GLOBAL IDENTIFIABILITY OF OVERCOMPLETE DICTIONARY LEARNING VIA L1 AND VOLUME MINIMIZATION

Yuchen Sun & Kejun Huang
Department of Computer and Information Science and Engineering
University of Florida
Gainesville, FL 32611, USA
{yuchen.sun,kejun.huang}@ufl.edu

ABSTRACT

We propose a novel formulation for dictionary learning with an overcomplete dictionary, i.e., when the number of atoms is larger than the dimension of the dictionary. The proposed formulation consists of a weighted sum of ℓ_1 norms of the rows of the sparse coefficient matrix plus the log of the matrix volume of the dictionary matrix. The main contribution of this work is to show that this novel formulation guarantees global identifiability of the overcomplete dictionary, under a mild condition that the sparse coefficient matrix satisfies a strong scattering condition in the hypercube. Furthermore, if every column of the coefficient matrix is sparse and the dictionary guarantees ℓ_1 recovery, then the coefficient matrix is identifiable as well. This is a major breakthrough for not only dictionary learning but also general matrix factorization models as identifiability is guaranteed even when the latent dimension is higher than the ambient dimension. We also provide a probabilistic analysis and show that if the sparse coefficient matrix is generated from the widely adopted sparse-Gaussian model, then the $m \times k$ overcomplete dictionary is globally identifiable if the sample size is bigger than a constant times $(k^2/m)\log(k^2/m)$ with overwhelming probability. Finally, we propose an algorithm based on alternating minimization to solve the new proposed formulation.

1 INTRODUCTION

Dictionary learning (DL) amounts to factor a data matrix as $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S}$ where $\boldsymbol{S}$ is sparse (Tošić & Frossard, 2011), which may also be known as sparse coding (Olshausen & Field, 1997) or sparse component analysis (Georgiev et al., 2005) in various fields. Treating $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ or $\mathbb{C}^{m\times n}$ as a collection of data samples as its columns, this factorization means that each sample is a *sparse* combination of the columns of $\boldsymbol{A}$, or in other words atoms of the dictionary. Unlike the task of compressive sensing or sparse vector recovery, in which case the dictionary matrix $\boldsymbol{A}$ is given, dictionary learning tries to find both $\boldsymbol{A}$ and $\boldsymbol{S}$, therefore the problem is a lot more challenging. Depending on the shape of the dictionary matrix $\boldsymbol{A}$, we may seek to find a complete dictionary if $\boldsymbol{A}$ is square or an overcomplete dictionary if $\boldsymbol{A}$ is wide. In this paper we focus on overcomplete dictionary learning, therefore $\boldsymbol{A} \in \mathbb{R}^{m\times k}$ and $\boldsymbol{S} \in \mathbb{R}^{k\times n}$ (or $\mathbb{C}^{m\times k}$ and $\mathbb{C}^{k\times n}$, respectively) where $m < k$.

Dictionary learning has found numerous applications in signal denoising (Elad & Aharon, 2006), audio coding (Plumbley et al., 2009), and medical imaging (Tošić et al., 2010), to name just a few. On the theory side, most of the existing works have focused on algorithm design. Famous algorithms include k-SVD (Aharon et al., 2006a) and online dictionary learning (Mairal et al., 2009), among numerous other algorithms based on generic nonconvex algorithm design with guarantee of convergence to a stationary point. More recently, there has appeared a line of research that attempts to show global optimality for dictionary learning under more restrictive assumptions, such as (Spielman et al., 2012; Agarwal et al., 2016; Arora et al., 2014; 2015; Sun et al., 2016a;b; Rambhatla et al., 2019; Bai et al., 2019; Zhai et al., 2020a;b; Shen et al., 2020; Tolooshams & Ba, 2022).

1.1 PRIOR WORK ON IDENTIFIABILITY OF DL

A matrix factorization model without any additional assumptions on the latent factors is known to be not unique, since we can always "insert" an invertible matrix $\boldsymbol{W}$ and $\boldsymbol{W}^{-1}$ as $\boldsymbol{X} = \widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{S}}$ where $\widetilde{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{W}^{-1}$ and $\widetilde{\boldsymbol{S}} = \boldsymbol{W}\boldsymbol{S}$, and one cannot distinguish whether $\boldsymbol{S}$ or $\widetilde{\boldsymbol{S}}$ are the groundtruth sources. Furthermore, if the dictionary $\boldsymbol{A}$ is overcomplete with $m < k$, columns of $\widetilde{\boldsymbol{S}}$ could include additional components that are orthogonal to the row space of $\boldsymbol{A}$, i.e., $\widetilde{\boldsymbol{S}} = \boldsymbol{W}(\boldsymbol{S} + \boldsymbol{B})$ where $\boldsymbol{A}\boldsymbol{B} = 0$, while we still have $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S} = \boldsymbol{A}\boldsymbol{W}^{-1}\boldsymbol{W}(\boldsymbol{S} + \boldsymbol{B}) = \widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{S}}$. If, however, a learning criterion $q(\boldsymbol{A}, \boldsymbol{S})$ is imposed so that the resulting ambiguities can only be permutations and scaling, then we say the model is identifiable, as is formalized as follows. Notice that we differentiate the identifiability of just the dictionary $\boldsymbol{A}$ and the whole factorization model, which is indeed not equivalent when the dictionary is overcomplete.

Definition 1 (Identifiability)**.** Consider the generative model $\boldsymbol{X} = \boldsymbol{A}_\natural\boldsymbol{S}_\natural$, where $\boldsymbol{A}_\natural$ and $\boldsymbol{S}_\natural$ are the groundtruth latent factors. Let $(\boldsymbol{A}_\star, \boldsymbol{S}_\star)$ be optimal for an identification criterion q

$$(\boldsymbol{A}_\star, \boldsymbol{S}_\star) = \operatorname*{arg\,min}_{\boldsymbol{X}=\boldsymbol{A}\boldsymbol{S}} q(\boldsymbol{A}, \boldsymbol{S}).$$

If $\boldsymbol{A}_\natural$ and/or $\boldsymbol{S}_\natural$ satisfy some condition such that for any $(\boldsymbol{A}_\star, \boldsymbol{S}_\star)$, there exist a permutation matrix $\boldsymbol{\Pi}$ and a diagonal matrix $\boldsymbol{D}$ such that $\boldsymbol{A}_\natural = \boldsymbol{A}_\star\boldsymbol{D}\boldsymbol{\Pi}$, then we say $\boldsymbol{A}_\natural$ is essentially identifiable, up to permutation and scaling, under that condition; if we further have that $\boldsymbol{S}_\natural = \boldsymbol{\Pi}^\top\boldsymbol{D}^{-1}\boldsymbol{S}_\star$, then we say that the matrix factorization model is essentially identifiable, up to permutation and scaling, under that condition.

When dictionary learning was first proposed, a common learning criterion $q(\boldsymbol{A}, \boldsymbol{S})$ is simply the total number of nonzeros in $\boldsymbol{S}$, sometimes also called the ℓ_0 (pseudo-)norm $\|\boldsymbol{S}\|_0$. If every column of $\boldsymbol{S}_\natural$ is s-sparse, then via some combinatorial calculation, it has been shown that the ℓ_0-norm minimization criterion guarantees identifiability if the spark of $\boldsymbol{A}$ is at least $2s$ and the sample size n is $O((s+1)\binom{k}{s})$ (Aharon et al., 2006b; Hillar & Sommer, 2015; Garfinkle & Hillar, 2019). The main drawback is that the required sample size n is usually too large to be practical. Cohen & Gillis (2019) reduced the sample complexity down to $O(k^3/(k-s)^2)$, but also restricted the dictionary to be complete, i.e., $m \geq k$.

Another famous learning criterion for DL, inspired by the success of compressive sensing (Donoho, 2006; Candès & Wakin, 2008), is the following formulation with $q(\boldsymbol{A}, \boldsymbol{S})$ being the summation of the absolute values of $\boldsymbol{S}$ plus indicator functions that columns of $\boldsymbol{A}$ have bounded ℓ_2 norms:

$$\operatorname*{minimize}_{\boldsymbol{A},\boldsymbol{S}} \ \|\boldsymbol{S}\|_1 \qquad \text{subject to} \ \ \boldsymbol{X} = \boldsymbol{A}\boldsymbol{S}, \|\boldsymbol{A}\boldsymbol{e}_c\|_2 \leq 1, c = 1, \ldots, k. \tag{1}$$

Identifiability results based on the ℓ_1 norm formulation have been predominantly local, meaning the model is identifiable within a neighborhood of the groundtruth factors $(\boldsymbol{A}^\natural, \boldsymbol{S}^\natural)$, while the dictionary is restricted to be complete (and incoherent) (Gribonval & Schnass, 2010; Wu & Yu, 2017; Wang et al., 2020), with the sole exception of (Geng & Wright, 2014) for overcomplete DL. The advantage is that the sample size requirement is typically down to $O(k \log k)$ and allows the existence of dense outliers. Global identifiability is achieved by Hu & Huang (2023a); Sun & Huang (2024) by using a matrix volume criterion $|\det \boldsymbol{A}|$ while constraining the ℓ_1 norms of the rows of $\boldsymbol{S}$ with the same sample complexity, although as the criterion suggests it only applies to complete dictionaries.

1.2 THIS PAPER

In this paper, we propose the following novel formulation for overcomplete dictionary learning, and show that global identifiability can be achieved under mild conditions:

$$\operatorname*{minimize}_{\boldsymbol{A},\boldsymbol{S}} \ \frac{1}{2}\log\det\boldsymbol{A}\boldsymbol{A}^\top + \max_{\|\boldsymbol{d}\|_2^2=m} \sum_{c=1}^{k} d_c\|\boldsymbol{e}_c^\top\boldsymbol{S}\|_1 \qquad \text{subject to} \ \ \boldsymbol{X} = \boldsymbol{A}\boldsymbol{S} \tag{2}$$

This means the learning criterion $q(\boldsymbol{A}, \boldsymbol{S})$ consists of two parts: a weighted sum of the ℓ_1 norms of the rows of $\boldsymbol{S}$ and a term that is proportional to the "volume" of the dictionary matrix $\boldsymbol{A}$—for an overcomplete dictionary, the volume is defined as $\det\boldsymbol{A}\boldsymbol{A}^\top$ (Ben-Israel, 1992). Our contributions are as follows:

1. We give a deterministic characterization of global identifiability of overcomplete dictionary learning via solving (2). Our analysis shows that a sufficient condition is that 1) every column of $\boldsymbol{S}_\natural$ is at most s-sparse and $\boldsymbol{A}$ is a dictionary that guarantees exact recovery of all s-sparse vectors via ℓ_1 minimization, and 2) the cellular hull of $\boldsymbol{S}_\natural$ is m-sufficiently scattered in the k-hypercube $[-1,1]^k$. The resulting identifiability condition is almost minimal: the first condition is obviously necessary as otherwise $\boldsymbol{S}_\natural$ would not be identifiable even if the overcomplete $\boldsymbol{A}_\natural$ is correctly recovered; the second condition is a slightly stronger condition than that of complete dictionary learning. It is appealing to see that no other conditions are needed to guarantee global identifiability.
2. We further provide a probabilistic characterization of when a randomly generated factor matrix satisfies the aforementioned identifiability conditions. Since we only require $\boldsymbol{A}_\natural$ to guarantee exact recovery of s-sparse vectors via ℓ_1 minimization, for which there exist numerous work on this topic (such as when an i.i.d. Gaussian matrix satisfies the restricted isometry property with high probability), we will be focusing on studying the sample complexity of $\boldsymbol{S}$. We adopt the sparse-Gaussian model, i.e., every column contains at most s nonzero values that are drawn from i.i.d. standard normal and show that the resulting $\boldsymbol{S}$ satisfies the m-strongly scattered in the k-hypercube $[-1,1]^k$ with overwhelming probability if $k < m$ and n is $O((k^2/m)\log(k^2/m))$. Notice that it is again a sharp generalization of complete dictionary learning with sample complexity $O(k\log(k))$ (Hu & Huang, 2023a), and a factor of k better than the works that focus on global optimality guarantees of overcomplete DL (Agarwal et al., 2016; Rambhatla et al., 2019).
3. We propose an alternating minimization algorithm for the novel identification criterion of overcomplete DL. The formulation is first modified slightly by moving the exact factorization constraint as a data fidelity term in the objective function, then the overcomplete dictionary and the sparse coefficient matrices are updated alternatingly via a gradient-type step. As the problem is NP-hard, no known algorithm is able to guarantee convergence to a global optimum. The proposed algorithm is applied to synthetically generated data to demonstrate that global identifiability can indeed be guaranteed via solving (2).

