SIAM J. MATH. DATA SCI.
Vol. 1, No. 3, pp. 518–536



# Identifiability of Complete Dictionary Learning*

Jeremy E. Cohen† and Nicolas Gillis‡

**Abstract.** Sparse component analysis (SCA), also known as complete dictionary learning, is the following problem: Given an input matrix $M$ and an integer $r$, find a dictionary $D$ with $r$ columns and a matrix $B$ with $k$-sparse columns (that is, each column of $B$ has at most $k$ nonzero entries) such that $M \approx DB$. A key issue in SCA is identifiability, that is, characterizing the conditions under which $D$ and $B$ are essentially unique (that is, they are unique up to permutation and scaling of the columns of $D$ and rows of $B$). Although SCA has been vastly investigated in the last two decades, only a few works have tackled this issue in the deterministic scenario, and no work provides reasonable bounds in the minimum number of samples (that is, columns of $M$) that leads to identifiability. In this work, we provide new results in the deterministic scenario when the data has a low-rank structure, that is, when $D$ is (under)complete. While previous bounds feature a combinatorial term $\binom{r}{k}$, we exhibit a sufficient condition involving $\mathcal{O}(r^3/(r-k)^2)$ samples that yields an essentially unique decomposition, as long as these data points are well spread among the subspaces spanned by $r-1$ columns of $D$. We also exhibit a necessary lower bound on the number of samples that contradicts previous results in the literature when $k$ equals $r-1$. Our bounds provide a drastic improvement compared to the state of the art, and imply, for example, that for a fixed proportion of zeros (constant and independent of $r$, e.g., 10% of zero entries in $B$), one only requires $\mathcal{O}(r)$ data points to guarantee identifiability.

**Key words.** matrix factorization, dictionary learning, sparse component analysis, identifiability, uniqueness

**AMS subject classifications.** 15A23, 65F50, 94A12

**DOI.** 10.1137/18M1233339

## 1. Introduction.

In the last two decades, dictionary learning has had tremendous success in various fields, such as image processing, neuroimaging, and remote sensing; see, e.g., [16, 28, 29, 30, 35, 40] and the references therein. In fact, many applications in source separation involve data expressed as a combination of a few atoms from an appropriate but unknown basis. After the pioneer work of Olshausen and Field [33], efforts were made to derive conditions under which a "true" underlying dictionary and sparse coefficients could be recovered with certainty [5, 19] or almost surely [20, 37]. Algorithmic aspects of dictionary learning have also been extensively studied [4, 29], sometimes under the name sparse component analysis (SCA) in the signal processing community [21, 31].

The focus of this work is on the *identifiability* of SCA in the case of an undercomplete

*Received by the editors December 14, 2018; accepted for publication (in revised form) July 10, 2019; published electronically September 12, 2019.
https://doi.org/10.1137/18M1233339
**Funding:** This work was supported by the Fonds de la recherche scientifique-FNRS (incentive grant for scientific research F.4501.16). The second author's research was also supported by the European Research Council (ERC starting grant 679515) and by the Fonds de la Recherche Scientifique - FNRS and the Fonds Wetenschappelijk Onderzoek - Vlanderen (FWO) under EOS project O005318F-RG47.

†CNRS, Université de Rennes, Inria, IRISA Campus de Beaulieu, 35042 Rennes, France (jeremy.cohen@irisa.fr).
‡Corresponding author. Department of Mathematics and Operational Research, Faculté polytechnique, Université de Mons, 7000 Mons, Belgium (nicolas.gillis@umons.ac.be).

dictionary (that is, the number of atoms is smaller than the ambient dimension) in a deterministic scenario and without noise. By deterministic we mean that we will derive conditions under which the decomposition is always essentially unique. We will refer to this model as low-rank SCA (LRSCA). To the best of our knowledge, the identifiability of LRSCA has been treated in only three early works [5, 19, 22]. The main contribution of this paper is to provide new strong identifiability results for LRSCA.

In section 2.1, we formally introduce LRSCA and recall previous results for this problem. We also show some examples that give a geometric intuition for LRSCA. In section 2.1, we show how LRSCA relates to other low-rank matrix factorization models, and describe some applications of both LRSCA and the proposed identifiability results. In section 3, we prove our main results, namely Theorems 3.6 and 3.8. Both provide sufficient conditions based on lower bounds on the number of data points required to guarantee identifiability. Both theorems asymptotically lead to the same bound and require $\mathcal{O}(r^3/(r-k)^2)$ data points well spread among the subspaces spanned by $r-1$ atoms of the dictionary, where $r$ is the rank of the input matrix and $k$ is the maximum number of atoms used by each data point. For convenience, let us introduce the notation $\ell = r - k$, where the integer $\ell$ is the number of zero coefficients of the columns of $B$ and is commonly referred to as the co-sparsity level with respect to the pseudoinverse of $D$. Using this notation, our result requires $\mathcal{O}(r^3/\ell^2)$ data points to guarantee identifiability. We prove that this bound is tight in the following two cases: when $\ell$ is constant, in which case $\mathcal{O}(r^3)$ points are sufficient to guarantee identifiability, and when $\ell$ is a fixed proportion of $r$, in which case $\mathcal{O}(r)$ points are sufficient. Theorem 3.6 is weaker than Theorem 3.8, but its proof is easier to derive and the bound it provides on the number of points can be easily computed. Theorem 3.8 is based on a sequential construction of subspaces containing the data points and therefore requires construction-dependent hypotheses harder to verify. We conclude the paper by discussing some directions of further research, in particular the generalization of our results in the presence of noise, for overcomplete dictionaries, and for nonnegative coefficients.

**Notation.** Vectors are denoted as small letters $x$ and matrices as capital letters $M$. The $i$th column of matrix $B$ is denoted as $b_i$, and the $j$th entry in that column is denoted as $b_{i,j}$. The quantity $\|x\|_0$ is the so-called $\ell_0$ norm of the vector $x$ defined as the number of nonzero entries in $x$. The spark of a matrix is the smallest number $p$ such that there exists a set of $p$ columns which are linearly dependent. By extension, if a matrix $M \in \mathbb{R}^{p\times n}$ has rank $n$, we define $\operatorname{spark}(M) = n+1$. Note that if $\operatorname{spark}(M) = r+1$, then $\operatorname{rank}(M) \geq r$. Given a finite set $\mathcal{S}$, we denote its cardinality by $|\mathcal{S}|$. Given a set of vectors $\mathcal{X} = \{x_1, \ldots, x_n\}$ or a matrix $X = [x_1, \ldots, x_n]$, $\operatorname{span}(X) = \operatorname{span}(\mathcal{X})$ is the linear subspace spanned by $\mathcal{X}$.

## 2. Formalism, previous results, and geometric intuition.

Let $M$ be a real $p \times n$ matrix. The working assumption of dictionary learning is that there exist a real matrix $D$ in $\mathbb{R}^{p\times r}$ and a sparse coefficient matrix $B$ in $\mathbb{R}^{r\times n}$ such that $M = DB$. In this paper, we impose a strict sparsity constraint on the coefficient matrix $B$, namely $\|b_i\|_0 \leq k$ for all $i \leq n$ for some $k < r$. This requires that each column of $B$ has at least $\ell = r - k$ zero entries. Furthermore, we assume that matrix $M$ admits a low-rank dictionary-based representation, so that $r$ is the rank of $M$ and $r \leq p$. This leads to the following model, which we will refer to as low-rank

sparse component analysis (LRSCA):

$$\text{LRSCA:} \begin{cases} M = DB, \\ M \in \mathbb{R}^{p \times n}, D \in \mathbb{R}^{p \times r}, B \in \mathbb{R}^{r \times n}, \\ \|b_i\|_0 \leq k < r \leq p \text{ for all } i, \\ \operatorname{rank}(M) = \operatorname{rank}(D) = r. \end{cases} \tag{2.1}$$

LRSCA (2.1) does not take the noise into account. Note that without loss of generality (w.l.o.g.) a dimensionality reduction step may be performed on $M$ that leads to a square/complete dictionary learning problem with $p = r$ (this requires one to premultiply $M$ by an $r$-by-$p$ matrix which does not destroy the sparsity structure of $B$). In other words, studying undercomplete dictionary learning in the absence of noise boils down to studying square dictionary learning. Hence, one may assume w.l.o.g. that $p = r$ when analyzing (2.1).

