb{R}^m$, B ,and $\tilde{B} \in \mathbb{R}^{m\times s}$ be such that $||B - \tilde{B}||_{1,2} \leq \bar{\epsilon} \leq ||B||_{1,2}$, and let f satisfy Assumption 2. Then,*

$$\left\|\mathcal{R}^f_B(x) - \mathcal{R}^f_{\tilde{B}}(x)\right\|_2^2 \leq 12\,\frac{L}{\mu}\,\bar{\epsilon}\,||B||_{1,2}.$$

Proof. Let us denote $z = \mathcal{P}^f_B(x)$, $\tilde{z} = \mathcal{P}^f_{\tilde{B}}(x)$, and $z^* = \mathcal{P}^f_{[B,\tilde{B}]}(x)$. We have

$$\left\|\mathcal{R}^f_B(x) - \mathcal{R}^f_{\tilde{B}}(x)\right\|_2 = \left\|(x - \mathcal{P}^f_B(x)) - (x - \mathcal{P}^f_{\tilde{B}}(x))\right\|_2 = \left\|\mathcal{P}^f_B(x) - \mathcal{P}^f_{\tilde{B}}(x)\right\|_2 = \|z - \tilde{z}\|_2 .$$

Since $z^* = [B, \tilde{B}]w^*$ for some $w^* \in \Delta$, there exist $y = Bw$ and $\tilde{y} = \tilde{B}\tilde{w}$ with $w, \tilde{w} \in \Delta$ such that $||z^* - y||_2 \leq \bar{\epsilon}$ and $||z^* - \tilde{y}||_2 \leq \bar{\epsilon}$. In fact, it suffices to take $w = \tilde{w} = w^*(1{:}r) + w^*(r+1{:}2r)$ since

$$\begin{aligned} ||z^* - y||_2 &= ||[B, \tilde{B}]w^* - Bw||_2 \\ &= ||Bw^*(1:r) + \tilde{B}w^*(r+1:2r) - Bw^*(1:r) - Bw^*(r+1:2r)||_2 \\ &= ||(\tilde{B} - B)w^*(r+1:2r)||_2 \leq ||B - \tilde{B}||_{1,2} \leq \bar{\epsilon}, \end{aligned}$$

and similarly for $\tilde{y}$. Therefore, there exist some $n, \tilde{n}$ such that $z^* = y + n = \tilde{y} + \tilde{n}$ with $||n||_2, ||\tilde{n}||_2 \leq \bar{\epsilon}$. By Lemma 3.14 and the fact that $||y||_2 \leq ||B||_{1,2}$ and $||\tilde{y}||_2 \leq ||\tilde{B}||_{1,2} \leq ||B||_{1,2} + \bar{\epsilon} \leq 2||B||_{1,2}$, we have

$$f(z^*) = f(y + n) \geq f(y) - ||B||_{1,2}L\bar{\epsilon} \quad \text{and} \quad f(z^*) = f(\tilde{y} + \tilde{n}) \geq f(\tilde{y}) - 2||B||_{1,2}L\bar{\epsilon}.$$

Therefore, by the definitions of z and $\tilde{z}$,

$$f(z^*) \geq \frac{1}{2}f(y) + \frac{1}{2}f(\tilde{y}) - \frac{3}{2}||B||_{1,2}L\bar{\epsilon} \geq \frac{1}{2}f(z) + \frac{1}{2}f(\tilde{z}) - \frac{3}{2}||B||_{1,2}L\bar{\epsilon}.$$

Moreover, by the definition of z^* and strong convexity of f, we obtain

$$f(z^*) \leq f\left(\frac{1}{2}z + \frac{1}{2}\tilde{z}\right) \leq \frac{1}{2}f(z) + \frac{1}{2}f(\tilde{z}) - \frac{\mu}{8}||z - \tilde{z}||_2^2.$$

Hence, combining the above two inequalities, $||z - \tilde{z}||_2^2 \leq 12\frac{L}{\mu}||B||_{1,2}\bar{\epsilon}$. ■

Lemma B.2. *Let $x, y \in \mathbb{R}^m$, B, and $\tilde{B} \in \mathbb{R}^{m\times s}$ be such that $||B - \tilde{B}||_{1,2} \leq \bar{\epsilon} \leq ||B||_{1,2}$, and let f satisfy Assumption 2. Then,*

$$\left\|\mathcal{R}^f_{\tilde{B}}(x) - \mathcal{R}^f_{\tilde{B}}(y)\right\|_2 \geq \left\|\mathcal{R}^f_B(x) - \mathcal{R}^f_B(y)\right\|_2 - 4\sqrt{\frac{3KL}{\mu}\bar{\epsilon}}.$$