2 Identifiability Analysis

In this section, we provide analysis on when solving (2) guarantees the exact recovery of the overcomplete dictionary $\boldsymbol{A}_\natural$ and/or the sparse coefficient matrix $\boldsymbol{S}_\natural$. In Definition 1 we mentioned that it is acceptable to recover $\boldsymbol{A}_\natural$ up to column permutation and scaling. While column permutation does not affect the objective value of (2), column scaling does. Therefore, we first study the optimal scaling that is induced from solving (2), which provides important insights into the subsequent analysis of identifiability. We provide both a deterministic condition and a probabilistic generative model that guarantees identifiability with overwhelming probability.

2.1 Optimal scaling

Lemma 1. *Let* $(\boldsymbol{A}_\star, \boldsymbol{S}_\star)$ *be an optimal solution of* (2)*, then*

$$\|\boldsymbol{e}_c^\top \boldsymbol{S}_\star\|_1 = \sqrt{\left[\boldsymbol{A}_\star^\top \left(\boldsymbol{A}_\star \boldsymbol{A}_\star^\top\right)^{-1} \boldsymbol{A}_\star\right]_{cc}} = d_{\star c}, \qquad c = 1,\ldots,k. \tag{3}$$

where $d_{\star c}$ *are the optimal weights that reach the maximum of* $\sum_c d_c \|\boldsymbol{e}_c^\top \boldsymbol{S}_\star\|_1$.

Proof. If $(\boldsymbol{A}_\star, \boldsymbol{S}_\star)$ is feasible for (2), then so is $(\boldsymbol{A}_\star \boldsymbol{\Psi}, \boldsymbol{\Psi}^{-1}\boldsymbol{S}_\star)$ where $\boldsymbol{\Psi}$ is a diagonal matrix with cth diagonal entry ψ_c. Plugging $(\boldsymbol{A}_\star \boldsymbol{\Psi}, \boldsymbol{\Psi}^{-1}\boldsymbol{S}_\star)$ into the objective of (2) and optimize with respect to $\boldsymbol{\Psi}$ while fixing $(\boldsymbol{A}_\star, \boldsymbol{S}_\star)$, then $\boldsymbol{\Psi} = \boldsymbol{I}$ should be an optimal solution. Taking the derivative with respect to ψ_c and setting it equal to zero, we get

$$\boldsymbol{a}_c^\top \left(\boldsymbol{A}_\star \boldsymbol{\Psi}^2 \boldsymbol{A}_\star^\top\right)^{-1} \boldsymbol{a}_c \psi_c - d_c \|\boldsymbol{e}_c^\top \boldsymbol{S}_\star\|_1 / \psi_c^2 = 0,$$

where $\boldsymbol{a}_c^\top$ is the cth row of $\boldsymbol{A}_\star$. If $\psi_c = 1$ is optimal, then we must have

$$d_c \|\boldsymbol{e}_c^\top \boldsymbol{S}_\star\|_1 = \left[\boldsymbol{A}_\star^\top \left(\boldsymbol{A}_\star \boldsymbol{A}_\star^\top\right)^{-1} \boldsymbol{A}_\star\right]_{cc}. \tag{4}$$

To maximize $\sum_c d_c \|\boldsymbol{e}_c^\top \boldsymbol{S}_\star\|_1$ subject to $\|\boldsymbol{d}\|_2^2 = m$, we know from the Cauchy-Schwarz inequality that $d_{\star c}$ should be chosen as some scalar α times $\|\boldsymbol{e}_c^\top \boldsymbol{S}_\star\|_1$ such that

$$\|\boldsymbol{d}_\star\|_2^2 = \sum_{c=1}^{k} \alpha^2 \|\boldsymbol{e}_c^\top \boldsymbol{S}_\star\|_1^2 = m.$$

Plugging $d_{\star c} = \alpha \|\boldsymbol{e}_c^\top \boldsymbol{S}_\star\|_1$ into (4) and sum over $c = 1, \ldots, k$ shows

$$\sum_{c=1}^{k} \alpha \|\boldsymbol{e}_c^\top \boldsymbol{S}_\star\|_1^2 = \sum_{c=1}^{k} \left[\boldsymbol{A}_\star^\top \left(\boldsymbol{A}_\star \boldsymbol{A}_\star^\top\right)^{-1} \boldsymbol{A}_\star\right]_{cc} = \operatorname{Tr} \boldsymbol{A}_\star^\top \left(\boldsymbol{A}_\star \boldsymbol{A}_\star^\top\right)^{-1} \boldsymbol{A}_\star = \operatorname{Tr} \left(\boldsymbol{A}_\star \boldsymbol{A}_\star^\top\right)^{-1} \boldsymbol{A}_\star \boldsymbol{A}_\star^\top = m.$$

This means $\alpha = 1$, and therefore (3) holds. □

2.2 Identifiability analysis

Assumption 1. The columns of $\boldsymbol{A}_\natural$ and rows of $\boldsymbol{S}_\natural$ are scaled and counter-scaled to satisfy:

$$\|\boldsymbol{e}_c^\top \boldsymbol{S}_\natural\|_1 = \sqrt{\left[\boldsymbol{A}_\natural^\top \left(\boldsymbol{A}_\natural \boldsymbol{A}_\natural^\top\right)^{-1} \boldsymbol{A}_\natural\right]_{cc}}, \qquad c = 1, \ldots, k. \tag{5}$$

Assumption 2. Rows of $\boldsymbol{A}_\natural$ and $\boldsymbol{S}_\natural$ are both linearly independent. Matrix $\boldsymbol{A}_\natural$ does not contain zero columns.

Assumption 3 (m-strongly scattered in the k-hypercube)**.** Let C_k denote the k-hypercube $C_k = \{\boldsymbol{x} \in \mathbb{R}^k \mid \|\boldsymbol{x}\|_\infty \le 1\}$. Define $\mathcal{B}_m$ as the following set

$$\mathcal{B}_m = \left\{\operatorname{Diag}(\|\boldsymbol{q}_1\|_2, \ldots, \|\boldsymbol{q}_k\|_2)^\dagger \boldsymbol{Q}\boldsymbol{p} \mid \forall\, \boldsymbol{Q} \in \mathbb{R}^{k\times m} : \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I},\ \boldsymbol{p} \in \mathbb{R}^m : \|\boldsymbol{p}\|_2 = 1\right\},$$

where $\boldsymbol{q}_c$ denotes the cth row of $\boldsymbol{Q}$. A set $\mathcal{S} \in \mathbb{R}^k$ is m-strongly scattered in the k-hypercube if:

1. $\mathcal{B}_m \subseteq \mathcal{S} \subseteq C_k$;
2. $\partial\mathcal{B}_m \cap \partial\mathcal{S} = \{\operatorname{Diag}(\|\boldsymbol{q}_1\|_2, \ldots, \|\boldsymbol{q}_k\|_2)^\dagger \boldsymbol{Q}\boldsymbol{q}/\|\boldsymbol{q}\|_2 \mid \boldsymbol{q} \text{ are rows of } \boldsymbol{Q} \text{ with } \boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}\}$, where ∂ denotes the boundary of the set.

The definition of the set $\mathcal{B}_m$ involves all $k \times m$ matrices with orthonormal columns $\boldsymbol{Q}$ and normalized m-vectors $\boldsymbol{p}$. To see that $\mathcal{B}_m$ is indeed a subset of C, notice that the rows of $\boldsymbol{\Psi}\boldsymbol{Q}$ are either zero or unit norms by construction, so all elements of the vector $\boldsymbol{\Psi}\boldsymbol{Q}\boldsymbol{p}$ are in $[-1, 1]$ using Cauchy-Schwarz inequality. Assumption 3 is equivalent to the sufficiently scattered condition proposed by Hu & Huang (2023a) when $m = k$, and more restrictive when $m < k$. As we will see, such restriction will be useful to establish identifiability for overcomplete DL. The sufficiently scattered condition has many variations for various identifiable unsupervised learning models, such as nonnegative matrix factorization (Huang et al., 2013; Fu et al., 2015; Huang et al., 2016; 2018), simplicial representation learning (Fu et al., 2015; Lin et al., 2015; Huang & Fu, 2019), and bounded component analysis (Tatli & Erdogan, 2021; Hu & Huang, 2023b; 2024). However, to the best of our knowledge, this is the first variation that is capable of guaranteeing identifiability when the latent dimension is *higher* than the ambient dimension. An illustration of 2-strongly scattered in the 3-hypercube is shown on the right of Figure 1, compared with 3-strongly scattered on the left, which is equivalent to the sufficiently scattered condition presented by Hu & Huang (2023a).

Assumption 3 will be imposed on the cellular hull of $\boldsymbol{S}_\natural$, which is defined as follows:

Definition 2 (Cellular hull)**.** The cellular hull of a finite set of vectors $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$, stacked as the columns of the matrix $\boldsymbol{S}$, is

$$\operatorname{cell}(\boldsymbol{S}) = \left\{\boldsymbol{S}\boldsymbol{\theta} \;\middle|\; \|\boldsymbol{\theta}\|_\infty \le 1\right\}.$$

Consider the groundtruth sparse coefficient matrix $\boldsymbol{S}_\natural$, if we rescale its rows to have unit ℓ_1 norms, denoted as $\widetilde{\boldsymbol{S}}_\natural$, then $\operatorname{cell}(\widetilde{\boldsymbol{S}}_\natural) \subseteq C_k$ due to Hölder's inequality $|\boldsymbol{a}^\top \boldsymbol{b}| \le \|\boldsymbol{a}\|_1 \|\boldsymbol{b}\|_\infty$. For identifiability of overcomplete DL, we would require $\operatorname{cell}(\widetilde{\boldsymbol{S}}_\natural)$ to be m-strongly scattered in the k-hypercube, as formally stated as follows:

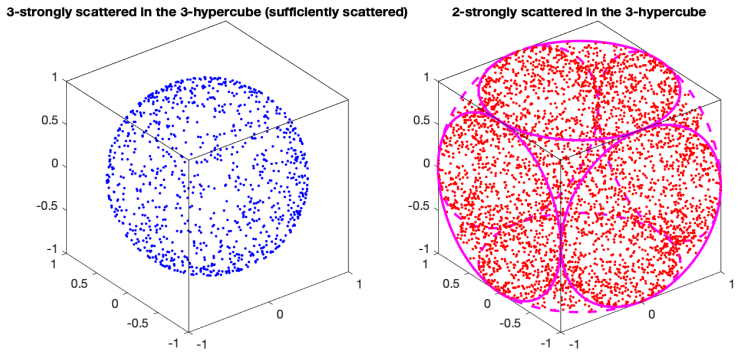

Figure 1: An illustration of $\mathcal{B}_2$ in $\mathbb{R}^3$ on the right. In comparison, $\mathcal{B}_3$ in $\mathbb{R}^3$ is the Euclidean ball illustrated on the left. While $\mathcal{B}_3$ touches the boundary of $[-1,1]^3$ at only 6 points $\pm\boldsymbol{e}_1, \pm\boldsymbol{e}_2$, and $\pm\boldsymbol{e}_3$, $\mathcal{B}_2$ touches each face of $[-1,1]^3$ at a circle with radius 1, as shown in magenta on the right. In the context of dictionary learning, a 3×3 complete dictionary is identifiable if $\mathrm{cell}(\boldsymbol{S}_\natural) \supseteq \mathcal{B}_3$, shown on the left, while a 2×3 overcomplete dictionary is identifiable if $\mathrm{cell}(\boldsymbol{S}_\natural) \supseteq \mathcal{B}_2$, shown on the right.