**2.1. LRSCA: Related models and applications.** LRSCA is a sparse low-rank matrix factorization model [36], and it is also closely related to sparse PCA [13, 24, 43]. In the literature, to obtain sparse decompositions, the most widely used approach is to add an $\ell_1$ norm penalty term in the objective function [43], which has also been used for tensor factorization models [12, 27]. As far as we know, most of these works do not discuss the identifiability of the LRSCA model. Note that for tensor low-rank factorization models such as PARAFAC [25], identifiability is satisfied under mild conditions without enforcing sparsity constraint; see [14] and the references therein.

LRSCA is also related to other constrained matrix factorization models, in particular to nonnegative matrix factorization (NMF) [26]. NMF imposes that both factors $D$ and $B$ are componentwise nonnegative, which, in most cases, leads to sparse decompositions. Nonnegativity allows one to interpret the factors (see below for some examples), which is only meaningful if the NMF solution is essentially unique. Hence, finding identifiability conditions for NMF has been an active field of research; see the recent survey [18] and the references therein. However, as far as we know, there is no identifiability result specific for sparse NMF. Our results (Theorems 3.6 and 3.8) apply to sparse NMF, and therefore LRSCA could be used for the following applications:

- Spectral unmixing: given a hyperspectral image $M$ containing pixels rowwise and reflectance spectra columnwise, the dictionary $D$ contains the spectral signatures of constitutive materials (called endmembers), and coefficient matrix $B$ contains the abundance of each endmember in each pixel. Since only a few endmembers are present in each pixel (typically at most five), $B$ is sparse [10].
- Document classification: given a term-by-document matrix $M$, the dictionary $D$ is a collection of topics, while the coefficient matrix $B$ indicates which document discusses which topic. Since most documents do not discuss most topics, $B$ is sparse [26]. Note that, in this case, $D$ is also sparse, as most topics use only a small proportion of all the words.
- Audio source separation: given a spectrogram $M$ of a recording of several audio sources (rows correspond to frequency and columns to time), the dictionary $D$ contains the spectra of each source and the coefficient matrix $B$ contains the temporal activation of each source. If audio sources (speakers, instruments) are not active all the time in

the recordings, then $B$ must be sparse [34].

LRSCA is also closely related to subspace clustering [42]. In fact, as already noted in the literature, the exact sparsity constraint on the columns of $B$ is equivalent to imposing that the columns of the data matrix $M$ belong to the union of subspaces generated by subsets of the columns of $D$ with cardinality smaller than $k$. Therefore, given $M$, finding a decomposition $(D, B)$ as in (2.1) may be written as a subspace clustering problem, and many algorithm solutions have been derived using this observation [19, 21, 31]; see also [17, 41] for recent results and algorithms for subspace clustering. However, as far as we know, these works do not discuss the identifiability of LRSCA.

**2.2. Identifiability.** In this paper, we focus on the following question: Given a matrix $M \in \mathbb{R}^{p\times n}$ satisfying the LRSCA model (2.1), is the decomposition $(D, B)$ essentially unique? A decomposition $(D, B)$ is said to be *essentially unique* if for any other decomposition $(D', B')$ satisfying (2.1) we have $D = D'\Pi\mathbf{\Sigma}$ and $B = \mathbf{\Sigma}^{-1}\Pi^T B'$ for a diagonal scaling matrix $\mathbf{\Sigma}$ and a permutation matrix $\Pi$. In the remainder of this paper, we will say that two decompositions are distinct if they cannot be obtained by permutation and scaling of one another.

**2.2.1. State-of-the-art results.** Surprisingly, to the best of our knowledge, most works focusing on the identifiability of dictionary learning have not tackled directly the undercomplete case, although numerous algorithms have been proposed for this problem and variants; see section 2.1.

Most recent works on dictionary learning have tackled the identifiability question using probabilistic approaches under various a priori distributions for the locus and values of the nonzero entries of $B$; see [20] for a summary of such results. Algorithmic recovery results are also available and usually require strong assumptions on the entries of $B$. For example, the authors of [1, 11, 37] show that when $k = \mathcal{O}(\sqrt{r})$, roughly $r\log(r)$ samples are sufficient for recovery, and they propose an efficient dictionary learning algorithm, referred to as Exact Recovery of Sparsely-Used Dictionaries (ER-SpUD). Arora, Ge, and Moitra [9], Agarwal, Anandkumar, and Netrapalli [3], and Agarwal et al. [2] proposed provable learning algorithms for overcomplete dictionaries that run in polynomial time. They are also based on probabilistic distributions on the entries of $B$ and require $B$ to be sufficiently sparse. Roughly speaking, a main result in [9] requires the columns of $B$ to be $O(n^{\frac{1}{2}-\epsilon})$-sparse with $O(r^2/k^2\log r + rk^2\log r + r\log r\log 1/\epsilon)$ samples, and the main result in [2] requires the columns of $B$ to be $O(n^{\frac{1}{4}})$-sparse with $O(r^2)$ samples. Later, Arora et al. [7] provided a provable algorithm for signals that are $n/\mathrm{poly}(\log n)$-sparse. Other algorithms include a Riemannian trust-region method proposed in [38, 39] which assumes $\mathcal{O}(r)$ zeros per column of $B$, a model based on tensor Tucker decompositions [6] which requires structured sparsity, and an algorithm similar to basis pursuit which requires separability and nonnegativity [8].

These results do not consider degeneracies that could occur when studying unions of subspaces (because degeneracies happen with probability zero). This makes the study of the deterministic case rather different. To the best of our knowledge, only three works [5, 19, 22] have studied specifically the identifiability of both $D$ and $B$ under sparsity constraints in the absence of noise and without any a priori distribution on the entries of $B$, that is, no assumption is made on $B$ beyond sparsity. The first result is due to Aharon, Elad, and Bruckstein [5], which states that if

- $k < \frac{\text{spark}(D)}{2}$,
- every $k$-dimensional subspace spanned by $k$ columns of $D$ contains at least $k+1$ columns of $M$ whose spark is $k+1$, and
- no $k$-dimensional subspace contains $k+1$ columns of $M$ except those generated by $D$ (nondegeneracy),

then the decomposition $(D, B)$ is essentially unique. This result is, however, somewhat not satisfying since the number of points required in total is extremely large, proportional to $(k+1)\binom{r}{k}$, and the nondegeneracy condition, although not restrictive in practice, basically means that uniqueness is assumed from the start.

The second result by Hillar and Sommar [22, Theorem 1] does not improve on the above bound in the low-rank and deterministic setting. It states that if

- $k < \frac{\text{spark}(D)}{2}$ and
- every $k$-dimensional subspace spanned by $k$ columns of $D$ contains at least $k\binom{r}{k}$ columns of $M$ in a general position,

then the decomposition $(D, B)$ is essentially unique. Note that the notion of general position in the results of Hillar and Sommar implies the conditions of Aharon, Elad, and Bruckstein. Again, the number of samples required is large, namely $k\binom{r}{k}^2$, since there are $\binom{r}{k}$ subspaces spanned by $k$ columns of $D$.

A third seemingly more powerful result is due to Georgiev, Theis, and Cichocki [19]. Among the proposed results on identifiability in their contribution, the most commonly used one is the following. If

- $D$ is full column rank,
- $\|b_i\|_0 = r - 1$ for all $i$, and
- every subspace spanned by $r-1$ columns of $D$ contains at least $r$ columns of $M$ whose spark is equal to $r$,

then the decomposition $(D, B)$ is essentially unique. We show in the next subsection that this result is incorrect.

**2.2.2. Examples and lower bounds on the number of columns of $M$.** Before exposing our identifiability results, let us provide the reader with some geometric intuition by presenting a few simple examples. To allow two-dimensional representations of three-dimensional problems ($r = 3$), all drawings are done in a projective space, where the vectors were normalized to have their entries summing to 1 (hence, in three dimensions, hyperplanes are represented by lines).

*Example* 1 (nine points lying on three hyperplanes in dimension 3). In Figure 1, we provide an example of a matrix $M$ which admits exactly two decompositions of the form

$$M = [d_1, d_2, d_3] \begin{bmatrix} * & * & * & * & * & * & 0 & 0 & 0 \\ * & * & * & 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * & * & * & * & * \end{bmatrix},$$

where each hyperplane generated by two columns of $D$ contains exactly three points. In other words, the matrix $M$ admits two LRSCA decompositions (2.1) with $p = r = 3$, $k = 2$, $\ell = 1$, and $n = 9$. This provides a counterexample in the case $r = 3$ to the result from [19] (see section 2.2.1). In fact, the matrix $M$ contains nine data points with two

distinct decompositions, although $M$ satisfies the conditions in [19]; namely, $D$ has full column rank, and each hyperplane contains three columns of $M$ whose spark is 3. This example will generically not happen since observing three aligned points generated randomly on three two-dimensional subspaces has probability zero. Hence, most low-rank SCA models with three points on each hyperplane in the case $k = 2 = r - 1$ do not suffer from identifiability issues.