Proof. This follows directly from Lemma B.1:

$$\begin{aligned}\left\|\mathcal{R}^f_{\tilde{B}}(x)-\mathcal{R}^f_{\tilde{B}}(y)\right\|_2 &= \left\|\mathcal{R}^f_{\tilde{B}}(x)-\mathcal{R}^f_{B}(x)+\mathcal{R}^f_{B}(x)-\mathcal{R}^f_{\tilde{B}}(y)+\mathcal{R}^f_{B}(y)-\mathcal{R}^f_{B}(y)\right\|_2\\ &\geq \left\|\mathcal{R}^f_{B}(x)-\mathcal{R}^f_{B}(y)\right\|_2 - 2\sqrt{\frac{12||B||_{1,2}L\bar{\epsilon}}{\mu}}. \qquad \blacksquare\end{aligned}$$

Lemma B.3. *Let $A \in \mathbb{R}^{m\times k}$, B, and $\tilde{B} \in \mathbb{R}^{m\times s}$ be such that $||B-\tilde{B}||_{1,2} \leq \bar{\epsilon} \leq ||B||_{1,2}$, and let f satisfy Assumption 2. Then,*

$$\omega\left(\mathcal{R}^f_{\tilde{B}}(A)\right) \geq \beta([A,B]) - 2\sqrt{6\frac{L}{\mu}||B||_{1,2}\bar{\epsilon}}.$$

Proof. This follows directly from Lemmas B.1 and B.2. In fact, for all i,

$$\beta([A,B]) - \left\|\mathcal{R}^f_{\tilde{B}}(a_i)\right\|_2 \leq \left\|\mathcal{R}^f_{B}(a_i)\right\|_2 - \left\|\mathcal{R}^f_{\tilde{B}}(a_i)\right\|_2 \leq \left\|\mathcal{R}^f_{B}(a_i)-\mathcal{R}^f_{\tilde{B}}(a_i)\right\|_2 \leq \sqrt{12\,\frac{L}{\mu}\,\bar{\epsilon}\,||B||_{1,2}},$$

while, for all i, j,

$$\begin{aligned}\frac{1}{\sqrt{2}}\left\|\mathcal{R}^f_{\tilde{B}}(a_i)-\mathcal{R}^f_{\tilde{B}}(a_j)\right\|_2 &\geq \frac{1}{\sqrt{2}}\left\|\mathcal{R}^f_{B}(a_i)-\mathcal{R}^f_{B}(a_j)\right\|_2 - \frac{4}{\sqrt{2}}\sqrt{\frac{3||B||_{1,2}L}{\mu}\bar{\epsilon}}\\ &\geq \beta([A,B]) - 2\sqrt{\frac{6||B||_{1,2}L}{\mu}\bar{\epsilon}}. \qquad \blacksquare\end{aligned}$$

Acknowledgments. The author would like to thank Abhishek Kumar and Vikas Sindhwani (IBM T.J. Watson Research Center) for motivating him to study the robustness of algorithms based on nonnegative projections, for insightful discussions, and for performing some numerical experiments on document data sets. The author would also like to thank Wing-Kin Ma (The Chinese University of Hong Kong) for insightful discussions and for suggesting to analyze the noiseless case separately. Finally, the author is grateful to the reviewers for their insightful comments, which helped improve the paper significantly.

REFERENCES

[1] A. Ambikapathi, T.-H. Chan, C.-Y. Chi, and K. Keizer, *Hyperspectral data geometry based estimation of number of endmembers using p-norm based pure pixel identification*, IEEE Trans. Geosci. Remote Sensing, 51 (2013), pp. 2753–2769.

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

[3] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu, *A practical algorithm for topic modeling with provable guarantees*, in Proceedings of the International Conference on Machine Learning (ICML '13), Vol. 28, 2013, pp. 280–288.

[4] S. Arora, R. Ge, R. Kannan, and A. Moitra, *Computing a nonnegative matrix factorization – provably*, in Proceedings of the 44th Annual ACM Symposium on Theory of Computing (STOC '12), 2012, pp. 145–162.

[5] S. Arora, R. Ge, and A. Moitra, *Learning topic models – going beyond SVD*, in Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science (FOCS '12), 2012, pp. 1–10.

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

[7] V. Bittorf, B. Recht, E. Ré, and J.A. Tropp, *Factoring nonnegative matrices with linear programs*, in Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS '12), 2012, pp. 1223–1231.

[8] T.-H. Chan, W.-K. Ma, A. Ambikapathi, and C.-Y. Chi, *A simplex volume maximization framework for hyperspectral endmember extraction*, IEEE Trans. Geosci. Remote Sensing, 49 (2011), pp. 4177–4193.

[9] T.-H. Chan, W.-K. Ma, C.-Y. Chi, and Y. Wang, *A convex analysis framework for blind separation of non-negative sources*, IEEE Trans. Signal Process., 56 (2008), pp. 5120–5134.