Theorem 1. *Consider the overcomplete DL model $\boldsymbol{X} = \boldsymbol{A}_\natural \boldsymbol{S}_\natural$, where $\boldsymbol{A}_\natural \in \mathbb{R}^{m\times k}$ is the groundtruth mixing matrix and $\boldsymbol{S}_\natural \in \mathbb{R}^{k\times n}$ is the groundtruth sparse coefficient matrix. Suppose $\boldsymbol{A}_\natural$ and $\boldsymbol{S}_\natural$ satisfies Assumptions 1 and 2. Furthermore, let $\widetilde{\boldsymbol{S}}_\natural$ denote the matrix obtained from rescaling the rows of $\boldsymbol{S}_\natural$ to have unit ℓ_1 norms, and assume that $\mathrm{cell}(\widetilde{\boldsymbol{S}}_\natural)$ is m-strongly scattered in the k-hypercube. Then for any solution of (2), denoted as $(\boldsymbol{A}_\star, \boldsymbol{S}_\star)$, there exist a permutation matrix $\boldsymbol{\Pi}$ and a diagonal matrix $\boldsymbol{D}$ such that $\boldsymbol{A}_\natural = \boldsymbol{A}_\star \boldsymbol{D}\boldsymbol{\Pi}$. In other words, an overcomplete dictionary $\boldsymbol{A}_\natural$ is identifiable if the groundtruth $\boldsymbol{A}_\natural$ and $\boldsymbol{S}_\natural$ satisfies Assumptions 1, 2, and $\mathrm{cell}(\widetilde{\boldsymbol{S}}_\natural)$ satisfies Assumption 3.*

Proof sketch. Assumption 2 asserts that rows of $\boldsymbol{A}_\natural$ are linearly independent, so there exists a $k\times m$ matrix $\boldsymbol{Q}$ with orthonormal columns that spans the row space of $\boldsymbol{A}_\natural$, then $\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}$ and

$$\boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{A}_\natural^\top \left(\boldsymbol{A}_\natural \boldsymbol{A}_\natural^\top\right)^{-1} \boldsymbol{A}_\natural.$$

Using the diagonal matrix $\boldsymbol{D}_\natural$, which is defined as

$$[\boldsymbol{D}_\natural]_{cc} = \sqrt{\left[\boldsymbol{A}_\natural^\top \left(\boldsymbol{A}_\natural \boldsymbol{A}_\natural^\top\right)^{-1} \boldsymbol{A}_\natural\right]_{cc}}, \qquad c = 1,\dots,k, \tag{6}$$

and $\boldsymbol{q}_c^\top$ as rows of $\boldsymbol{Q}$ defined in Assumption 3, we have

$$[\boldsymbol{D}_\natural]_{cc} = \|\boldsymbol{e}_c^\top \boldsymbol{S}_\natural\|_1 = \sqrt{[\boldsymbol{Q}\boldsymbol{Q}^\top]_{cc}} = \|\boldsymbol{q}_c\|_2. \tag{7}$$

For two wide $k\times n$ matrices $\boldsymbol{S}_\natural$ and $\boldsymbol{S}_\star$, there exist a $k\times k$ matrix $\boldsymbol{W}$ and $k\times n$ matrix $\boldsymbol{B}$ such that

$$\boldsymbol{S}_\star = \boldsymbol{W}\boldsymbol{S}_\natural + \boldsymbol{B},$$

where $\boldsymbol{W} = \boldsymbol{S}_\natural^\dagger \boldsymbol{S}_\star$ and rows of $\boldsymbol{B}$ are in the null space of $\boldsymbol{S}_\natural$, i.e., $\boldsymbol{S}_\natural \boldsymbol{B}^\top = 0$. Denote $\boldsymbol{w}_c^\top$ and $\boldsymbol{b}_c^\top$ as the cth row of $\boldsymbol{W}$ and $\boldsymbol{B}$, respectively, then

$$\|\boldsymbol{e}_c^\top \boldsymbol{S}_\star\|_1 = \|\boldsymbol{w}_c^\top \boldsymbol{S}_\natural + \boldsymbol{b}_c^\top\|_1 \geq \boldsymbol{w}_c^\top \boldsymbol{S}_\natural \boldsymbol{\theta} + \boldsymbol{b}_c^\top \boldsymbol{\theta} = \boldsymbol{w}_c^\top \boldsymbol{D}_\natural \widetilde{\boldsymbol{S}}_\natural \boldsymbol{\theta} + \boldsymbol{b}_c^\top \boldsymbol{\theta}, \ \ \forall \|\boldsymbol{\theta}\|_\infty \leq 1.$$

If $\mathrm{cell}(\widetilde{\boldsymbol{S}}_\natural)$ is m-strongly scattered in the k-hypercube, then for all $\|\boldsymbol{p}\|_2 = 1$ we can find a $\|\boldsymbol{\theta}\|_\infty$ such that $\boldsymbol{D}_\natural^{-1}\boldsymbol{Q}\boldsymbol{p} = \widetilde{\boldsymbol{S}}_\natural \boldsymbol{\theta}$. In Appendix A we show that $\|\boldsymbol{e}_c^\top \boldsymbol{S}_\star\|_1 = \|\boldsymbol{w}_c^\top \boldsymbol{S}_\natural\|_1$, as a result,

$$\|\boldsymbol{e}_c^\top \boldsymbol{S}_\star\|_1 = \|\boldsymbol{w}_c^\top \boldsymbol{S}_\natural\|_1 = \|\boldsymbol{w}_c^\top \boldsymbol{D}_\natural \widetilde{\boldsymbol{S}}_\natural\|_1 \geq \boldsymbol{w}_c^\top \boldsymbol{D}_\natural \boldsymbol{D}_\natural^{-1} \boldsymbol{Q}\boldsymbol{p} = \boldsymbol{w}_c^\top \boldsymbol{Q}\boldsymbol{p}, \ \ \forall \|\boldsymbol{p}\|_2 = 1.$$

Choose $\boldsymbol{p} = \boldsymbol{Q}^\top \boldsymbol{w}_c / \|\boldsymbol{Q}^\top \boldsymbol{w}_c\|_2$, we have $\|\boldsymbol{e}_c^\top \boldsymbol{S}_\star\|_1 \geq \|\boldsymbol{Q}^\top \boldsymbol{w}_c\|_2$. Square both sides and sum over $c = 1,\dots,k$, we get

$$\|\boldsymbol{Q}^\top \boldsymbol{W}^\top\|_{\mathrm{F}}^2 = \sum_{c=1}^k \|\boldsymbol{Q}^\top \boldsymbol{w}_c\|_2^2 \leq \sum_{c=1}^k \|\boldsymbol{e}_c^\top \boldsymbol{S}_\star\|_1^2 = \mathrm{Tr}\, \boldsymbol{A}_\star^\top \left(\boldsymbol{A}_\star \boldsymbol{A}_\star^\top\right)^{-1} \boldsymbol{A}_\star = m. \tag{8}$$

Since both $(\boldsymbol{A}_\natural, \boldsymbol{S}_\natural)$ and $(\boldsymbol{A}_\star, \boldsymbol{S}_\star)$ are feasible, $\boldsymbol{X} = \boldsymbol{A}_\natural \boldsymbol{S}_\natural = \boldsymbol{A}_\star \boldsymbol{S}_\star = \boldsymbol{A}_\star(\boldsymbol{W}\boldsymbol{S}_\natural + \boldsymbol{B})$. Multiplying both sides by $\boldsymbol{S}_\natural^\dagger$ gives

$$\boldsymbol{A}_\natural = \boldsymbol{A}_\natural \boldsymbol{S}_\natural \boldsymbol{S}_\natural^\dagger = \boldsymbol{A}_\star(\boldsymbol{W}\boldsymbol{S}_\natural + \boldsymbol{B})\boldsymbol{S}_\natural^\dagger = \boldsymbol{A}_\star \boldsymbol{W}.$$

Then we have

$$\log\det \boldsymbol{A}_\star \boldsymbol{A}_\star^\top \geq \log\det \boldsymbol{A}_\natural \boldsymbol{W}^\dagger (\boldsymbol{W}^\dagger)^\top \boldsymbol{A}_\natural^\top = \log\det \boldsymbol{A}_\natural \boldsymbol{A}_\natural^\top + \log\det \boldsymbol{Q}^\top \boldsymbol{W}^\dagger (\boldsymbol{W}^\dagger)^\top \boldsymbol{Q} \tag{9}$$

where the first inequality is shown in Appendix A. Regarding the second term, we have

$$\log\det \boldsymbol{Q}^\top \boldsymbol{W}^\dagger (\boldsymbol{W}^\dagger)^\top \boldsymbol{Q} \geq -\log\det \boldsymbol{Q}^\top \boldsymbol{W}^\top \boldsymbol{W} \boldsymbol{Q} \tag{10a}$$

$$\geq -m \log \frac{1}{m} \|\boldsymbol{Q}^\top \boldsymbol{W}^\top\|_{\mathrm{F}}^2, \tag{10b}$$

where (10a) and (10b) are shown Appendix A. Combining (8), (9), and (10) shows that

$$\log\det \boldsymbol{A}_\star \boldsymbol{A}_\star^\top \geq \log\det \boldsymbol{A}_\natural \boldsymbol{A}_\natural^\top. \tag{11}$$

On the other hand, since $(\boldsymbol{A}_\star, \boldsymbol{S}_\star)$ is optimal for (2), we have

$$\frac{1}{2}\log\det \boldsymbol{A}_\star \boldsymbol{A}_\star^\top + \max_{\|\boldsymbol{d}\|_2^2 = m} \sum_{c=1}^k d_c \|\boldsymbol{e}_c^\top \boldsymbol{S}_\star\|_1 \leq \frac{1}{2}\log\det \boldsymbol{A}_\natural \boldsymbol{A}_\natural^\top + \max_{\|\boldsymbol{d}\|_2^2 = m} \sum_{c=1}^k d_c \|\boldsymbol{e}_c^\top \boldsymbol{S}_\natural\|_1.$$

Lemma 1 and Assumption 1 shows that

$$\max_{\|\boldsymbol{d}\|_2^2 = m} \sum_{c=1}^k d_c \|\boldsymbol{e}_c^\top \boldsymbol{S}_\star\|_1 = \max_{\|\boldsymbol{d}\|_2^2 = m} \sum_{c=1}^k d_c \|\boldsymbol{e}_c^\top \boldsymbol{S}_\natural\|_1 = m,$$

therefore

$$\log\det \boldsymbol{A}_\star \boldsymbol{A}_\star^\top \leq \log\det \boldsymbol{A}_\natural \boldsymbol{A}_\natural^\top \tag{12}$$

Combining (11) and (12) shows that $\boldsymbol{A}_\natural$ when scaled according to Assumption 1, or any of its column permutation and/or sign flips, is optimal for (2). In Appendix A, we complete the proof that the second requirement of m-strongly scattered in the k-hypercube guarantees that every solution must satisfy $\boldsymbol{A}_\natural = \boldsymbol{A}_\star \boldsymbol{D}\boldsymbol{\Pi}$, where $\boldsymbol{D}$ is a diagonal matrix with only ± 1 on the diagonal, hence the overcomplete dictionary is identifiable. □

Our analysis so far has not explicitly mentioned the sparsity of $\boldsymbol{S}_\natural$, which may seem somewhat counter-intuitive. The explanation is two-fold: in the next subsection, we will show that if $\boldsymbol{S}_\natural$ follows a sparse generative model, then $\mathrm{cell}(\widetilde{\boldsymbol{S}}_\natural)$ will be m-strongly scattered in the k-hypercube with very high probability, thus sparsity is implicitly implied in Assumption 3. In particular, our analysis shows that it is necessary that every column of $\boldsymbol{S}_\natural$ contains no more than m nonzeros. On the other hand, Theorem 1 only shows that the dictionary $\boldsymbol{A}_\natural$ is identifiable, but for an overcomplete dictionary it does not necessarily mean that the sparse coefficient $\boldsymbol{S}_\natural$ is identifiable. Fortunately, with the knowledge of the dictionary, the identifiability of the sparse coefficients has been studied extensively (Donoho, 2006; Candès & Wakin, 2008). Here we provide a general result.