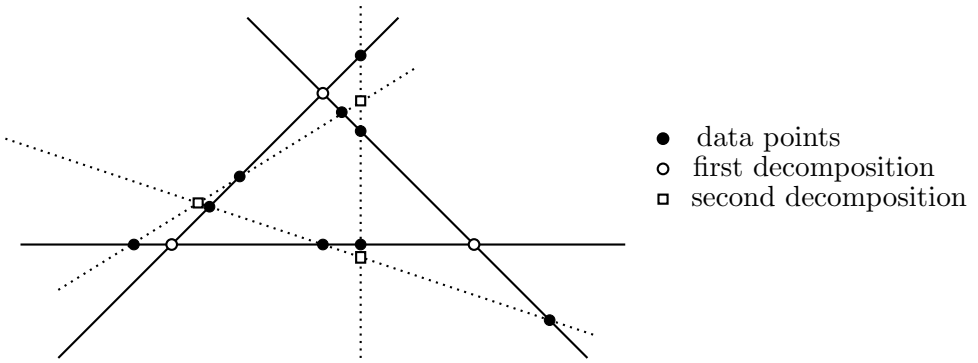


**Figure 1.** *A scenario where SCA is not unique, although it would be unique generically (that is, if the points were generated randomly on the hyperplanes generated by the combinations of any two columns of $D$).*

Inspired by Example 1, for any $r$ and $k = r - 1$, it is possible to construct a matrix $M$ with two distinct LRSCA decompositions that has $n = r^3 - 2r^2$ columns, with $r^2 - 2r$ columns with spark $r$ on each subspace spanned by $r - 1$ columns of $D$; see Algorithm 2.1.[1] Lemma 2.1 proves that the construction of Algorithm 2.1 will generate such examples with probability one. For example, for $r = 4$ and $k = 3$, Algorithm 2.1 generates a matrix $M$ with two distinct LRSCA decompositions with $n = 32$, with eight data points whose spark is 4 on each subspace spanned by three columns of $D$.

Lemma 2.1. *When using the Gaussian distribution, Algorithm 2.1 generates with probability one a matrix $M = D^{(1)}B^{(1)} = D^{(2)}B^{(2)} \in \mathbb{R}^{r \times n}$ with $n = r^3 - 2r^2$ columns, where each subspace spanned by $r - 1$ columns of $D^{(t)}$ $(t = 1, 2)$ contains $r^2 - 2r$ columns of $M$ that have spark $r$.*

*Proof.* Under the i.i.d. Gaussian distribution, with probability one, $D^{(1)}$ and $D^{(2)}$ are full rank and the $2r$ subspaces $\mathbb{F}_j^{(t)} = \text{span}(\{d_i^{(t)}\}_{i \neq j})$ for $j = 1, 2, \ldots, r$ and $t = 1, 2$ do not coincide. If the $r-2$ points on each intersection $\mathbb{F}_j^{(1)} \bigcap \mathbb{F}_l^{(2)}$ for $j, l \in \{1, 2, \ldots, r\}$ are generated as follows: (i) compute a basis of the intersection, and (ii) use i.i.d. Gaussian distribution for the weights of the linear combination in that subspace, then these points have spark $r$, with probability one. First, note that these intersections exist and have dimension $r - 2$, so that the $r - 2$ points have spark $r - 1$ with probability one. Note that the intersections $\mathbb{F}_j^{(1)} \bigcap \mathbb{F}_l^{(2)}$ of dimension $r - 2$ define $r^2$ subspaces that do not coincide, with probability one. Note also that each subspace $\mathbb{F}_j^{(k)}$ contains exactly $r(r - 2)$ points whose spark cannot be larger than $r$ since it has dimension $r - 1$. With probability one, the spark is exactly $r$: $r - 1$ columns

[1] Algorithm 2.1 is available online from https://sites.google.com/site/nicolasgillis/.

**Algorithm 2.1.** Generating matrix $M$ with distinct LRSCA decompositions (2.1) with $r^3 - 2r^2$ columns in the case $k = r - 1$.

**INPUT:** An integer $r$.

**OUTPUT:** Data matrix $M \in \mathbb{R}^{r \times n}$ with $n = r^3 - 2r^2$ columns, and two decompositions $(D^{(1)}, B^{(1)})$ and $(D^{(2)}, B^{(2)})$ for $M = D^{(1)}B^{(1)} = D^{(2)}B^{(2)}$ satisfying (2.1) and with $r^2 - 2r$ data points with spark $r$ on each subspace spanned by $r - 1$ columns of $D^{(i)}$ $(i = 1, 2)$.

1/ Generate at random two full-rank matrices $D^{(1)}$ and $D^{(2)}$ in $\mathbb{R}^{r \times r}$. For example, use i.i.d. Gaussian distribution for the entries of $D^{(1)}$ and $D^{(2)}$.

2/ For $t = 1, 2$, define the $r$ hyperplanes $\mathbb{F}_j^{(t)} = \text{span}(\{d_i^{(t)}\}_{i \neq j})$ for $j = 1, 2, \ldots, r$.

3/ Generate at random $r - 2$ points on each intersection $\mathbb{F}_j^{(1)} \bigcap \mathbb{F}_l^{(2)}$ for $j, l \in \{1, 2, \ldots, r\}$ (there are $r^2$ such intersections), for a total of $r^2(r - 2)$ points; e.g., compute a basis of the intersection, and then use i.i.d. Gaussian distribution for the weights of the linear combinations of the $r - 2$ points. Let the columns of $M$ consist of these $r^2(r - 2)$ points.

4/ For $t = 1, 2$, compute the coefficient matrices $B^{(t)}$ such that $M = D^{(t)}B^{(t)}$.

cannot be linearly dependent (and hence have rank $r - 2$) since only subsets of $(r - 2)$ points were generated on the same $(r - 2)$-dimensional subspace.

Finally, by construction, $B^{(t)}$ $(t = 1, 2)$ exists and each column is at worst $k - 1$ sparse since every column of $M$ belongs to $\mathbb{F}_j^{(t)}$ for some $j$. In fact, it is exactly $k - 1$ sparse with probability one since the points were picked at random in the intersections and hence have nonzero coefficients in the corresponding columns of $D^{(t)}$. ∎

Lemma 2.1 implies that, in the deterministic case and for $k = r - 1$, at least $\mathcal{O}(r^3)$ points are necessary to guarantee essential uniqueness of LRSCA decompositions. Interestingly, we will prove in Theorem 3.8 that adding a single point to any subspace spanned by $r-1$ columns of $D$ will make the decomposition essentially unique. The next example illustrates this fact in dimension 3, with a hyperplane containing four points.

*Example* 2 ($4 + 3 + 2$ points lying on three hyperplanes in dimension 3). Figure 2 shows an example in dimension 3 with four points on a single hyperplane (the four aligned points). It turns out that there cannot be three hyperplanes covering these four points that do not contain the span of these four points. Hence, the hyperplane containing these four points must be identified (in the figure, the line containing the four aligned points must be identified), meaning that the span of these four points must coincide with the span of two columns of $D$ in any LRSCA decomposition of $M$; see Lemma 3.5. Using a similar argument, the line containing the three aligned points not on the first identified line must be identified (because these three points cannot be covered by two other lines). Finally, the last two points define a single line that must be identified as well. This implies that the decomposition is essentially unique since identifying all hyperplanes generated by $r - 1$ columns of $D$ makes $D$ unique up to permutation and scaling of its columns (see Corollary 3.4), while $B$ is unique since $D$ has full column rank. This will be proved rigorously and for general dimensions and sparsity in Theorem 3.8.

