Journal of Machine Learning Research 18 (2018) 1-56 Submitted 3/16; Revised 7/17; Published 4/18

Local Identifiability of ℓ_1-minimization Dictionary Learning: a Sufficient and Almost Necessary Condition

Siqi Wu SIQI@STAT.BERKELEY.EDU

Bin Yu * BINYU@BERKELEY.EDU

Department of Statistics
University of California
Berkeley, CA 94720-1776, USA

Editor: Hui Zou

Abstract

We study the theoretical properties of learning a dictionary from N signals $\mathbf{x}_i \in \mathbb{R}^K$ for $i = 1, \ldots, N$ via ℓ_1-minimization. We assume that $\mathbf{x}_i$'s are *i.i.d.* random linear combinations of the K columns from a complete (i.e., square and invertible) reference dictionary $\mathbf{D}_0 \in \mathbb{R}^{K \times K}$. Here, the random linear coefficients are generated from either the s-sparse Gaussian model or the Bernoulli-Gaussian model. First, for the population case, we establish a sufficient and almost necessary condition for the reference dictionary $\mathbf{D}_0$ to be locally identifiable, i.e., a strict local minimum of the expected ℓ_1-norm objective function. Our condition covers both sparse and dense cases of the random linear coefficients and significantly improves the sufficient condition by Gribonval and Schnass (2010). In addition, we show that for a complete μ-coherent reference dictionary, i.e., a dictionary with absolute pairwise column inner-product at most $\mu \in [0, 1)$, local identifiability holds even when the random linear coefficient vector has up to $O(\mu^{-2})$ nonzero entries. Moreover, our local identifiability results also translate to the finite sample case with high probability provided that the number of signals N scales as $O(K \log K)$.

Keywords: dictionary learning, ℓ_1-minimization, local minimum, non-convex optimization, sparse decomposition

1. Introduction

Expressing signals as sparse linear combinations of a dictionary basis has enjoyed great success in applications ranging from image denoising to audio compression. Given a known dictionary matrix $\mathbf{D} \in \mathbb{R}^{d \times K}$ with K columns or atoms, one popular method to recover the sparse coefficients $\boldsymbol{\alpha} \in \mathbb{R}^K$ of a signal $\mathbf{x} \in \mathbb{R}^d$ is through solving the convex ℓ_1-minimization problem

$$\text{minimize } \|\boldsymbol{\alpha}\|_1 \text{ subject to } \mathbf{x} = \mathbf{D}\boldsymbol{\alpha}.$$

This approach, known as *basis pursuit* (Chen et al., 1998), along with many of its variants, has been studied extensively in statistics and signal processing communities. See e.g., Donoho and Elad (2003); Fuchs (2004); Candes and Tao (2005).

For certain data types such as natural image patches, predefined dictionaries like the wavelets (Mallat, 2008) are usually available. However, for a less-known data type, a new

∗. Also in the Department of Electrical Engineering & Computer Science.

dictionary has to be designed to effectively represent the data. Dictionary learning, or sparse coding, learns adaptively a dictionary from a set of training signals such that they have sparse representations under this dictionary (Olshausen and Field, 1997). One formulation of dictionary learning involves solving a non-convex ℓ_1-minimization problem (Zibulevsky et al., 2001; Plumbley, 2007; Gribonval and Schnass, 2010; Geng et al., 2011). Concretely, define

$$l(\mathbf{x}, \mathbf{D}) = \min_{\boldsymbol{\alpha} \in \mathbb{R}^K} \{\|\boldsymbol{\alpha}\|_1, \text{ subject to } \mathbf{x} = \mathbf{D}\boldsymbol{\alpha}\}. \tag{1}$$

We learn a dictionary from the N signals $\mathbf{x}_i \in \mathbb{R}^d$ for $i = 1, \ldots, N$ by solving

$$\min_{\mathbf{D} \in \mathcal{D}} \ L_N(\mathbf{D}) = \min_{\mathbf{D} \in \mathcal{D}} \frac{1}{N} \sum_{i=1}^{N} l(\mathbf{x}_i, \mathbf{D}). \tag{2}$$

Here, $\mathcal{D} \subset \mathbb{R}^{d \times K}$ is a constraint set for candidate dictionaries. In many signal processing tasks, learning an adaptive dictionary via the optimization problem (2) and its variants is empirically demonstrated to have superior performance over fixed standard dictionaries (Elad and Aharon, 2006; Peyré, 2009; Grosse et al., 2012). For a review of dictionary learning algorithms and applications, see Elad (2010); Rubinstein et al. (2010); Mairal et al. (2014).