Assumption 4. Every column of $\boldsymbol{S}_\natural$ contains at most s nonzeros. In addition, $\boldsymbol{A}_\natural$ is a dictionary such that for every $\boldsymbol{s}_0$ with no more than s nonzeros, $\boldsymbol{s}_0$ is the unique solution to the following optimization problem

$$\underset{\boldsymbol{s}}{\text{minimize}} \quad \|\boldsymbol{s}\|_1 \qquad \text{subject to} \quad \boldsymbol{A}_\natural \boldsymbol{D}_\natural^{-1} \boldsymbol{s} = \boldsymbol{A}_\natural \boldsymbol{D}_\natural^{-1} \boldsymbol{s}_0,$$

where $\boldsymbol{D}_\natural$ is a diagonal matrix with cth diagonal defined in (6).

Corollary 1. *Consider the overcomplete DL model $\boldsymbol{X} = \boldsymbol{A}_\natural \boldsymbol{S}_\natural$, where $\boldsymbol{A}_\natural \in \mathbb{R}^{m\times k}$ is the groundtruth mixing matrix and $\boldsymbol{S}_\natural \in \mathbb{R}^{k\times n}$ is the groundtruth sparse coefficient matrix. Suppose $\boldsymbol{A}_\natural$ and $\boldsymbol{S}_\natural$ satisfies Assumptions 1–4. Then for any solution of* (2)*, denoted as $(\boldsymbol{A}_\star, \boldsymbol{S}_\star)$, there exist a permutation matrix $\boldsymbol{\Pi}$ and a diagonal matrix $\boldsymbol{D}$ such that $\boldsymbol{A}_\natural = \boldsymbol{A}_\star \boldsymbol{D}\boldsymbol{\Pi}$ and $\boldsymbol{S}_\natural = \boldsymbol{\Pi}^\top \boldsymbol{D}^{-1} \boldsymbol{S}_\star$.*

Proof. Theorem 1 shows that $\boldsymbol{A}_\natural$ is identifiable if Assumption 1–3 are satisfied. Assumption 4 guarantees that $\boldsymbol{S}_\natural$ is uniquely determined if $\boldsymbol{A}_\natural$ is given. To see this, we fix $\boldsymbol{A} = \boldsymbol{A}_\natural$ in (2). From Lemma 1, we know that the optimal $\boldsymbol{d}$ should be the diagonal of $\boldsymbol{D}_\natural$ as defined in Assumption 4. Then optimizing (2) with respect to $\boldsymbol{S}$ is equivalent to the following problem with a change-of-variable $\widetilde{\boldsymbol{S}} = \boldsymbol{D}_\natural \boldsymbol{S}$

$$\underset{\widetilde{\boldsymbol{S}}}{\text{minimize}} \ \ \|\widetilde{\boldsymbol{S}}\|_1 \qquad \text{subject to} \ \ \boldsymbol{X} = \boldsymbol{A}_\natural \boldsymbol{S}_\natural = \boldsymbol{A}_\natural \boldsymbol{D}_\natural^{-1} \widetilde{\boldsymbol{S}}.$$

If every column of $\boldsymbol{S}_\natural$ is at most s-sparse, then the optimal $\widetilde{\boldsymbol{S}}_\star = \boldsymbol{D}_\natural \boldsymbol{S}_\natural$, therefore $\boldsymbol{S}_\star = \boldsymbol{S}_\natural$. The result still holds if columns of $\boldsymbol{A}_\natural$ are permuted and/or multiplied with ± 1. □

2.3 Sample complexity analysis

Theorem 1 states that an overcomplete dictionary $\boldsymbol{A}_\natural$ is identifiable if Assumptions 1–3 are satisfied. Assumptions 1 and 2 are quite easy to satisfy, as it is very reasonable to assume that rows of wide matrices $\boldsymbol{A}_\natural$ and $\boldsymbol{S}_\natural$ are linearly independent, and given any $\boldsymbol{A}_\natural$ and $\boldsymbol{S}_\natural$ one can always find the scaling to satisfy (5). The most crucial condition is Assumption 3, or the fact that $\text{cell}(\widetilde{\boldsymbol{S}}_\natural)$ is m-strongly scattered in the k-hypercube. In this section we assume that a sparse coefficient matrix $\boldsymbol{S}$ is generated from the sparse-Gaussian model, which has appeared in (Wu & Yu, 2017; Wang et al., 2020), and show that in this case Assumption 3 is satisfied with high probability. This is a different generative model than prior works that also use a volume criterion for complete DL (Hu & Huang, 2023a; Sun & Huang, 2024), in which a Bernoulli-Gaussian model is considered. This is because such a generative model cannot guarantee that every column of $\boldsymbol{S}$ is at least s-sparse, thus Corollary 1 cannot be invoked to identify both $\boldsymbol{A}_\natural$ and $\boldsymbol{S}_\natural$.

Assumption 5 (Sparse-Gaussian model)**.** The matrix $\boldsymbol{S} \in \mathbb{R}^{k\times n}$ is generated from a sparse-Gaussian model with parameter $s < k$, denoted as $\boldsymbol{S} \sim \mathcal{SG}(s)$, if every column of $\boldsymbol{S}$ is independently and identically distributed from the following process: a subset $\mathcal{I}$ of size s is uniformly drawn from all size-s subsets of $\{1, \ldots, k\}$, let $\boldsymbol{s} \in \mathbb{R}^k$ be such that $s_i = 0$ if $i \in \mathcal{I}$ and $s_i \sim \mathcal{N}(0,1)$ if $i \notin \mathcal{I}$, where $\mathcal{N}(0,1)$ stands for a standard normal distribution.

To check whether $\mathcal{B}_m \subseteq \text{cell}(\widetilde{\boldsymbol{S}})$, where $\widetilde{\boldsymbol{S}}$ is obtained from scaling its rows to have unit ℓ_1 norms, it is easier to equivalently check its polar version $\text{cell}(\widetilde{\boldsymbol{S}})^\circ \subseteq \mathcal{B}_m^\circ$, where the polar of set $\mathcal{S}$ is defined as $\mathcal{S}^\circ = \{\boldsymbol{x} \mid \boldsymbol{x}^\top \boldsymbol{y} \le 1, \forall \boldsymbol{y} \in \mathcal{S}\}$. For $\text{cell}(\widetilde{\boldsymbol{S}})$, its polar has a relatively simple form

$$\text{cell}(\widetilde{\boldsymbol{S}})^\circ = \left\{\boldsymbol{w} \mid \|\boldsymbol{w}^\top \widetilde{\boldsymbol{S}}\|_1 \le 1\right\}.$$

The polar for $\mathcal{B}_m$ has a more complicated form

$$\mathcal{B}_m^\circ = \left\{\boldsymbol{w} \mid \|\boldsymbol{Q}^\top \text{Diag}(\|\boldsymbol{q}_1\|_2, \ldots, \|\boldsymbol{q}_k\|_2)^\dagger \boldsymbol{w}\|_2 \le 1, \ \forall\, \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}\right\}.$$

Therefore, checking whether $\text{cell}(\widetilde{\boldsymbol{S}})^\circ \subseteq \mathcal{B}_m^\circ$ is equivalent to checking whether the optimal value of the following problem equals 1:

$$\underset{\boldsymbol{Q},\boldsymbol{w}}{\text{maximize}} \ \ \|\boldsymbol{Q}^\top \boldsymbol{D}^\dagger \boldsymbol{w}\|_2^2 \qquad \text{subject to} \ \ \|\boldsymbol{w}^\top \widetilde{\boldsymbol{S}}\|_1 \le 1, \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}, \tag{13}$$

where $\boldsymbol{D} = \text{Diag}(\|\boldsymbol{q}_1\|_2, \ldots, \|\boldsymbol{q}_k\|_2)$.

Theorem 2. *Suppose $\boldsymbol{S} \in \mathbb{R}^{k\times n}$ is generated from the sparse-Gaussian model $\mathcal{SG}(s)$, where $s < m$, and $\widetilde{\boldsymbol{S}}$ is obtained by scaling its rows to have unit ℓ_1 norm. Then*

$$\Pr\left[\sup_{\substack{\|\boldsymbol{w}^\top \widetilde{\boldsymbol{S}}\|_1 \le 1 \\ \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}}} \|\boldsymbol{Q}^\top \boldsymbol{D}^\dagger \boldsymbol{w}\| > 1\right] \le 4\exp\left(\frac{k}{2}\log\frac{k^2}{m} - n\frac{s^2 m}{k^3}\right). \tag{14}$$

The probability goes to zero exponentially fast as

$$n = O\left(\frac{k^2}{m}\log\frac{k^2}{m}\right).$$

The proof is relegated to Appendix B. Comparing this result to prior work on complete DL (Hu & Huang, 2023a), in which case the sample complexity is $O(k \log k)$, we see that the bounds agree when $m = k$, which is a good sign that the bound is tight. On the other hand, there is one step in the proof that shows that for overcomplete DL, it is necessary that every column of $\boldsymbol{S}$ is at most s-sparse, where $s < m$; this is not required for complete DL. This shows the necessity of adopting the sparse-Gaussian model rather than the Bernoulli-Gaussian model, even if identifiability of $\boldsymbol{S}_{\natural}$ is not required. In fact, even the most relaxed condition on sparse recovery would require $s < m/2$, so assuming $\boldsymbol{S} \sim \mathcal{SG}(s)$ with $s < m$ is a very reasonable assumption in practice.

3 Algorithm via Alternating Minimization

We will now design an algorithm for the novel formulation (2) for overcomplete DL, whose global correctness in recovering the ground-truth dictionary has been theoretically established above. The main idea is similar to the (inexact) alternating optimization framework that most DL algorithms adopt. First of all, since most practical applications admit approximate factorization, we move the constraint $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S}$ in (2) as a penalty term as follows

$$\underset{\boldsymbol{A},\boldsymbol{S}}{\text{minimize}} \quad \frac{\lambda}{2}\|\boldsymbol{A}\boldsymbol{S} - \boldsymbol{X}\|_{\mathrm{F}}^2 + \frac{1}{2}\log\det \boldsymbol{A}\boldsymbol{A}^\top + \max_{\|\boldsymbol{d}\|_2^2 = m} \sum_{c=1}^{k} d_c \|\boldsymbol{e}_c^\top \boldsymbol{S}\|_1, \tag{15}$$

where the hyper-parameter λ balances data fidelity and the identifiability criterion. The rest of this section will focus on designing an iterative algorithm for solving (15). We denote $(\boldsymbol{A}_t, \boldsymbol{S}_t)$ as the updates obtained at the tth iteration. In an alternating fashion, $\boldsymbol{A}_{t+1}$ is obtained by fixing $\boldsymbol{S} = \boldsymbol{S}_t$, and $\boldsymbol{S}_{t+1}$ is obtained by fixing $\boldsymbol{A} = \boldsymbol{A}_{t+1}$.