To summarize, in this section, we have made the following observations:

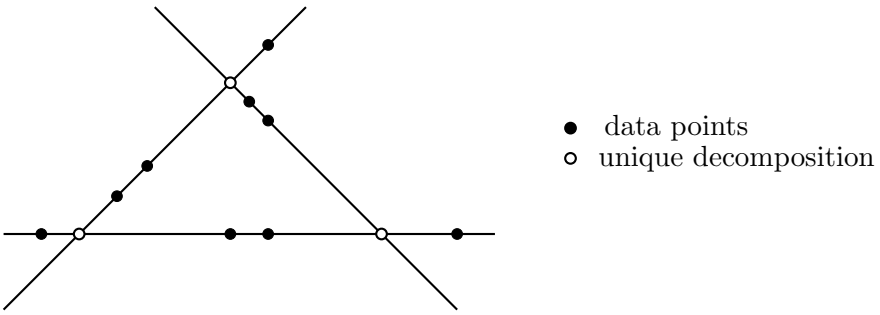


**Figure 2.** *A scenario where $M$ admits an essentially unique LRSCA (2.1) when $k = r - 1 = 2$.*

- If $M$ follows an LRSCA model with sparsity set to $r-1$ where each hyperplane contains strictly less than $r$ points, then the factorization of $M$ is never unique.
- If each of these hyperplanes contains exactly $r$ points, it may happen that $M$ does not admit an essentially unique decomposition, although this is unlikely in practice. In fact, these hyperplanes can contain up to $r(r-2)$ points, while identifiability is not guaranteed; see Algorithm 2.1 and Lemma 2.1.

## 3. Identifiability of LRSCA.

In this section, we prove our main identifiability results for LRSCA (2.1). In section 3.1, we provide some definitions and properties that will be useful for our purpose. Although these properties are well known, we provide the proofs to have the paper self-contained. In section 3.2, we prove our main theorems and discuss the tightness of our bounds on the number of data points needed to guarantee identifiability.

### 3.1. Definitions and properties.

Let us define the hyperplanes spanned by the columns of a matrix $D$ and a covering of a set of points spanned by a set of hyperplanes.

Definition 3.1. *Given $D \in \mathbb{R}^{p\times r}$ with rank $r$, we refer to the subspaces*

$$\mathbb{F}_i(D) = \operatorname{span}\left(\{d_j\}_{j\neq i}\right), \quad 1 \leq i \leq r, \tag{3.1}$$

*as the $r$ hyperplanes generated by $D$. To simplify notation and when it is clear from the context, we will drop the argument and simply write $\mathbb{F}_i = \mathbb{F}_i(D)$. Note that when $p > r$, $\mathbb{F}_i$ are not hyperplanes but rather $(r-1)$-dimensional subspaces in $\mathbb{R}^p$. However, as explained in the beginning of section 2, we could assume w.l.o.g. that $p = r$, which justifies this slight abuse of language ($\mathbb{F}_i$'s are hyperplanes within* $\operatorname{span}(D)$*).*

Definition 3.2. *A set of subspaces $\{\mathbb{F}_i\}_{i=1}^r$ forms a covering set of the data points $\{m_j\}_{j=1}^n$ if and only if for all $j = 1, 2, \ldots, n$ there exists $1 \leq i \leq r$ such that $m_j \in \mathbb{F}_i$.*

Hyperplanes generated by $D$ in LRSCA (2.1) have the following properties.

Lemma 3.3. *Let $M = DB$ follow the LRSCA model (2.1), and let $\{\mathbb{F}_j\}_{j=1}^r$ be the hyperplanes generated by $D$. Then the following hold:*

1. *The hyperplanes $\mathbb{F}_j$ are distinct $(r-1)$-dimensional subspaces.*
2. *The matrix $D$ is uniquely defined by its hyperplanes $\{\mathbb{F}_j\}_{j=1}^r$ up to scaling and permu-*

*tation:*

$$d_i \in \bigcap_{j \neq i} \mathbb{F}_j, \quad 1 \leq i \leq r, \tag{3.2}$$

*where* $\bigcap_{j\neq i} \mathbb{F}_j$ *are one-dimensional subspaces.*

3. *The set* $\{\mathbb{F}_j\}_{j\leq r}$ *is a covering of the columns of* $M$*.*

*Proof.* Claims 1 and 2 follow directly from standard linear algebra and the LRSCA model (2.1) since $\text{rank}(D) = r$. Claim 3 follows from the sparsity of $B$: since $||\mathbf{b}_i||_0 = k < r$ for all $i$, all points $m_i = Db_i$ belong to at least one hyperplane $\mathbb{F}_j$ (namely, $j$ such that $b_{i,j} = 0$). ∎

Using the lemma above, one can easily link the essential uniqueness of an LRSCA decomposition and the uniqueness of the hyperplanes generated by $D$.

Corollary 3.4. *Let* $M = DB$ *follow the LRSCA model* (2.1)*. Then the following are equivalent:*

1. $(D, B)$ *is essentially unique.*
2. *There is a unique covering set of* $M$ *involving* $r$ *subspaces of dimension* $r - 1$*.*

*Proof.* This follows directly from Lemma 3.3. In fact, given an essentially unique LRSCA decomposition, there is a unique set of hyperplanes that contains the columns of $M$. Reversely, given a unique set of hyperplanes, a unique factor $D$ is obtained. Since the data is contained in the union of these hyperplanes, the factor $B$ exists. It can be uniquely determined knowing $D$ and $M$ since $\text{rank}(D) = r$. ∎

In summary, studying the identifiability of LRSCA (2.1) is equivalent to studying the uniqueness of $r$ hyperplanes whose union contains all the data points.

**3.2. Main identifiability results.** The following key lemma provides a condition under which a hyperplane generated by $D$ spanned by the data points contained in the submatrix $M^{(1)}$ of $M$ has to be a hyperplane of any LRSCA decomposition of $M$. In other words, Lemma 3.5 focuses on the identifiability of a single hyperplane generated by $r-1$ columns of $D$. Theorem 3.6 will use this result to uniquely identify all hyperplanes generated by $D$ implying that $D$ is uniquely identified (Lemma 3.3). Theorem 3.8 refines the results of Theorem 3.6, but features more involved technical conditions, which are relaxed in Corollary 3.10.

Lemma 3.5. *Let* $M^{(1)} \in \mathbb{R}^{p\times n_1}$ *be a set of* $n_1$ *data points lying on an* $(r-1)$*-dimensional subspace, that is,* $\text{rank}(M^{(1)}) = r - 1$*, with* $\text{spark}(M^{(1)}) = r$*. Let* $D \in \mathbb{R}^{p\times r}$*,* $B^{(1)} \in \mathbb{R}^{r\times n_1}$*, and* $k < r$ *be such that* $M^{(1)} = DB^{(1)}$*,* $D$ *is full column rank, and* $\|b_j^{(1)}\|_0 \leq k$ *for all* $j$*. Then the following holds:*

$$n_1 \geq \left\lfloor \frac{r(r-2)}{r-k} \right\rfloor + 1 \quad \Rightarrow \quad \text{there exists } j \text{ such that } \mathbb{F}_j(D) = \text{span}(M^{(1)}). \tag{3.3}$$

*Proof.* Let us define $S_j = \{m_i^{(1)} \mid m_i^{(1)} \in \mathbb{F}_j\}$ as the set of columns of $M^{(1)}$ contained in the $j$th hyperplane $\mathbb{F}_j$ generated by $D$. If there exists some $j$ such that $|S_j| \geq r - 1$, then $\text{span}(S_j) = \mathbb{F}_j$ since, by assumption, $\text{spark}(M^{(1)}) = r$ and $\dim(\mathbb{F}_j) = r - 1$. Because

$\operatorname{rank}(M^{(1)}) = r-1$, this implies that a hyperplane $\mathbb{F}_j$ containing strictly more than $r-2$ data points of $M^{(1)}$ satisfies $\mathbb{F}_j = \operatorname{span}(M^{(1)})$. Moreover, every column of $M^{(1)}$ lies on at least $r-k$ hyperplanes since $\|b_i\|_0 \leq k$ for all $i$. Hence, one can check using the pigeonhole principle that the maximum number of columns that $M^{(1)}$ can contain such that each of the $r$ hyperplanes generated by $D$ contains at most $r-2$ columns of $M^{(1)}$ is given by

$$n_{\max} = \left\lfloor \frac{r(r-2)}{r-k} \right\rfloor.$$

Hence, $n_1 \geq n_{\max} + 1 > n_{\max}$ implies $\mathbb{F}_j = \operatorname{span}(M^{(1)})$ for some $j$, which completes the proof. ∎

Lemma 3.5 provides a bound on the number of points on a hyperplane that ensures that this hyperplane is contained in any LRSCA decomposition of a matrix containing these points. For instance, for $r = 3$ and $k = 2$, a hyperplane containing $r(r-2)+1 = 4$ columns of $M$ with spark 3 must be included in any LRSCA decomposition of $M$, while if it contains only three columns, it is not necessarily the case; see Examples 1 and 2. For $r = 4$ and $k = 3$, having nine columns or more of $M$ belonging to an $(r-1)$-dimensional subspace and having spark $r$ implies that the corresponding hyperplane will be contained in any 3-sparse decomposition of $M$. If such a hyperplane contains only eight columns, its identifiability is not guaranteed; see the construction in Algorithm 2.1.