Apart from the empirical success of many dictionary learning formulations, recently there is a growing body of work on the theory of dictionary learning. One line of research treats the problem of *dictionary identifiability*: if the signals are generated using a dictionary $\mathbf{D}_0$ referred to as the *reference dictionary*, under what conditions can we recover $\mathbf{D}_0$ by solving the dictionary learning problem? Being able to identify the reference dictionary is important when there is a need to interpret the learned dictionary, see for example Wu et al. (2016). Let $\boldsymbol{\alpha}_i \in \mathbb{R}^K$ for $i = 1, \ldots, N$ be some random vectors. A popular signal generation model assumes that a signal vector can be expressed as a linear combination of the columns of the reference dictionary: $\mathbf{x}_i \approx \mathbf{D}_0\boldsymbol{\alpha}_i$ (Gribonval and Schnass, 2010; Geng et al., 2011; Gribonval et al., 2015). In this paper, we will study the problem of *local identifiability* of ℓ_1-minimization dictionary learning (2) under this generating model.

Local identifiability. A reference dictionary $\mathbf{D}_0$ is said to be *locally identifiable* with respect to an objective function $L(\mathbf{D})$ if $\mathbf{D}_0$ is one of the strict local minima of L. The pioneer work of Gribonval and Schnass (2010) (referred to as GS henceforth) analyzed the ℓ_1-minimization problem (2) for noiseless signals ($\mathbf{x}_i = \mathbf{D}_0\boldsymbol{\alpha}_i$) and complete ($K = d$ and full rank) dictionaries. Under the sparse Bernoulli-Gaussian model for the linear coefficients $\boldsymbol{\alpha}_i$'s, they showed that for a sufficiently incoherent reference dictionary $\mathbf{D}_0$, $N = O(K \log K)$ samples can guarantee local identifiability with respect to $L_N(\mathbf{D})$ in (2) with high probability. Still in the noiseless setting, Geng et al. (2011) extended the analysis to over-complete ($K > d$) dictionaries. More recently under the noisy linear generative model with possible outliers, Gribonval et al. (2015) established local identifiability results for (2) with $l(\mathbf{x}, \mathbf{D})$ replaced by the LASSO objective function of Tibshirani (1996). Other related works on local identifiability include Schnass (2014) and Schnass (2015), who gave respectively sufficient conditions for the local correctness of the K-SVD (Aharon et al., 2006b) algorithm and a maximum response formulation of dictionary learning.

Contributions. There has not been much work on necessary conditions for local dictionary identifiability. Numerical experiments demonstrate that local identifiability undergoes a phase transition (Figure 1; see also Figure 3 of GS). The bound implied by the sufficient condition in GS falls well below the empirical phase boundary. Thus, even though theoretical results for the more general scenarios are available, we adopt the noiseless signals and complete dictionary setting of GS in order to find better local identifiability conditions. We summarize our main contributions below:

- For the population case where $N = \infty$, we establish a sufficient and almost necessary condition for local identifiability under both the s-sparse Gaussian and the Bernoulli-Gaussian models. For the Bernoulli-Gaussian model, the phase boundary implied by our condition significantly improves the GS bound and agrees well with the empirical phase boundary (Figure 1).

- We provide lower and upper bounds to approximate the quantities in our local identifiability condition, as it generally requires to solve a series of second-order cone programs to compute those quantities.