3.1 Update of A

Fixing $\boldsymbol{S} = \boldsymbol{S}_t$, the update of $\boldsymbol{A}$ amounts to solving a log-determinant regularized least squares, which is nonconvex optimization problem. We propose two types of updates:

- Gradient descent. As the gradient of $(1/2)\log\det \boldsymbol{A}\boldsymbol{A}^\top$ at $\boldsymbol{A}_t$ is $(\boldsymbol{A}_t^\dagger)^\top$, a simple choice of update is to move along the negative gradient direction with step size γ

$$\boldsymbol{A}_{t+1} \leftarrow \boldsymbol{A}_t - \gamma\left(\lambda(\boldsymbol{A}\boldsymbol{S}_t - \boldsymbol{X})\boldsymbol{S}_t^\top + (\boldsymbol{A}_t^\dagger)^\top\right)$$

- Majorization minimization. Notice that the log determinant of a positive definite matrix is concave, therefore

$$\log\det \boldsymbol{A}\boldsymbol{A}^\top \le \log\det \boldsymbol{A}_t\boldsymbol{A}_t^\top + \mathrm{Tr}[(\boldsymbol{A}_t\boldsymbol{A}_t^\top)^{-1}(\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{A}_t\boldsymbol{A}_t^\top)].$$

  This defines a quadratic majorization function for $(1/2)\log\det \boldsymbol{A}\boldsymbol{A}^\top$. Minimizing this term plus the fitting term with respect to $\boldsymbol{A}$ amounts to solving the following linear equation:

$$\lambda(\boldsymbol{A}\boldsymbol{S}_t - \boldsymbol{X})\boldsymbol{S}_t^\top + (\boldsymbol{A}_t\boldsymbol{A}_t^\top)^{-1}\boldsymbol{A} = 0.$$

  This is Sylvester's equation, which can be solved by taking the eigen-decomposition of $\boldsymbol{S}_t\boldsymbol{S}_t^\top$ and $\boldsymbol{A}_t\boldsymbol{A}_t^\top$.

For simplicity, we resort to the gradient descent update in the rest of this paper.

3.2 Update of S

If $\boldsymbol{d}$ is fixed, then the subproblem of $\boldsymbol{S}$ is a ℓ_1 regularized least squares problem, which has been extensively studied. It is known to have no closed-form solutions, which is not preferable as one step of an iterative algorithm. Thus, we propose to update $\boldsymbol{S}$ with a proximal gradient step. Since it amounts to getting a linear approximation at $\boldsymbol{S}_t$, we also set $\boldsymbol{d}$ as the optimal choice with $\boldsymbol{S}_t$, which is easy to obtain from Cauchy-Schwarz:

$$d_c = \sqrt{\frac{m}{\sum_{j=1}^{k}\|\boldsymbol{e}_j^\top \boldsymbol{S}_t\|_1^2}}\|\boldsymbol{e}_c^\top \boldsymbol{S}_t\|_1. \tag{16}$$

As a result, the proximal gradient update of $\boldsymbol{S}$ with step size γ take the form

$$\boldsymbol{S}_{t+1} \leftarrow \mathcal{T}_{\gamma \boldsymbol{d}} \left(\boldsymbol{S}_t - \gamma \lambda \boldsymbol{A}_{t+1}^\top (\boldsymbol{A}_{t+1} \boldsymbol{S}_t - \boldsymbol{X}) \right),$$

where $\mathcal{T}_{\gamma \boldsymbol{d}}(\cdot)$ is the soft-thresholding operator on a matrix with k rows, and the threshold for components on the cth row is d_c/γ.

3.3 Summary and experimental demonstration

The proposed algorithm based on alternating minimization is summarized in

Algorithm 1 Solving (14) via alternating minimization

initialize $\boldsymbol{A}_0$ and $\boldsymbol{S}_0$
for $t = 0, 1, 2, \ldots$ until convergence **do**
  $\boldsymbol{A}_{t+1} \leftarrow \boldsymbol{A}_t - \gamma \left(\lambda (\boldsymbol{A}\boldsymbol{S}_t - \boldsymbol{X}) \boldsymbol{S}_t^\top + (\boldsymbol{A}_t^\dagger)^\top \right)$
  for $c = 1, \ldots, k$ **do**
    $d_c = \sqrt{\frac{m}{\sum_{j=1}^k \|\boldsymbol{e}_j^\top \boldsymbol{S}_t\|_1^2}} \|\boldsymbol{e}_c^\top \boldsymbol{S}_t\|_1$
  end for
  $\boldsymbol{S}_{t+1} \leftarrow \mathcal{T}_{\gamma \boldsymbol{d}} \left(\boldsymbol{S}_t - \gamma \lambda \boldsymbol{A}_{t+1}^\top (\boldsymbol{A}_{t+1} \boldsymbol{S}_t - \boldsymbol{X}) \right)$
end for

As (14) is nonconvex and NP-hard, no known algorithm is able to guarantee convergence to a global optimum. In the following, we provide a brief demonstration of the performance of the proposed algorithm. Admittedly, the proposed formulation with identifiability guarantees opens up a new direction for research on algorithm design for dictionary learning, which is a challenging task in itself as it involves several terms that are nontrivial to handle.

We synthetically generate random problems with $s = 5$, $m = 10$, $k = 20$, and $n = 200$. A groundtruth sparse coefficient matrix $\boldsymbol{S}_\natural \in \mathbb{R}^{k \times n}$ is generated from the sparse-Gaussian model $\mathcal{SG}(s)$, while the groundtruth dictionary $\boldsymbol{A}_\natural \in \mathbb{R}^{m \times k}$ is simply generated from a standard normal distribution. The data matrix is then generated as $\boldsymbol{X} = \boldsymbol{A}_\natural \boldsymbol{S}_\natural$. Algorithm 1 runs on $\boldsymbol{X}$ with $\lambda = 1000$ (since we want the data fidelity term to be almost zero) and $\gamma = 10^{-5}$. At the end, both columns of $\boldsymbol{A}_\natural$ and $\boldsymbol{A}_\star$ are scaled to unit ℓ_2 norm, and the Hungarian algorithm (Kuhn, 1955) is used to find the best column matching. The resulting estimation error $\|\boldsymbol{A}_\natural - \boldsymbol{A}_\star\|_{\mathrm{F}}^2$ remains approximately 10^{-8} over multiple runs. Considering the $\boldsymbol{A}_\natural$ is 10×20 with unit column norms this is a satisfactory initial result.

4 Conclusion

In this paper, we provide perhaps the first identifiability analysis of a matrix factorization model when the latent dimension is *higher* than the ambient dimension, namely the overcomplete dictionary learning problem. Classical works on this problem rely on combinatorial mathematics, which in turn requires the sample size to be factorial to the latent dimension. Our work is based on a novel formulation that uses a hybrid of weighted ℓ_1 norm of the sparse coefficients and the volume of the overcomplete dictionary as the identification criterion, and we show that identifiability of the overcomplete dictionary of size $m \times k$ can be guaranteed if a geometric condition is satisfied, namely the cellular hull of the sparse coefficient matrix is m-strongly scattered in the k-hypercube. If the sparse coefficient matrix is generated from the sparse-Gaussian model, then such identifiability condition can be satisfied with very high probability if the sample size is $O((k^2/m)\log(k^2/m))$, which is a huge improvement compared to prior work with factorial complexity. We also propose an algorithm for the novel overcomplete dictionary learning formulation. The proposed novel formulation for overcomplete DL with a global identifiability guarantee leaves much room for faster and more efficient algorithm design.

Acknowledgments

This work is supported in part by NSF ECCS-2237640 and NIH R01LM014027.

REFERENCES

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, and Praneeth Netrapalli. Learning sparsely used overcomplete dictionaries via alternating minimization. *SIAM Journal on Optimization*, 26 (4):2775–2799, 2016.

Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11): 4311–4322, 2006a.

Michal Aharon, Michael Elad, and Alfred M Bruckstein. On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. *Linear algebra and its applications*, 416(1): 48–67, 2006b.

Sanjeev Arora, Rong Ge, and Ankur Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *Conference on Learning Theory*, pp. 779–806. PMLR, 2014.

Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. In *Conference on Learning Theory*, pp. 113–149. PMLR, 2015.

Yu Bai, Qijia Jiang, and Ju Sun. Subgradient descent learns orthogonal dictionaries. In *International Conference on Learning Representations*, 2019.

Adi Ben-Israel. A volume associated with mxn matrices. *Linear algebra and its applications*, 167: 87–111, 1992.

George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

Emmanuel J Candès and Michael B Wakin. An introduction to compressive sampling. *IEEE signal processing magazine*, 25(2):21–30, 2008.

Jeremy E Cohen and Nicolas Gillis. Identifiability of complete dictionary learning. *SIAM Journal on Mathematics of Data Science*, 1(3):518–536, 2019.

David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12):3736–3745, 2006.

Xiao Fu, Wing-Kin Ma, Kejun Huang, and Nicholas D Sidiropoulos. Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain. *IEEE Transactions on Signal Processing*, 63(9):2306–2320, 2015.

Charles J Garfinkle and Christopher J Hillar. On the uniqueness and stability of dictionaries for sparse representation of noisy signals. *IEEE Transactions on Signal Processing*, 67(23):5884–5892, 2019.

Quan Geng and John Wright. On the local correctness of l1-minimization for dictionary learning. In *2014 IEEE International Symposium on Information Theory*, pp. 3180–3184. IEEE, 2014.

Pando Georgiev, Fabian Theis, and Andrzej Cichocki. Sparse component analysis and blind source separation of underdetermined mixtures. *IEEE transactions on neural networks*, 16(4):992–996, 2005.

Rémi Gribonval and Karin Schnass. Dictionary identification——sparse matrix-factorization via ℓ_1-minimization. *IEEE Transactions on Information Theory*, 56(7):3523–3539, 2010.

Christopher J Hillar and Friedrich T Sommer. When can dictionary learning uniquely recover sparse data from subsamples? *IEEE Transactions on Information Theory*, 61(11):6290–6297, 2015.

Jingzhou Hu and Kejun Huang. Global identifiability of ℓ_1-based dictionary learning via matrix volume optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023a.

Jingzhou Hu and Kejun Huang. Identifiable bounded component analysis via minimum volume enclosing parallelotope. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023b.

Jingzhou Hu and Kejun Huang. Complex bounded component analysis: Identifiability and algorithm. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.

Kejun Huang and Xiao Fu. Detecting overlapping and correlated communities without pure nodes: Identifiability and algorithm. In *International Conference on Machine Learning*, pp. 2859–2868, 2019.

Kejun Huang, Nicholas D Sidiropoulos, and Ananthram Swami. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. *IEEE Transactions on Signal Processing*, 62(1):211–224, 2013.

Kejun Huang, Xiao Fu, and Nikolaos D Sidiropoulos. Anchor-free correlated topic modeling: Identifiability and algorithm. *Advances in Neural Information Processing Systems*, 29, 2016.

Kejun Huang, Xiao Fu, and Nicholas D. Sidiropoulos. Learning hidden Markov models from pairwise co-occurrences with application to topic modeling. In *International Conference on Machine Learning (ICML)*, pp. 2073–2082, 2018.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Chia-Hsiang Lin, Wing-Kin Ma, Wei-Chiang Li, Chong-Yung Chi, and ArulMurugan Ambikapathi. Identifiability of the simplex volume minimization criterion for blind hyperspectral unmixing: The no-pure-pixel case. *IEEE Transactions on Geoscience and Remote Sensing*, 53(10):5530–5546, 2015.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, pp. 689–696, 2009.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311–3325, 1997.

Gilles Pisier. *The volume of convex bodies and Banach space geometry*, volume 94. Cambridge University Press, 1999.

Mark D Plumbley, Thomas Blumensath, Laurent Daudet, Rémi Gribonval, and Mike E Davies. Sparse representations in audio and music: from coding to source separation. *Proceedings of the IEEE*, 98 (6):995–1005, 2009.