We can now use Lemma 3.5 to give a sufficient condition to the uniqueness of LRSCA (2.1) by applying Lemma 3.5 to each hyperplane of a decomposition of $M$.

Theorem 3.6. *Let $M = DB$ satisfy the LRSCA model (2.1). The decomposition $(D, B)$ is essentially unique if there exists a collection of subsets $\{I_j\}_{j=1}^r$ such that $M^{(j)} = M(:, I_j)$ satisfies the following conditions: every column of $M^{(j)}$ belongs to the $j$th hyperplane generated by $D$, that is, $B(j, I_j) = 0$ for all $j$, and for all $j$,*

$$\operatorname{spark}(M^{(j)}) = r \quad \text{and} \quad |I_j| \geq \left\lfloor \frac{r(r-2)}{r-k} \right\rfloor + 1. \tag{3.4}$$

*Proof.* By assumption, $\mathbb{F}_j = \operatorname{span}(M^{(j)})$. Then uniqueness of $(B, D)$ follows directly from Corollary 3.4 (it is equivalent to identify $D$ or its hyperplanes) and Lemma 3.5, which implies that all hyperplanes generated by $D$ are identified under the condition (3.4). ∎

A simpler way to phrase Theorem 3.6 is the following: an LRSCA decomposition $M = DB$ is essentially unique if on each subspace spanned by all but one column of $D$ there are $\left\lfloor \frac{r(r-2)}{r-k} \right\rfloor + 1$ data points with spark $r$. Note that Theorem 3.6 does not require any assumption on the dictionary $D$ beyond that it has full-rank $r$.

**3.2.1. Tightness of Theorem 3.6.** Theorem 3.6 can be used to compute a minimum value of the number $n$ of columns of a matrix $M$ satisfying the assumptions of Theorem 3.6. Since each column of $M$ belongs to $r-k$ hyperplanes of $D$, it may belong to $r-k$ subsets $I_j$, and hence the condition (3.4) implies that

$$n \geq \frac{\sum_{j=1}^r |I_j|}{r-k} \geq \frac{r}{r-k} \left( \left\lfloor \frac{r(r-2)}{r-k} \right\rfloor + 1 \right).$$

For $k = r - 1$, the bound gives $n \geq r^3 - 2r^2 + r$, which is tight up to a constant $r$ since Algorithm 2.1 provides distinct LRSCA decompositions with $n = r^3 - 2r^2$ satisfying the conditions $\text{rank}(M^{(j)}) = r - 1$ and $\text{spark}(M^{(j)}) = r$ with $|I_j| = r(r-2)$. For $k = 1$, clearly $|I_j| = 1$ for all $j$ ensures uniqueness, and hence $n = r$ is enough. Our bound can be simplified as follows: since $(r-1)^2 > r(r-2)$, we have $\left\lfloor \frac{r(r-2)}{r-1} \right\rfloor = r - 2$, and hence

$$\frac{r}{r-1}\left(\left\lfloor \frac{r(r-2)}{r-1} \right\rfloor + 1\right) = \frac{r}{r-1}(r-1) = r,$$

so that the bound of Theorem 3.6 is tight.

If the columns of the matrix $B$ contain a number of zero entries proportional to $r$, that is, if the co-sparsity level satisfies $\ell = \alpha r$ for some constant $\alpha \in [1/r, (r-1)/r]$ so that $k = r(1-\alpha)$, our bound requires $n \geq \frac{r-2+\alpha}{\alpha^2}$, which is proportional to $r$ and hence is tight up to a (possibly large) constant factor $\frac{1}{\alpha^2}$ (since clearly $n \geq r$ is a necessary condition for essential uniqueness). This is rather interesting to observe since in many practical problems the sparsity is often proportional to $r$ (e.g., 10% of zero entries). This is interesting to put in perspective with the $\mathcal{O}(r \log(r))$ data points required in a probabilistic setting, when columns of $B$ are generated randomly, which can be seen as an instance of the coupon collector problem as shown by Spielman, Wang, and Wright [37].

*Remark* 3.7. Note that Theorem 3.6 can be combined with probabilistic arguments to obtain probabilistic bounds. However, such bounds would be weaker, for example, than the one of Spielman, Wang, and Wright [37] because Theorem 3.6 does use the assumption that the data points are sampled in general position. For example, let us consider the case $k = r-1$ and assume that the coefficients are sampled from the product of a uniform distribution (choose $k$ nonzero positions uniformly at random) and a Gaussian distribution (for sampling nonzeros). We need to have $m = \mathcal{O}(r^2)$ points in each subspace. The coupon collector problem requires on average $O(r \log(r))$ points to have a single point on each of the $r$ hyperplanes. Hence, to have $m$ points on each hyperplane, we need on average to sample at most $O(mr \log(r))$ points (this can be proved using the linearity of the expected value), that is, $O(r^3 \log(r))$ points, which is, up to the factor $\log(r)$, asymptotically the same as our deterministic bound. Note that slightly tighter bounds can be obtained using a more refined analysis of the generalized coupon problem that needs to collect $m$ coupons of each type, also known as the double dixie cup problem [32].

For $\alpha = 1/r$ (resp., $(r-1)/r$), that is, $\ell = 1$ and $k = r - 1$, we recover the bound $n = \Omega(r^3)$ (resp., $n \geq r$). For other values, it is more difficult to prove tightness of the bound. For example, for a number of zero entries in the columns of $B$ proportional to $\sqrt{r}$, that is, $\alpha = 1/\sqrt{r}$, we get $n = \Omega(r^2)$ and we do not know whether this bound is tight. Generalizing the construction of Algorithm 2.1 will require us to intersect $\ell$ hyperplanes among the $\mathbb{F}_j^{(1)}$'s with $\ell$ hyperplanes among the $\mathbb{F}_j^{(2)}$'s (to guarantee $B^{(1)}$ and $B^{(2)}$ to be $k$-sparse): there are many such intersections, and it is not clear how to count the points that can be generated to avoid degeneracy, that is, to guarantee that the spark of the data points on each hyperplane generated by $D$ is $r-1$ (note that $\ell < r/2$ is required for these intersections to be nonempty). Proving the tightness of the bounds in these cases is a direction of further research.

*Brute-force algorithm for LRSCA.* As a by-product of the proposed identifiability result, we also exhibit an algorithm that, under the conditions of Theorem 3.6, outputs the unique solution to the LRSCA problem in a finite number of steps; this is similar to what was proposed in [5, 19]. This algorithm iterates the following steps, until either all data points belong to an identified hyperplane (in which case the algorithm has successfully found the unique solution) or when all different possible combinations of data points are checked:

1. Choose $p = \left\lfloor \frac{r(r-2)}{r-k} \right\rfloor + 1$ data points in $M$.
2. Check whether these data points satisfy the spark condition. If not, return to step 1.
3. Check whether these data points belong to an $(r-1)$-dimensional subspace. If this is the case, then their span must be one of the hyperplanes $\mathcal{F}_j(D)$.

This provides a way, albeit computationally impractical with up to $\binom{n}{p}$ combinations of points to check, to assert the existence of a solution to LRSCA and find the unique solution under the assumptions of Theorem 3.6.

**3.2.2. Stronger identifiability result.** Theorem 3.6 is a significant improvement to already known sufficient identifiability conditions for complete dictionary learning. However, looking back on Example 2 ($r = 3$, $k = 2$), it appears that the sample complexity bound given by Theorem 3.6 may be improved. Indeed, in Example 2, we have shown that only $n = 9$ points may be sufficient for LRSCA to be identifiable: four points on a first hyperplane, three points on a second one, and two points on a third one. Conversely, Theorem 3.6 predicts identifiability when four points belong to each hyperplane. In other words, in Example 2, only $n = 9$ data points are required for identifiability, whereas Theorem 3.6 requires at least $n = 12$.

In what follows, we propose Theorem 3.8, featuring somewhat involved conditions but being weaker than Theorem 3.6. This allows us to obtain Corollary 3.10, whose conditions are as simple as that of Theorem 3.6 while reducing roughly by half the sample size required by Theorem 3.6 in the particular case $k = r - 1$. In fact, Corollary 3.10 provides a complete explanation of the observations made in Example 2.

The idea is the following: in Theorem 3.6, we have identified each hyperplane independently of the others. However, the fact that a hyperplane is identified influences the bound on the number of points needed on the other hyperplanes. Hence, using Lemma 3.5 sequentially, instead of simultaneously, for all hyperplanes, the following stronger result can be derived.