[10] K. Devarajan, *Nonnegative matrix factorization: An analytical and interpretive tool in computational biology*, PLoS Comput. Biol., 4 (2008), e1000029.

[11] W. Ding, M.H. Rohban, P. Ishwar, and V. Saligrama, *Topic discovery through data dependent and random projections*, in Proceedings of the International Conference on Machine Learning (ICML '13), Vol. 28, 2013, pp. 471–479.

[12] E. Elhamifar, G. Sapiro, and R. Vidal, *See all by looking at a few: Sparse modeling for finding representative objects*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '12), 2012, pp. 1600–1607.

[13] E. Esser, M. Moller, S. Osher, G. Sapiro, and J. Xin, *A convex model for nonnegative matrix factorization and dimensionality reduction on physical space*, IEEE Trans. Image Process., 21 (2012), pp. 3239–3252.

[14] X. Fu, W.-K. Ma, T.-H. Chan, J.M. Bioucas-Dias, and M.-D. Iordache, *Greedy algorithms for pure pixel identification in hyperspectral unmixing: A multiple-measurement vector viewpoint*, in Proceedings of the European Signal Processing Conference (EUSIPCO '13), 2013.

[15] N. Gillis, *Sparse and unique nonnegative matrix factorization through data preprocessing*, J. Mach. Learn. Res., 13 (2012), pp. 3349–3386.

[16] N. Gillis, *Robustness analysis of Hottopixx, a linear programming model for factoring nonnegative matrices*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 1189–1212.

[17] N. Gillis, *The why and how of nonnegative matrix factorization*, in Regularization, Optimization, Kernels, and Support Vector Machines, J.A.K. Suykens, M. Signoretto, and A. Argyriou, eds., Chapman & Hall/CRC, Boca Raton, FL, to appear (arXiv:1401.5226).

[18] N. Gillis and F. Glineur, *Accelerated multiplicative updates and hierarchical* ALS *algorithms for nonnegative matrix factorization*, Neural Comput., 24 (2012), pp. 1085–1105.

[19] N. Gillis and R. Luce, *Robust near-separable nonnegative matrix factorization using linear optimization*, J. Mach. Learn. Res., 15 (2014), pp. 1249–1280.

[20] N. Gillis and S.A. Vavasis, *Semidefinite Programming Based Preconditioning for More Robust Near-Separable Nonnegative Matrix Factorization*, preprint, arXiv:1310.2273, 2013.

[21] N. Gillis and S.A. Vavasis, *Fast and robust recursive algorithms for separable nonnegative matrix factorization*, IEEE Trans. Pattern Anal. Mach. Intell., 36 (2014), pp. 698–714.

[22] G.H. Golub and C.F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[23] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*, Springer, Berlin, 2001.

[24] A. Kumar, *private communication*, 2013.

[25] A. Kumar, V. Sindhwani, and P. Kambadur, *Fast conical hull algorithms for near-separable nonnegative matrix factorization*, in Proceedings of the International Conference on Machine Learning (ICML '13), Vol. 28, 2013, pp. 231–239.

[26] D.D. Lee and H.S. Seung, *Learning the parts of objects by nonnegative matrix factorization*, Nature, 401 (1999), pp. 788–791.

[27] W.-K. Ma, J.M. Bioucas-Dias, P. Gader, T.-H. Chan, N. Gillis, A. Plaza, A. Ambikapathi, and C.-Y. Chi, *A signal processing perspective on hyperspectral unmixing: Insights from remote sensing*, IEEE Signal Process. Mag., 31 (2014), pp. 67–81.

[28] T. Mizutani, *Ellipsoidal rounding for nonnegative matrix factorization under noisy separability*, J. Mach. Learn. Res., 15 (2014), pp. 1011–1039.

[29] Yu. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer Academic Publishers, Boston, MA, 2004.

[30] V.P. Pauca, J. Piper, and R.J. Plemmons, *Nonnegative matrix factorization for spectral data analysis*, Lin. Alg. Appl., 406 (2006), pp. 29–47.

[31] H. Ren and C.-I. Chang, *Automatic spectral target recognition in hyperspectral imagery*, IEEE Trans. Aero. Electron. Syst., 39 (2003), pp. 1232–1249.

[32] F. Shahnaz, M.W. Berry, V.P. Pauca, and R.J. Plemmons, *Document clustering using nonnegative matrix factorization*, Inform. Process. Manag., 42 (2006), pp. 373–386.

[33] S.A. Vavasis, *On the complexity of nonnegative matrix factorization*, SIAM J. Optim., 20 (2009), pp. 1364–1377.

[34] F. Wang, T. Li, X. Wang, S. Zhu, and C. Ding, *Community discovery using nonnegative matrix factorization*, Data Min. Knowl. Discov., 22 (2011), pp. 493–521.