Sirisha Rambhatla, Xingguo Li, and Jarvis Haupt. Noodl: Provable online dictionary learning and sparse coding. In *International Conference on Learning Representations*, 2019.

Yifei Shen, Ye Xue, Jun Zhang, Khaled Letaief, and Vincent Lau. Complete dictionary learning via lp-norm maximization. In *Conference on Uncertainty in Artificial Intelligence*, pp. 280–289. PMLR, 2020.

Daniel A. Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In Shie Mannor, Nathan Srebro, and Robert C. Williamson (eds.), *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pp. 37.1–37.18, Edinburgh, Scotland, 25–27 Jun 2012. PMLR.

Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere i: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2016a.

Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere ii: Recovery by riemannian trust-region method. *IEEE Transactions on Information Theory*, 63(2):885–914, 2016b.

Yuchen Sun and Kejun Huang. Improved identifiability and sample complexity analysis of complete dictionary learning. In *2024 IEEE 13rd Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pp. 1–5. IEEE, 2024.

Gokcan Tatli and Alper T Erdogan. Polytopic matrix factorization: Determinant maximization based criterion and identifiability. *IEEE Transactions on Signal Processing*, 69:5431–5447, 2021.

Bahareh Tolooshams and Demba E Ba. Stable and interpretable unrolled dictionary learning. *Transactions on Machine Learning Research*, 2022.

Ivana Tošić and Pascal Frossard. Dictionary learning. *IEEE Signal Processing Magazine*, 28(2): 27–38, 2011.

Ivana Tošić, Ivana Jovanović, Pascal Frossard, Martin Vetterli, and Neb Durić. Ultrasound tomography with learned dictionaries. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5502–5505. IEEE, 2010.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Yu Wang, Siqi Wu, and Bin Yu. Unique sharp local minimum in l1-minimization complete dictionary learning. *Journal of Machine Learning Research*, 21(63):1–52, 2020.

Siqi Wu and Bin Yu. Local identifiability of l1-minimization dictionary learning: a sufficient and almost necessary condition. *The Journal of Machine Learning Research*, 18(1):6121–6176, 2017.

Yuexiang Zhai, Hermish Mehta, Zhengyuan Zhou, and Yi Ma. Understanding l4-based dictionary learning: Interpretation, stability, and robustness. In *International Conference on Learning Representations*, 2020a.

Yuexiang Zhai, Zitong Yang, Zhenyu Liao, John Wright, and Yi Ma. Complete dictionary learning via l4-norm maximization over the orthogonal group. *Journal of Machine Learning Research*, 21 (165):1–68, 2020b.

A PROOF OF THEOREM 1

In the proof sketch of Theorem 1, we showed that

$$\frac{1}{2}\log\det \boldsymbol{A}_\natural \boldsymbol{A}_\natural^\top + \max_{\|\boldsymbol{d}\|_2^2=m} \sum_{c=1}^k d_c \|\boldsymbol{e}_c^\top \boldsymbol{S}_\natural\|_1 = \frac{1}{2}\log\det \boldsymbol{A}_\star \boldsymbol{A}_\star^\top + \max_{\|\boldsymbol{d}\|_2^2=m} \sum_{c=1}^k d_c \|\boldsymbol{e}_c^\top \boldsymbol{S}_\star\|_1,$$

so $(\boldsymbol{A}_\natural, \boldsymbol{S}_\natural)$ is at least one candidate solution for (2). In this section, we first complete the proof by showing that if the second requirement of Assumption 3 is satisfied, then any optimal solution $(\boldsymbol{A}_\star, \boldsymbol{S}_\star)$ must satisfy that $\boldsymbol{A}_\star = \boldsymbol{A}_\natural \boldsymbol{\Pi}\boldsymbol{D}$, where $\boldsymbol{\Pi}$ is a permutation matrix and $\boldsymbol{D}$ is a diagonal matrix with ± 1 on the diagonal.

In (8) we proved that $\|\boldsymbol{e}_c^\top \boldsymbol{S}_\star\|_1 \geq \|\boldsymbol{Q}^\top \boldsymbol{w}_c\|_2$ by choosing $\|\boldsymbol{\theta}\|_\infty \leq 1$ such that

$$\widetilde{\boldsymbol{S}}\boldsymbol{\theta} = \boldsymbol{D}_\natural^{-1}\boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{w}_c / \|\boldsymbol{Q}^\top \boldsymbol{w}_c\|_2.$$

Suppose $\boldsymbol{Q}^\top \boldsymbol{w}_c/\|\boldsymbol{Q}^\top \boldsymbol{w}_c\|_2 \neq \pm \boldsymbol{q}_c/\|\boldsymbol{q}_c\|_2$ for all $c = 1, \ldots, k$, then according to the second requirement of Assumption 3, it is in the interior of $\mathrm{cell}(\widetilde{\boldsymbol{S}}_\natural)$, which means there exists $\alpha > 1$ and $\|\tilde{\boldsymbol{\theta}}\|_\infty \leq 1$ such that

$$\widetilde{\boldsymbol{S}}\tilde{\boldsymbol{\theta}} = \alpha \boldsymbol{D}_\natural^{-1}\boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{w}_c / \|\boldsymbol{Q}^\top \boldsymbol{w}_c\|_2.$$

Therefore

$$\|\boldsymbol{e}_c^\top \boldsymbol{S}_\star\|_1 = \|\boldsymbol{w}_c^\top \boldsymbol{S}\|_1 \geq \boldsymbol{w}_c \boldsymbol{D}_\natural \widetilde{\boldsymbol{S}}\tilde{\boldsymbol{\theta}} = \alpha \|\boldsymbol{Q}^\top \boldsymbol{w}_c\|_2 > \|\boldsymbol{Q}^\top \boldsymbol{w}_c\|_2.$$

This means equality of $\|\boldsymbol{e}_c^\top \boldsymbol{S}_\star\|_1 \geq \|\boldsymbol{Q}^\top \boldsymbol{w}_c\|_2$ is only attained when $\boldsymbol{Q}^\top \boldsymbol{w}_c = \pm \boldsymbol{q}_c$ for some $c = 1, \ldots, k$.

Finally, it is proven in (10b) that

$$\log\det \boldsymbol{Q}^\top \boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{Q} \leq m \log \frac{1}{m}\|\boldsymbol{Q}^\top \boldsymbol{W}\|_{\mathrm{F}}^2,$$

and if $\det \boldsymbol{A}_\natural \boldsymbol{A}_\natural^\top = \det \boldsymbol{A}_\star \boldsymbol{A}_\star^\top$ the above inequality must hold as an equality. In the proof of (10b) we showed that equality holds only if all the eigenvalues of $\boldsymbol{Q}^\top \boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{Q}$ are equal, meaning columns of $\boldsymbol{W}\boldsymbol{Q}$ are orthonormal. Since $\boldsymbol{Q}$ itself is orthonormal, it is possible only if $\boldsymbol{W}\boldsymbol{Q} = \boldsymbol{D}\boldsymbol{\Pi}^\top \boldsymbol{Q}$, where $\boldsymbol{D}$ is a diagonal matrix with ± 1 on the diagonals and $\boldsymbol{\Pi}$ is a permutation matrix. This shows that

$$\boldsymbol{A}_\star = \boldsymbol{A}_\natural \boldsymbol{\Pi}\boldsymbol{D}.$$

Q.E.D.

The remaining of this section shows some key equalities and inequalities in the proof sketch that are skipped for clarity.

Proof that $\|\boldsymbol{e}_c^T \boldsymbol{S}_\star\|_1 = \|\boldsymbol{w}_c^\top \boldsymbol{S}_\natural\|_1$. As we argued in the proof sketch of Theorem 1, the constraint $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S}$ is equivalent to $\boldsymbol{A}\boldsymbol{W} = \boldsymbol{A}_\natural$ and $\boldsymbol{S} = \boldsymbol{W}\boldsymbol{S}_\natural + \boldsymbol{B}$ where $\boldsymbol{S}_\natural \boldsymbol{B}^\top = 0$. Substituting them into (2) eliminates the constraint $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S}$:

$$\underset{\boldsymbol{A},\boldsymbol{W},\boldsymbol{B}}{\text{minimize}} \quad \frac{1}{2}\log\det \boldsymbol{A}\boldsymbol{A}^\top + \max_{\|\boldsymbol{d}\|_2^2=m} \sum_{c=1}^k d_c \|\boldsymbol{w}_c^\top \boldsymbol{S}_\natural + \boldsymbol{b}_c^\top\|_1 \qquad \text{subject to } \boldsymbol{A}\boldsymbol{W} = \boldsymbol{A}_\natural, \boldsymbol{S}_\natural \boldsymbol{B}^\top = 0.$$

Taking the Clarke generalized derivative of the Lagrange function with respect to $\boldsymbol{b}_c$ and setting it equal to zero gives

$$d_c \boldsymbol{\theta}_c + \boldsymbol{S}_\natural^\top \boldsymbol{\mu}_c = 0,$$

where $\boldsymbol{\theta}_c = \mathrm{sign}(\boldsymbol{S}_\natural^\top \boldsymbol{w}_c + \boldsymbol{b}_c)$ and $\boldsymbol{\mu}_c$ is the Lagrange multiplier for the cth column of $\boldsymbol{S}_\natural \boldsymbol{B}^\top$. This shows that $\boldsymbol{\theta}_c$ is a linear combination of rows of $\boldsymbol{S}_\natural$, which means it is orthogonal to $\boldsymbol{b}_c$. As a result,

$$\|\boldsymbol{e}_c^\top \boldsymbol{S}_\star\|_1 = \|\boldsymbol{w}_c^\top \boldsymbol{S}_\natural + \boldsymbol{b}_c^\top\|_1 = \boldsymbol{w}_c^\top \boldsymbol{S}_\natural \boldsymbol{\theta}_c + \boldsymbol{b}_c^\top \boldsymbol{\theta}_c = \boldsymbol{w}_c^\top \boldsymbol{S}_\natural \boldsymbol{\theta}_c = \|\boldsymbol{w}_c^\top \boldsymbol{S}_\natural\|_1.$$

□

Proof of inequality (9). The matrix $\boldsymbol{W}\boldsymbol{W}^\dagger$ defines a projection matrix, which is symmetric, with the following properties:

$$\boldsymbol{W}\boldsymbol{W}^\dagger \preceq \boldsymbol{I}, \quad \boldsymbol{W}\boldsymbol{W}^\dagger = \boldsymbol{W}\boldsymbol{W}^\dagger\boldsymbol{W}\boldsymbol{W}^\dagger = \boldsymbol{W}\boldsymbol{W}^\dagger(\boldsymbol{W}\boldsymbol{W}^\dagger)^\top = \boldsymbol{W}\boldsymbol{W}^\dagger(\boldsymbol{W}^\dagger)^\top\boldsymbol{W}^\top.$$

As a result,

$$\det \boldsymbol{A}_\star\boldsymbol{A}_\star^\top \geq \det \boldsymbol{A}_\star\boldsymbol{W}\boldsymbol{W}^\dagger\boldsymbol{A}_\star^\top = \det \boldsymbol{A}_\star\boldsymbol{W}\boldsymbol{W}^\dagger(\boldsymbol{W}^\dagger)^\top\boldsymbol{W}^\top\boldsymbol{A}_\star^\top = \det \boldsymbol{A}_\natural\boldsymbol{W}^\dagger(\boldsymbol{W}^\dagger)^\top\boldsymbol{A}_\natural^\top,$$

where the last step is because $\boldsymbol{A}_\natural = \boldsymbol{A}_\star\boldsymbol{W}$. □