Theorem 3.8. *Let $M = DB$ satisfy the LRSCA model (2.1). The decomposition $(D, B)$ is essentially unique if there exists a collection of subsets $\{I_j\}_{j=1}^r$ such that $M^{(j)} = M(:, I_j)$ satisfies the following conditions:*

1. *Every column of $M^{(j)}$ belongs to the $j$th hyperplane $\mathbb{F}_j$ of $D$; that is, for all $j$, $B(j, I_j) = 0$.*
2. *We have for all $j$*

$$\text{spark}(M^{(j)}) = r \quad \text{and} \quad |I_j| \geq \left\lfloor \frac{(r-j+1)(r-2)+\sum_{i\in I_j} c_{i\to j}}{r-k} \right\rfloor + 1, \tag{3.5}$$

*where the quantity*

$$c_{i\to j} = |\{p \,|\, p < j, m_i \in \mathbb{F}_p\}| = |\{p \,|\, p < j, B(p, i) = 0\}|$$

*is defined for all* $1 \leq j \leq r$ *and all* $i \in I_j$. *Given a point* $m_i$ *that belongs to the hyperplane* $\mathbb{F}_j$, *the quantity* $c_{i\to j}$ *is the number of hyperplanes* $p$ *preceding* $j$ *(that is,* $p < j$*) to which* $m_i$ *belongs.*

Before we give the proof, let us try to shed some light on the conditions of Theorem 3.8. First, the order in which we sort the hyperplanes plays a crucial role, as opposed to Theorem 3.6, since $c_{i\to j}$ depends on this ordering. Second, the quantity $\sum_{i\in I_j} c_{i\to j}$ is smaller than $(j-1)(r-2)$: In fact, the intersection of two hyperplanes has dimension $r-2$, and hence one cannot pick more than $r-2$ points from hyperplane $j$ in hyperplane $k \neq j$ because of the spark condition $\text{spark}(M^{(j)}) = r$. Therefore, condition (3.5) of Theorem 3.8 is weaker than condition (3.4) of Theorem 3.6. For example, let us consider the case $k = r-1$ and assume $c_{i\to j} = 0$ for all $i, j$ (this can actually be assumed w.l.o.g.; see Lemma 3.9); hence the last requirement of the third inequality in condition (3.5) becomes $|I_j| \geq (r-j+1)(r-2)+1$. For $r = 3$ and $k = 2$, we know that having four points with spark 3 on each hyperplane is enough for identifiability (Theorem 3.6). From Theorem 3.8, we know that the following weaker conditions will be enough: (i) four points with spark 3 on a first hyperplane as in Theorem 3.6, (ii) three points with spark 3 on a second hyperplane that do not belong to the first hyperplane, and (iii) two points with spark 3 on a third hyperplane that do not belong to the first two hyperplanes. Figure 2 illustrates such a unique decomposition.

*Proof of Theorem* 3.8. Let us prove the result by induction.

First, consider the first set of indices $I_1$ satisfying (3.5). Since no hyperplane has already been identified, we have $c_{i\to 1} = 0$ for all $i \in I_1$. By applying Lemma 3.5 on $I_1$, as in Theorem 3.6, the first hyperplane $\mathbb{F}_1$ must be contained in any decomposition of $M$.

Now suppose that $j-1$ hyperplanes are correctly identified where $j \geq 2$, that is, the hyperplanes $\mathbb{F}_p$ for $p < j$ are identified (that is, they must belong to any decomposition). Assume $M$ admits (at least) two different decompositions: one corresponding to the sought hyperplanes $\mathbb{F}_p$ $(1 \leq p \leq r)$ and one with hyperplanes $\mathbb{F}'_p$ $(1 \leq p \leq r)$. Since the first $j-1$ hyperplanes are correctly identified, by induction, $\mathbb{F}'_p = \mathbb{F}_p$ for $p \leq j-1$ (w.l.o.g. we assume the first $j-1$ hyperplanes $\mathbb{F}'_p$ correspond to the first $j-1$ hyperplanes $\mathbb{F}_p$; otherwise, we reorder them accordingly). If $\mathbb{F}_j = \mathbb{F}'_k$ for some $k \geq j$, $\text{span}(M^{(j)})$ is a common hyperplane of both decompositions, and we move to the next $j$. Otherwise, $M^{(j)}$ is covered by the set of $r$ hyperplanes $\{\mathbb{F}'_l\}_{l=1}^r$ that does not contain $\text{span}(M^{(j)})$. Since $\mathbb{F}'_p = \mathbb{F}_p$ for $p < j$, the hyperplanes $\mathbb{F}'_p$ can be divided in two classes: hyperplanes that are identified $(p < j)$ and hyperplanes that are free $(p \geq j)$, that is, that are not necessarily identified.

Hyperplanes that are free may each contain only up to $r-2$ columns of $M^{(j)}$ by the same argument as in Lemma 3.5: In fact, if (say) $\mathbb{F}'_p$ for some $p \geq j$ contains $r-1$ columns of $M^{(j)}$, then because $\text{spark}(M^{(j)}) = r$, the span of these $r-1$ columns equals $\text{span}(M^{(j)})$, which contradicts our hypothesis that the hyperplane corresponding to $\text{span}(M^{(j)})$ is not identified. Therefore, the maximal number of columns of $M^{(j)}$ on all the free hyperplanes is $(r-j+1)(r-2)$.

The identified hyperplanes $\mathbb{F}_p$ for $p < j$ may also contain columns of $M^{(j)}$. By definition of $c_{i\to j}$, the number of times the identified hyperplanes touch points of $M^{(j)}$ in $I_j$ is given by

$\sum_{i\in I_j} c_{i\to j}$. Hence, the total number times all hyperplanes touch points in $I_j$ is at most

$$(r-j+1)(r-2)+\sum_{i\in I_j} c_{i\to j}.$$

Now, since the $r$ hyperplanes $\mathbb{F}'_p$ must correspond to a valid decomposition, that is, a decomposition that is $r-k$ sparse, each column of $M^{(j)}$ belongs to at least $r-k$ hyperplanes. This allows us to conclude that the total number of times the hyperplanes $\{\mathbb{F}'_l\}_{l=1}^r$ touch points in $M^{(j)}$ must be $(r-k)|I_j|$, while it is at most $(r-j+1)(r-2)+\sum_{i\in I_j} c_{i\to j}$. Therefore,

$$|I_j| > \frac{(r-j+1)(r-2)+\sum_{i\in I_j} c_{i\to j}}{r-k}$$

leads to a contradiction and the hyperplane $\mathbb{F}_j$ must be identified.

Finally, since all hyperplanes generated by $D$ have been identified, $(B,D)$ is essentially unique (Corollary 3.4), which concludes the proof. ∎

*Tightness of Theorem* 3.8. Since Theorem 3.8 is stronger than Theorem 3.6, it is also tight when $k=1$ and asymptotically tight when $k$ is a fixed proportion of $r$ (see the discussion after Theorem 3.6). However, Theorem 3.8 has the advantage of being tight for $k=r-1$ (which is not the case of Theorem 3.8, which is tight only up to a constant $r$). We will see in Lemma 3.9 that we may assume w.l.o.g. that $c_{i\to j}\le r-k-1$ and hence $c_{i\to j}=0$ for $k=r-1$. Therefore, when $k=r-1$, the condition on the cardinality of the sets $I_j$ simplifies to $|I_j|\ge (r-j+1)(r-2)+1$. Using a construction similar to that in Algorithm 2.1, we can generate matrices with two distinct LRSCA decompositions that satisfy all the conditions of Theorem 3.8 with $|I_j|=(r-j+1)(r-2)+1$ for all $j$ except for some $k$ for which $|I_k|=(r-k+1)(r-2)$. (Note that the two decompositions will have $k-1$ hyperplanes in common.)

Although Theorem 3.8 has a more relaxed condition than Theorem 3.6 on the cardinalities of the index sets $I_j$, it is more difficult to check, as we do not know a priori the values of the quantities $c_{i\to j}$ (except for $k=r-1$; see Lemma 3.9). In order to simplify this condition in Theorem 3.8, we derive an upper bound for the quantities $c_{i\to j}$. This will lead to a weaker identifiability condition than Theorem 3.8 but one that is easier to write down.

Lemma 3.9. *In Theorem 3.8, one may assume w.l.o.g. that $c_{i\to j}\le r-k-1$ for all $1\le j\le r$ and $i\in I_j$.*

*Proof.* Let $M=DB$ satisfy the conditions of Theorem 3.8, and suppose that there exist $j$ and $p\in I_j$ such that $c_{p\to j}\ge r-k$. Because of the spark condition on $M^{(j)}$, we have $|I_j|\ge r-1$. Let us consider two cases.