Proof of inequality (10a). Since $\boldsymbol{A}_\natural = \boldsymbol{A}_\star\boldsymbol{W}$, rows of $\boldsymbol{A}_\natural$ are in the row space of $\boldsymbol{W}$, and so are columns of $\boldsymbol{Q}$. Therefore

$$\boldsymbol{W}\boldsymbol{W}^\dagger\boldsymbol{Q} = \boldsymbol{Q}.$$

This means $\boldsymbol{W}\boldsymbol{Q}$ has linearly independent columns, therefore $\boldsymbol{Q}^\top\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{Q}$ is invertible and

$$(\boldsymbol{Q}^\top\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{Q})^{-1} = (\boldsymbol{W}\boldsymbol{Q})^\dagger(\boldsymbol{Q}^\top\boldsymbol{W}^\top)^\dagger.$$

On the other hand, since $\boldsymbol{Q}^\top\boldsymbol{W}\boldsymbol{W}^\dagger\boldsymbol{Q} = \boldsymbol{I}$, this means

$$\boldsymbol{Q}^\top\boldsymbol{W}^\dagger = (\boldsymbol{W}\boldsymbol{Q})^\dagger + \boldsymbol{C},$$

where rows of $\boldsymbol{C}$ are orthogonal to columns of $\boldsymbol{W}\boldsymbol{Q}$, i.e., $\boldsymbol{C}\boldsymbol{W}\boldsymbol{Q} = 0$. Columns of $\boldsymbol{W}\boldsymbol{Q}$ and rows of $(\boldsymbol{W}\boldsymbol{Q})^\dagger$ span the same subspace, so we also have $\boldsymbol{C}^\top(\boldsymbol{W}\boldsymbol{Q})^\dagger = 0$. As a result,

$$\begin{aligned}
\det \boldsymbol{Q}^\top\boldsymbol{W}^\dagger(\boldsymbol{W}^\dagger)^\top\boldsymbol{Q} &= \det\left((\boldsymbol{W}\boldsymbol{Q})^\dagger + \boldsymbol{C}\right)\left((\boldsymbol{W}\boldsymbol{Q})^\dagger + \boldsymbol{C}\right)^\top \\
&= \det\left((\boldsymbol{W}\boldsymbol{Q})^\dagger(\boldsymbol{Q}^\top\boldsymbol{W}^\top)^\dagger + \boldsymbol{C}\boldsymbol{C}^\top\right) \\
&\geq \det(\boldsymbol{W}\boldsymbol{Q})^\dagger(\boldsymbol{Q}^\top\boldsymbol{W}^\top)^\dagger = \det(\boldsymbol{Q}^\top\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{Q})^{-1}.
\end{aligned}$$

Taking the log on both sides shows (10a). □

Proof of inequality (10b). Denote the eigenvalues of $\boldsymbol{Q}^\top\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{Q}$ as $\lambda_1, \ldots, \lambda_m$, which are all nonnegative, then

$$\left(\det \boldsymbol{Q}^\top\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{Q}\right)^{1/m} = \left(\prod_{j=1}^m \lambda_j\right)^{1/m} \leq \frac{1}{m}\sum_{j=1}^m \lambda_j = \operatorname{Tr} \boldsymbol{Q}^\top\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{Q} = \|\boldsymbol{Q}^\top\boldsymbol{W}^\top\|_{\mathrm{F}}^2,$$

where the inequality in the middle is the geometric-arithmetic mean inequality. Taking the log on both sides and rearranging gives (10b). Notice that equality holds only if $\lambda_1 = \cdots = \lambda_m$, i.e., columns of $\boldsymbol{W}\boldsymbol{Q}$ are orthonormal. □

B Proof of Theorem 2

We assume that $\boldsymbol{S} \sim \mathcal{SG}(s)$, and the first thing we do is rescale its rows to have unit ℓ_1 norms and use it for the problem 13. To simplify the analysis, we can instead directly maximize $\|\boldsymbol{Q}^\top\boldsymbol{D}^{-1}\boldsymbol{w}\|^2$ subject to $\|\boldsymbol{w}^\top\tilde{\boldsymbol{S}}\|_1 \leq 1$, and compare it with the largest ℓ_1 norm of the rows of $\boldsymbol{S}$. The complement of the intended probability can be bounded as

$$\Pr\left[\sup_{\substack{\|\boldsymbol{w}^\top\tilde{\boldsymbol{S}}\|_1\leq 1\\ \boldsymbol{Q}^\top\boldsymbol{Q}=\boldsymbol{I}}} \|\boldsymbol{Q}^\top\boldsymbol{D}^{-1}\boldsymbol{w}\| \leq 1\right] \geq \Pr\left[\sup_{\substack{\|\boldsymbol{w}^\top\boldsymbol{S}\|_1\leq 1\\ \boldsymbol{Q}^\top\boldsymbol{Q}=\boldsymbol{I}}} \|\boldsymbol{Q}^\top\boldsymbol{D}^{-1}\boldsymbol{w}\| \leq \alpha \cap \max_j \|\boldsymbol{S}_{j,:}\|_1 \geq \alpha\right],$$

with an arbitrary choice of α. Conversely,

$$\begin{aligned}
\Pr\left[\sup_{\substack{\|\boldsymbol{w}^\top\tilde{\boldsymbol{S}}\|_1\leq 1\\ \boldsymbol{Q}^\top\boldsymbol{Q}=\boldsymbol{I}}} \|\boldsymbol{Q}^\top\boldsymbol{D}^{-1}\boldsymbol{w}\| > 1\right] &\leq \Pr\left[\sup_{\substack{\|\boldsymbol{w}^\top\boldsymbol{S}\|_1\leq 1\\ \boldsymbol{Q}^\top\boldsymbol{Q}=\boldsymbol{I}}} \|\boldsymbol{w}\| > \alpha \cup \max_j \|\boldsymbol{S}_{j,:}\|_1 < \alpha\right] \\
&\leq \Pr\left[\sup_{\substack{\|\boldsymbol{w}^\top\boldsymbol{S}\|_1\leq 1\\ \boldsymbol{Q}^\top\boldsymbol{Q}=\boldsymbol{I}}} \|\boldsymbol{Q}^\top\boldsymbol{D}^{-1}\boldsymbol{w}\| > \alpha\right] + \Pr\left[\max_j \|\boldsymbol{S}_{j,:}\|_1 < \alpha\right] \qquad (17)
\end{aligned}$$

where the second inequality is obtained from the union bound. The rest of this section is dedicated to bounding the above two terms. Both of these results rely on the following version of the Bernstein inequality (Bennett, 1962):

Theorem 3 (Bernstein's inequality)**.** *Let $Z_1, \ldots, Z_n$ be independent random variables with $\mathrm{E}[Z_i^2] \leq v^2$ and there exists some constant c such that for all integer $d > 2$*

$$\mathrm{E}[|Z_i|^d] \leq \frac{1}{2} d! v^2 c^{d-2}. \tag{18}$$

Then

$$\Pr\left[\left|\sum_{i=1}^{n} (Z_i - \mathrm{E}[Z_i])\right| > \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2}{2(nv^2 + c\epsilon)}\right)$$

Lemma 2 (Bounding the second term in (17))**.** *Suppose $\boldsymbol{S} \in \mathbb{R}^{k\times n}$ is generated from the sparse-Gaussian model $\mathcal{SG}(s)$. Then*

$$\Pr\left[\max_j \|\boldsymbol{S}_{j,:}\|_1 < n(s/k)(\sqrt{2/\pi} - \epsilon)\right] \leq 2k\exp\left(-\frac{n(s/k)\epsilon^2}{2 + \sqrt{2}\epsilon}\right)$$

Proof. Let $\boldsymbol{s} = (s_1, \ldots, s_n)$ be one row of $\boldsymbol{S}$ generated from $\mathcal{SG}(s)$, then each s_i has probability s/k to be standard normal and probability $1 - s/k$ to be zero. We will use Bernstein's inequality with $Z_i = |s_i|$. Let g denote a standard normal random variable, then $|g|$ follows a Chi-distribution of degree 1, so its moments are

$$\mathrm{E}[|g|^d] = 2^{d/2}\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)}. \tag{19}$$

As a result, we have $\mathrm{E}[Z_i] = (s/k)\sqrt{2/\pi}$ and $\mathrm{E}[Z_i^2] = s/k$. Using the recurrence relation for the Gamma function $\Gamma(t+1) = t\Gamma(t)$ and $\sqrt{2}/\Gamma(1/2) = \sqrt{2/\pi} < 1$ we can bound the rest of the moments with $d > 2$ as

$$\mathrm{E}[|Z_i|^d] \leq (s/k)\frac{d!}{2^{d/2}}. \tag{20}$$

Therefore the moments satisfy (18) with $c = 1/\sqrt{2}$, results in

$$\begin{aligned}
\Pr\left[\|\boldsymbol{s}\|_1 < n(s/k)(\sqrt{2/\pi} - \epsilon)\right] &\leq \Pr\left[\left|\sum_{i=1}^{n} (Z_i - \mathrm{E}[Z_i])\right| > n(s/k)\epsilon\right] \\
&\leq 2\exp\left(-\frac{n^2(s/k)^2\epsilon^2}{2(n(s/k) + n(s/k)\epsilon/\sqrt{2})}\right) \\
&= 2\exp\left(-\frac{n(s/k)\epsilon^2}{2 + \sqrt{2}\epsilon}\right).
\end{aligned}$$

Finally, using the union bound

$$\begin{aligned}
\Pr\left[\max_j \|\boldsymbol{S}_{j,:}\|_1 < n(s/k)(\sqrt{2/\pi} - \epsilon)\right] &\leq k\Pr\left[\|\boldsymbol{S}_{j,:}\|_1 < n(s/k)(\sqrt{2/\pi} - \epsilon)\right] \\
&\leq 2k\exp\left(-\frac{n(s/k)\epsilon^2}{2 + \sqrt{2}\epsilon}\right).
\end{aligned}$$

□

We now proceed to bound the first term in (17). First, we note the following equivalence:

$$\Pr\left[\sup_{\substack{\|\boldsymbol{w}^\top \boldsymbol{S}\|_1 \leq 1 \\ \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}}} \|\boldsymbol{Q}^\top \boldsymbol{D}^{-1} \boldsymbol{w}\| > \alpha\right] = \Pr\left[\inf_{\substack{\|\boldsymbol{Q}^\top \boldsymbol{D}^{-1} \boldsymbol{w}\| = 1 \\ \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}}} \|\boldsymbol{S}^\top \boldsymbol{w}\|_1 < 1/\alpha\right] \tag{21}$$

We are also going to use the following notion of δ-cover from convex geometry (Pisier, 1999):

Definition 3 (δ-cover)**.** A finite δ-cover of a set $\mathcal{S}$ in $\mathbb{R}^k$ is a finite set $\mathcal{N}(\mathcal{S}, \delta)$ of points on $\mathcal{S}$ such that any point on $\mathcal{S}$ is within ϵ away from an element in $\mathcal{N}(\mathcal{S}, \delta)$, i.e.

$$\min_{\boldsymbol{w}_i \in \mathcal{N}(\mathcal{S},\delta)} \|\boldsymbol{w} - \boldsymbol{w}_i\| < \delta, \ \ \forall \boldsymbol{b} \in \mathcal{S}.$$

A well-known result is that the δ-cover of the unit sphere $\mathcal{B}_k = \{\boldsymbol{w} \mid \|\boldsymbol{w}\|_2 = 1\}$ has bounded cardinality (Vershynin, 2018):

$$|\mathcal{N}(\mathcal{B}_k, \delta)| \leq \left(\frac{2}{\delta} + 1\right)^k.$$

Notice that $\mathcal{B}_k$ is a special case of $\mathcal{B}_m$ defined in Assumption 3 with $k = m$, implying that $\mathcal{B}_k \subseteq \mathcal{B}_m$ and thus $\mathcal{B}_m^\circ \subseteq \mathcal{B}_k^\circ$. We will use this result to prove a (rather loose) bound for the δ-cover of $\mathcal{B}_m^\circ$.