*Case* 1: $|I_j|\ge r$. In this case, we replace $I_j$ with $I'_j=I_j\setminus\{p\}$ and $M^{(j)}$ with $M'^{(j)}=M(:,I'_j)$ in Theorem 3.8. The two conditions

(i) $\operatorname{spark}(M'^{(j)})=r$ and

(ii) $|I'_j| > \frac{(r-j+1)(r-2)+\sum_{i\in I'_j} c_{i\to j}}{r-k}$

of Theorem 3.8 are still satisfied. In fact, since $\frac{c_{p\to j}}{r-k}\ge 1$ and by the condition on $|I_j|$, we have

$|I'_j| = |I_j| - 1$, while

$$|I_j| - 1 > \frac{(r-j+1)(r-2) + \sum_{i \in I_j} c_{i \to j}}{r-k} - 1 \geq \frac{(r-j+1)(r-2) + \sum_{i \in I'_j} c_{i \to j}}{r-k},$$

and hence $I'_j$ satisfies (ii). Then, because $M'^{(j)}$ has all columns of $M^{(j)}$ except for $m_p$, the spark of $M'^{(j)}$ can only decrease by one if $|I_j| = r-1$ (that is, if $M'^{(j)}$ has only $r-1$ columns); see the definition of the spark. Since we assume $|I_j| \geq r$ in this case, this is not possible.

*Case* 2: $|I_j| = r - 1$. In this case, we will construct $I'_j = I_j \backslash \{p\} \cup \{h\}$ by showing that there exists a point $m_h$ with $c_{h \to j} \leq r - k - 1$ and hence that two conditions (i)–(ii) are satisfied. By Theorem 3.8, the LRSCA decomposition of $M$ is unique. We have $|I_j| = r - 1$ and $c_{p \to j} \geq r - k$ for some $p$. Let us try to construct a new factorization $(D', B')$ where we replace the single hyperplane $\mathbb{F}_j = \mathbb{F}_j(D)$ with a different hyperplane $\mathbb{F}'_j = \mathbb{F}_j(D')$ defined as

$$\mathbb{F}'_j = \text{span}([M'^{(j)}, \zeta]),$$

where $M'^{(j)} = M(:, I_j \setminus \{p\})$, and $\zeta$ is any vector not in $\text{span}(M'^{(j)})$ and such that $\mathbb{F}'_j \neq \mathbb{F}_j$. By assumption, this cannot correspond to a valid decomposition $(D', B')$ of the form (2.1), and hence there exists a column of $M$, say $m_h$, such that the corresponding column of $B'$ does not contain $r - k$ zeros. Recall that $B(j, i) = 0$ if and only if the $i$th data point belongs to the $j$th hyperplane generated by $D$. We cannot have $h = p$ since $m_p$ belongs to $r-k$ hyperplanes other than $\mathbb{F}_j$ (since $c_{p \to j} \geq r - k$) which have not been modified in the new decomposition $(D', B')$; otherwise, the corresponding column of $B'$ has at least $r-k$ zeros. We must have $m_h \in \mathbb{F}_j$ since $\mathbb{F}_j$ is the only modified hyperplane in the new decomposition; otherwise, $b'_h$ contains at least $r - k$ zeros. For the same reason, we must have $m_h \notin \text{span}(M'^{(j)}) \subset \mathbb{F}'_j$. Also, $m_h$ belongs to exactly $r-k$ hyperplanes of $D$: if it belongs to $r-k+1$ hyperplanes, then it belongs to at least $r - k$ hyperplanes in the new decomposition since only one hyperplane is modified. Finally, for the decomposition to be unique, there must exist a point $m_h \in \mathbb{F}_j$ with $c_{h \to j} \leq r - k - 1$ such that $I'_j = I_j \backslash \{p\} \cup \{h\}$ satisfies the spark condition since $m_h \notin \text{span}(M'^{(j)})$. ∎

When $k = r - 1$, an interesting implication of Theorem 3.8 and Lemma 3.9 is that points lying on the $k$th hyperplane $\mathbb{F}_k$ (corresponding to the subset $I_k$) are useless to identify the hyperplanes $\mathbb{F}_j$ for $j > k$ and can therefore be discarded. Hence, this motivates and justifies the use of sequential algorithms that work as follows: until $r$ hyperplanes have not been identified, (i) identify sufficiently many points lying on a hyperplane, (ii) remove all these points, and go back to (i). Such algorithms have already been used for subspace clustering; see, for example, [42] and the references therein.

Corollary 3.10. *Let $M = DB$ satisfy (2.1). The factorization $(D, B)$ is essentially unique under the same conditions as in Theorem 3.8 except for the cardinalities of the index sets $I_j$ which are replaced with*

$$|I_j| > (r - j + 1)(r - 2), \tag{3.6}$$

*instead of the second condition in (3.5).*

*Proof.* Lemma 3.9 shows that one can assume w.l.o.g. in Theorem 3.8 that $c_{i\to j} \leq r-k-1$ for all $1 \leq j \leq r$ and $i \in I_j$. Hence, the second condition in (3.5), which is equivalent to

$$|I_j| > \frac{(r-j+1)(r-2) + \sum_{i\in I_j} c_{i\to j}}{r-k},$$

is implied by

$$|I_j| > \frac{(r-j+1)(r-2) + |I_j|(r-k-1)}{r-k} \tag{3.7}$$

since $\sum_{i\in I_j} c_{i\to j} \leq |I_j|(r-k-1)$. After simplifications, (3.7) is equivalent to

$$|I_j| > (r-j+1)(r-2),$$

which completes the proof. ∎

By Lemma 3.9, the bounds of Corollary 3.10 coincide with Theorem 3.8 when $k = r-1$ and hence is tight. For smaller $k$, Corollary 3.10 is weaker. In particular, for $k = 1$, it is much weaker since Corollary 3.10 would require $\mathcal{O}(r^2)$ data points instead of only $\mathcal{O}(r)$.

**4. Conclusion and further work.** In this paper, we have considered the identifiability of LRSCA (2.1), which is the SCA model with an (under)complete dictionary, without noise and in a deterministic setting. We provided new sufficient results (Theorems 3.6 and 3.8) guaranteeing identifiability as soon as the data points are sufficiently numerous and well-spread on the subspaces spanned by $r-1$ atoms of the dictionary. The total number of data points must be larger than $O(r^3/(r-k)^2)$, where $r$ is the number of atoms in the dictionary and $k$ is the maximum number of nonzeros in the coefficients used to reconstruct each data point using the dictionary atoms. These results improve drastically on what was known previously; see the discussion in section 2.2.1.

Further research includes the derivation of identifiability results in the presence of noise. Accounting for noise would be an important improvement to our results since it would allow us to identify the dictionary $D$ and the sparse coefficients $B$ in an approximate factorization setting, which is the model used for most applications, such as the ones described in section 2. Further research also includes the extension of our results in the case of overcomplete dictionaries. There are at least three additional issues in this scenario:

1. The uniqueness of $B$ becomes nontrivial, as the dictionary is not full rank, although sufficient conditions already exist; for example, $k < \frac{\text{spark}(D)}{2}$ [15].
2. Lemma 3.5 can be adapted but will require one to consider $\binom{r}{d-1}$ hyperplanes generated by the $r$ columns of $D$, where $d = \text{rank}(M)$. This drastically increases the number of data points needed to identify a hyperplane. Also, the case $\text{spark}(D) < d+1$ would have to be analyzed carefully since some subspaces generated by $d-1$ atoms will not be of dimension $d-1$ and will be contained in other subspaces spanned by $d-1$ atoms.
3. All hyperplanes generated by $r-1$ atoms of the dictionary need not to be identified to make the SCA essentially unique. In fact, it is sufficient to identify, for each atom, $d-1$ hyperplanes containing it. Since each hyperplane contains $d-1$ atoms, it is sufficient to identify $r$ well-chosen hyperplanes. For example, for a three-dimensional

problem with four atoms in the dictionary and $\ell = r - k = 2$, it is sufficient to identify four well-chosen hyperplanes (namely, each atom should be at the intersection of two identified hyperplanes) among the six hyperplanes to guarantee identifiability of the atoms.

It would also be interesting to study the identifiability of the LRSCA model with additional constraints. For example, LRSCA with nonnegativity on the factor $B$ is closely related to sparse nonnegative matrix factorization [23] and would lead to weaker conditions on the number of points needed on each hyperplane. It can be shown, for example, that in three dimensions and 2-sparse decompositions ($r = 3$, $k = 2$, $\ell = 1$), three points on a hyperplane are sufficient for it to be identifiable uniquely (as opposed to four data points in LRSCA; see Lemma 3.5). In fact, in the projective representation, the atoms are the vertices of a triangle whose segments contain the data points: a triangle cannot contain three aligned points unless one of its segments contains them all. Generalizing this observation in higher dimensions is a topic for further research.

Finally, given the sequential structure of our proofs, a study of partial identifiability, that is, on the uniqueness of some atoms in $D$ and the related coefficients in $B$, is an interesting direction of research. This kind of partial identifiability results could be useful for applications where only some atoms need to be identified and interpreted.

**Acknowledgments.** The authors would like to thank the editor and the reviewers for their insightful comments which helped improve the paper.

## REFERENCES

[1] R. Adamczak, *A note on the sample complexity of the Er-SpUD algorithm by Spielman, Wang and Wright for exact recovery of sparsely used dictionaries*, J. Mach. Learn. Res., 17 (2016), pp. 6153–6170.

[2] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon, *Learning sparsely used overcomplete dictionaries*, in Proceedings of the Conference on Learning Theory, 2014, pp. 123–137.

[3] A. Agarwal, A. Anandkumar, and P. Netrapalli, *A clustering approach to learning sparsely used overcomplete dictionaries*, IEEE Trans. Inform. Theory, 63 (2017), pp. 575–592.

[4] M. Aharon, M. Elad, and A. Bruckstein, *K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation*, IEEE Trans. Signal Process., 54 (2006), pp. 4311–4322.

[5] M. Aharon, M. Elad, and A. M. Bruckstein, *On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them*, Linear Algebra Appl., 416 (2006), pp. 48–67.

[6] A. Anandkumar, D. Hsu, and M. J. S. Kakade, *When are overcomplete topic models identifiable? Uniqueness of tensor Tucker decompositions with structured sparsity*, J. Mach. Learn. Res., 16 (2015), pp. 2643–2694.

[7] S. Arora, A. Bhaskara, R. Ge, and T. Ma, *More Algorithms for Provable Dictionary Learning*, preprint, https://arxiv.org/abs/1401.0579, 2014.

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

[9] S. Arora, R. Ge, and A. Moitra, *New algorithms for learning incoherent and overcomplete dictionaries*, in Proceedings of the Conference on Learning Theory, 2014, pp. 779–806.

[10] J. 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. Topics Appl. Earth Observ., 5 (2012), pp. 354–379.

[11] J. Blasiok and J. Nelson, *An improved analysis of the ER-SpUD dictionary learning algorithm*, in Proceedings of the 43rd International Colloquium on Automata, Languages, and Programming (ICALP),

Schloss Dagstuhl-Leibniz-Zentrum für Informatik, Wadern, Germany, 2016.
[12] C. F. Caiafa and A. Cichocki, *Multidimensional compressed sensing and their applications*, WIREs Data Min. Knowl., 3 (2013), pp. 355–380.
[13] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet, *A direct formulation for sparse PCA using semidefinite programming*, SIAM Rev., 49 (2007), pp. 434–448, https://doi.org/10.1137/050645506.
[14] I. Domanov and L. De Lathauwer, *On the uniqueness of the canonical polyadic decomposition of third-order tensors—Part* II: *Uniqueness of the overall decomposition*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 876–903, https://doi.org/10.1137/120877258.
[15] D. L. Donoho and M. Elad, *Optimally sparse representation in general (nonorthogonal) dictionaries via $\ell^1$ minimization*, Proc. Natl. Acad. Sci. USA, 100 (2003), pp. 2197–2202.
[16] M. Elad and M. Aharon, *Image denoising via sparse and redundant representations over learned dictionaries*, IEEE Trans. Image Process., 15 (2006), pp. 3736–3745.
[17] E. Elhamifar and R. Vidal, *Sparse subspace clustering: Algorithm, theory, and applications*, IEEE Trans. Pattern Anal. Mach. Intell., 35 (2013), pp. 2765–2781.
[18] X. Fu, K. Huang, N. D. Sidiropoulos, and W.-K. Ma, *Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications*, IEEE Signal Process. Mag., 36 (2019), pp. 59–80.
[19] P. Georgiev, F. Theis, and A. Cichocki, *Sparse component analysis and blind source separation of underdetermined mixtures*, IEEE Trans. Neural Netw., 16 (2005), pp. 992–996.
[20] R. Gribonval, R. Jenatton, and F. Bach, *Sparse and spurious: Dictionary learning with noise and outliers*, IEEE Trans. Inform. Theory, 61 (2015), pp. 6298–6319.
[21] R. Gribonval and M. Zibulevsky, *Sparse component analysis*, in Handbook of Blind Source Separation, Elsevier, Amsterdam, Boston, 2010, pp. 367–420.
[22] C. J. Hillar and F. T. Sommer, *When can dictionary learning uniquely recover sparse data from subsamples?*, IEEE Trans. Inform. Theory, 61 (2015), pp. 6290–6297.
[23] P. O. Hoyer, *Non-negative matrix factorization with sparseness constraints*, J. Mach. Learn. Res., 5 (2004), pp. 1457–1469.
[24] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, *Generalized power method for sparse principal component analysis*, J. Mach. Learn. Res., 11 (2010), pp. 517–553.
[25] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM Rev., 51 (2009), pp. 455–500, https://doi.org/10.1137/07070111X.
[26] D. D. Lee and H. S. Seung, *Learning the parts of objects by non-negative matrix factorization*, Nature, 401 (1999), pp. 788–791.
[27] L.-H. Lim and P. Comon, *Multiarray signal processing: Tensor decomposition meets compressed sensing*, C. R. Mécanique, 338 (2010), pp. 311–320.
[28] J. Mairal, F. Bach, and J. Ponce, *Task-driven dictionary learning*, IEEE Trans. Pattern Anal. Mach. Intell., 34 (2012), pp. 791–804.
[29] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, *Online dictionary learning for sparse coding*, in Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 689–696.
[30] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach, *Supervised dictionary learning*, in Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2009, pp. 1033–1040.
[31] F. M. Naini, G. H. Mohimani, M. Babaie-Zadeh, and C. Jutten, *Estimating the mixing matrix in Sparse Component Analysis (SCA) based on partial k-dimensional subspace clustering*, Neurocomputing, 71 (2008), pp. 2330–2343.
[32] D. J. Newman, *The double dixie cup problem*, Amer. Math. Monthly, 67 (1960), pp. 58–61.
[33] B. A. Olshausen and D. J. Field, *Sparse coding with an overcomplete basis set: A strategy employed by V1?*, Vis. Res., 37 (1997), pp. 3311–3325.
[34] A. Ozerov and C. Févotte, *Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation*, IEEE Trans. Speech Audio Process., 18 (2010), pp. 550–563.
[35] R. Rubinstein, A. M. Bruckstein, and M. Elad, *Dictionaries for sparse representation modeling*, Proc. IEEE, 98 (2010), pp. 1045–1057.
[36] H. Shen and J. Z. Huang, *Sparse principal component analysis via regularized low rank matrix approx-*

*imation*, J. Multivariate Anal., 99 (2008), pp. 1015–1034.
[37] D. A. Spielman, H. Wang, and J. Wright, *Exact recovery of sparsely-used dictionaries*, in Proceedings of the Conference on Learning Theory, 2012, pp. 37.1–37.18.
[38] J. Sun, Q. Qu, and J. Wright, *Complete dictionary recovery over the sphere* I: *Overview and the geometric picture*, IEEE Trans. Inform. Theory, 63 (2017), pp. 853–884.
[39] J. Sun, Q. Qu, and J. Wright, *Complete dictionary recovery over the sphere* II: *Recovery by Riemannian trust-region method*, IEEE Trans. Inform. Theory, 63 (2017), pp. 885–914.
[40] I. Tosic and P. Frossard, *Dictionary learning*, IEEE Signal Process. Mag., 28 (2011), pp. 27–38.
[41] M. C. Tsakiris and R. Vidal, *Hyperplane clustering via dual principal component pursuit*, in Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 3472–3481.
[42] R. Vidal, *Subspace clustering*, IEEE Signal Process. Mag., 28 (2011), pp. 52–68.
[43] H. Zou, T. Hastie, and R. Tibshirani, *Sparse principal component analysis*, J. Comput. Graph. Statist., 15 (2006), pp. 265–286.