Lemma 3. *The δ-cover of $\mathcal{B}_m^\circ$ satisfies*

$$|\mathcal{N}(\mathcal{B}_m, \delta)|_2 \leq \left(\frac{2}{\delta} + 1\right)^k$$

Proof. It is easy to see that $\mathcal{B}_k$ is self-polar, i.e., $\mathcal{B}_k^\circ = \mathcal{B}_k$, so $\mathcal{B}_m^\circ \subseteq \mathcal{B}_k$. Since the superset $\mathcal{B}_k$ satisfies $|\mathcal{N}(\mathcal{B}_k, \delta)| \leq (1 + 2/\delta)^k$, so is the subset $\mathcal{B}_m^\circ$. □

Lemma 4. *Let $\mathcal{N}(\mathcal{B}_m, \delta) = \{\boldsymbol{w}_i\}$ be a δ-cover for $\mathcal{B}_m^\circ$ in $\mathbb{R}^k$. Assume that we have both the lowerbound*

$$\|\boldsymbol{S}^\top \boldsymbol{w}_i\|_1 \geq \beta, \forall \boldsymbol{w}_i \in \mathcal{N}(\mathcal{B}_m, \delta)$$

and the upperbound

$$\|\boldsymbol{S}^\top\|_1 = \sup_{\|\boldsymbol{w}\|_1 \leq 1} \|\boldsymbol{S}^\top \boldsymbol{w}\|_1 \leq \gamma.$$

Then

$$\inf_{\boldsymbol{w} \in \mathcal{B}_m^\circ} \|\boldsymbol{S}^\top \boldsymbol{w}\|_1 \geq \beta - \gamma\delta\sqrt{k}$$

Proof. By definition of the δ-cover, for all $\boldsymbol{w} \in \mathcal{B}_m^\circ$ we can find $\boldsymbol{w}_i \in \mathcal{N}(\mathcal{B}_m^\circ, \delta)$ with $\|\boldsymbol{w} - \boldsymbol{w}_i\| < \delta$. Therefore

$$\begin{aligned}\|\boldsymbol{S}^\top \boldsymbol{w}\|_1 \geq \|\boldsymbol{S}^\top \boldsymbol{w}_i\|_1 - \|\boldsymbol{S}^\top(\boldsymbol{w} - \boldsymbol{w}_i)\|_1 &\geq \beta - \|\boldsymbol{S}^\top\|_1 \|\boldsymbol{w} - \boldsymbol{w}_i\|_1 \\ &\geq \beta - \|\boldsymbol{S}^\top\|_1 \|\boldsymbol{w} - \boldsymbol{w}_i\|_2 \sqrt{k} \geq \beta - \gamma\delta\sqrt{k}.\end{aligned}$$

□

Lemma 5 (Bounding the first term in (17))**.** *Suppose $\boldsymbol{S} \in \mathbb{R}^{k \times n}$ is generated from the sparse-Gaussian model $\mathcal{SG}(p)$. Then*

$$\Pr\left[\inf_{\boldsymbol{w} \in \mathcal{B}_m^\circ} \|\boldsymbol{S}^\top \boldsymbol{w}\|_1 < n(s/k)(\sqrt{2/\pi} - \epsilon) - \delta\sqrt{k} n(s/k)(\sqrt{2/\pi} + \epsilon)\right] \leq \left(\left(\frac{2}{\delta}\right)^k + 2\right) 2\exp\left(-\frac{n(s/k)^2\epsilon^2}{2 + \sqrt{2}\epsilon}\right),$$

where $\delta \in (0, 1)$ represents any choice of δ-cover for $\mathcal{B}_m^\circ$.

Proof. Following Lemma 4, we have

$$\Pr\left[\inf_{\boldsymbol{w} \in \mathcal{B}_m^\circ} \|\boldsymbol{S}^\top \boldsymbol{w}\|_1 < \beta - \gamma\delta\sqrt{k}\right] \leq \sum_{\boldsymbol{w}_i \in \mathcal{N}(\mathcal{B}_m^\circ, \delta)} \Pr\left[\|\boldsymbol{S}^\top \boldsymbol{w}_i\|_1 < \beta\right] + \Pr\left[\|\boldsymbol{S}^\top\|_1 > \gamma\right], \tag{22}$$

where $|\mathcal{N}(\mathcal{B}_m^\circ, \delta)| < (1 + 2/\delta)^k$ according to Lemma 3.

The bound to the first term in (22) is almost identical to Lemma 2. Dropping the subscript of $\boldsymbol{w}_i$, we write

$$\|\boldsymbol{S}^\top \boldsymbol{w}\|_1 = \sum_{i=1}^n \left|\sum_{j=1}^k s_{ij} w_j\right| := \sum_{i=1}^n |Z_i|.$$

Without the absolute value, Z_i is normally distributed with zero mean and variance $\sigma^2 = \sum_{j\in\mathcal{I}_i} w_j^2$, where $\mathcal{I}_i$ is the index set of nonzero elements in $\boldsymbol{s}_i$ with $|\mathcal{I}_i| = s < m$. Then $\sigma^2 \le 1$. To see this, suppose without loss of generality that $\mathcal{I}_i = \{1,\dots,s\}$ and let $\boldsymbol{Q} = [\boldsymbol{I}\ 0]^\top$, then $\boldsymbol{w} \in \mathcal{B}_m^\circ$ implies that $\sum_{j\in\mathcal{I}_i} w_j^2 = \|\boldsymbol{Q}\boldsymbol{w}\|_2 \le 1$. In other words, if $\boldsymbol{w} \in \mathcal{B}_m^\circ$, the squared sum of no more than m elements of $\boldsymbol{w}$ must be ≤ 1. Therefore for $d \ge 2$

$$\mathrm{E}[|Z_i|^d] \le \mathrm{E}[|Z_i|^2] \le 1,$$

while for $d = 1$ we have

$$\mathrm{E}[|Z_i|] = \frac{s}{k}\sqrt{\frac{2}{\pi}}.$$

Again, the moments satisfy (18) with $c = 1/\sqrt{2}$, results in

$$\begin{aligned}
\Pr\left[\|\boldsymbol{S}^\top\boldsymbol{w}\|_1 < n(s/k)(\sqrt{2/\pi}-\epsilon)\right] &\le \Pr\left[\left|\sum_{i=1}^n (Z_i - \mathrm{E}[Z_i])\right| > n(s/k)\epsilon\right] \\
&\le 2\exp\left(-\frac{n^2(s/k)^2\epsilon^2}{2(n+n(s/k)\epsilon/\sqrt{2})}\right) \\
&= 2\exp\left(-\frac{n(s/k)^2\epsilon^2}{2+\sqrt{2}\epsilon}\right). \qquad (23)
\end{aligned}$$

To bound $\|\boldsymbol{S}^\top\|_1$, we recall that this is the ℓ_1 induced norm for matrix $\boldsymbol{S}^\top$, which is shown to be the maximum of the ℓ_1 norms of the columns of $\boldsymbol{S}^\top$. This means we can use similar arguments used in Lemma 2 (but applied to the other direction) to have

$$\begin{aligned}
\Pr\left[\|\boldsymbol{S}^\top\|_1 > np(\sqrt{2/\pi}+\epsilon)\right] &= \Pr\left[\max_j \|\boldsymbol{S}_{j,:}\|_1 > n(s/k)(\sqrt{2/\pi}+\epsilon)\right] \\
&\le \Pr\left[\|\boldsymbol{S}_{j,:}\|_1 > n(s/k)(\sqrt{2/\pi}+\epsilon)\right] \\
&\le \Pr\left[\left|\sum_{i=1}^n (Z_i - \mathrm{E}[Z_i])\right| > n(s/k)\epsilon\right] \\
&\le 2\exp\left(-\frac{n(s/k)\epsilon^2}{2+\sqrt{2}\epsilon}\right) \le 2\exp\left(-\frac{n(s/k)^2\epsilon^2}{2+\sqrt{2}\epsilon}\right), \qquad (24)
\end{aligned}$$

where we pick an arbitrary $j \in [k]$ in the second line since this event implies that the maximum ℓ_1 norm of the rows is lowerbounded, and in the third line each Z_i satisfies (20). The proof is complete by combining (22), (23), and (24) with $\beta = n(s/k)(\sqrt{2/\pi}-\epsilon)$ and $\gamma = n(s/k)(\sqrt{2/\pi}+\epsilon)$. □

Proof of Theorem 2. We first instantiate Lemma 5 with

$$\delta = \frac{n^2(s/k)^2(\sqrt{2/\pi}-\epsilon)^2 - 1}{n^2(s/k)^2(\sqrt{2/\pi}-\epsilon)(\sqrt{2/\pi}+\epsilon)\sqrt{k}},$$

which satisfies $\delta > 0$ if

$$\epsilon < \sqrt{\frac{2}{\pi}} - \frac{1}{n(s/k)}.$$

Then we have

$$\Pr\left[\inf_{\boldsymbol{w}\in\mathcal{B}_m^\circ} \|\boldsymbol{S}^\top\boldsymbol{w}\|_1 < 1/n(s/k)^2(\sqrt{2/\pi}-\epsilon)\right] \le \left(\left(\frac{2}{\delta}\right)^k + 2\right) 2\exp\left(-\frac{n(s/k)\epsilon^2}{2+\sqrt{2}\epsilon}\right),$$

Combining (17), (21), and Lemma 2 with $\alpha = n(s/k)(\sqrt{2/\pi}-\epsilon)$, we obtain

$$\begin{aligned}
\Pr\left[\sup_{\|\boldsymbol{w}^\top\tilde{\boldsymbol{S}}\|_1\le 1} \|\boldsymbol{w}\| > 1\right] &\le \Pr\left[\inf_{\|\boldsymbol{w}\|=1} \|\boldsymbol{S}^\top\boldsymbol{w}\|_1 < 1/\alpha\right] + \Pr\left[\max_j \|\boldsymbol{S}_{j,:}\|_1 < \alpha\right] \\
&\le 2\left(k + \left(\frac{2}{\delta}\right)^k + 2\right)\exp\left(-\frac{n(s/k)^2\epsilon^2}{2+\sqrt{2}\epsilon}\right). \qquad (25)
\end{aligned}$$

Further, with

$$\epsilon < \frac{((m/k)^{1/4} - 1)\sqrt{2/\pi}}{(m/k)^{1/4} + 1},$$

where the right hand side is obviously positive, we have

$$\delta = \frac{n^2(s/k)^2(\sqrt{2/\pi} - \epsilon)^2 - 1}{n^2(s/k)^2(\sqrt{2/\pi} - \epsilon)(\sqrt{2/\pi} + \epsilon)\sqrt{k}} > \frac{n^2(s/k)^2(\sqrt{2/\pi} - \epsilon)^2}{n^2(s/k)^2(\sqrt{2/\pi} + \epsilon)^2\sqrt{k}} > \frac{\sqrt{m}}{k}.$$

We can further relax (25) to

$$\begin{aligned}
\Pr\left[\sup_{\boldsymbol{w} \in \mathcal{B}_m^\circ} \|\boldsymbol{w}\| > 1\right] &\leq 2\left(k + \left(2k/\sqrt{m}\right)^k + 2\right)\exp\left(-\frac{n(s/k)^2\epsilon^2}{2 + \sqrt{2}\epsilon}\right) \\
&\leq 4\left(2k/\sqrt{m}\right)^k \exp\left(-\frac{n(s/k)^2\epsilon^2}{2 + \sqrt{2}\epsilon}\right) \\
&\leq 4\exp\left((k/2)\log(k^2/m) - n(s/k)^2(m/k)\right),
\end{aligned}$$

where in the last inequality we simply picked a small enough ϵ so that

$$\frac{\epsilon^2}{2 + \sqrt{2}\epsilon} < \frac{m}{k}.$$

This completes the proof. □