- As a consequence, we show that a μ-coherent reference dictionary—a dictionary with absolute pairwise column inner-product at most $\mu \in [0, 1)$—is locally identifiable for sparsity level, measured by the average number of nonzeros in the random linear coefficient vectors, up to the order $O(\mu^{-2})$. Moreover, if the sparsity level is greater than $O(\mu^{-2})$, the reference dictionary is generally not locally identifiable. In comparison, the sufficient condition by GS demands the number of dictionary atoms $K = O(\mu^{-2})$, which is a much more stringent requirement. For over-complete dictionaries, Geng et al. (2011) requires the sparsity level to be of the order $O(\mu^{-1})$. It should also be noted that Schnass (2015) establishes the bound $O(\mu^{-2})$ for *approximate* local identifiability under a new response maximization formulation of dictionary learning. To the best of our knowledge, our result is the first in showing that $O(\mu^{-2})$ is achievable and optimal for *exact* local recovery under the ℓ_1-minimization criterion.

- We also extend our identifiability results to the finite sample case. We show that for a fixed sparsity level, we need $N = O(K \log K)$ *i.i.d.* signals to determine whether or not the reference dictionary can be identified locally. This sample requirement is the same as GS's and is the best known sample requirement among all previous studies on local identifiability.

Other related works. Apart from analyzing the local minima of dictionary learning, another line of research aims at designing provable algorithms for recovering the reference dictionary. Georgiev et al. (2005) and Aharon et al. (2006a) proposed combinatorial algorithms and gave deterministic conditions for dictionary recovery which require sample size N to be exponentially large in the number of dictionary atoms K. Spielman et al. (2012) established exact global recovery results for complete dictionaries through efficient convex programs. Agarwal et al. (2014c) and Arora et al. (2014) proposed clustering-based methods to estimate the reference dictionary in the over-complete setting. Agarwal et al. (2014a) and Arora et al. (2015) provided theoretical guarantees for their alternating minimization algorithms. Sun et al. (2017) proposed a non-convex optimization algorithm that provably

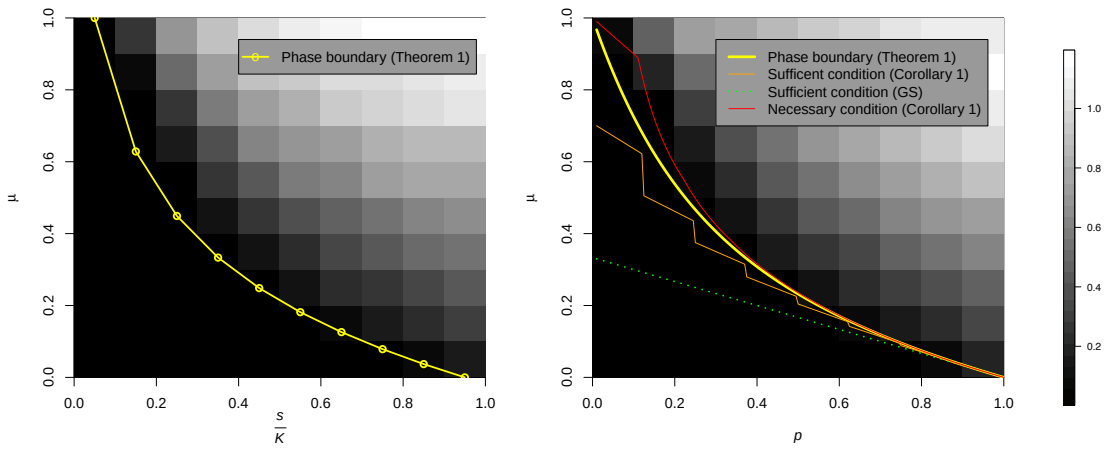

Figure 1: Local recovery errors for the s-sparse Gaussian model (Left) and the Bernoulli(p)-Gaussian model (Right). Under the s-sparse Gaussian model, the parameter $s \in \{1, \ldots, K\}$ is the number of nonzeros in each linear coefficient vector. Under the Bernoulli(p)-Gaussian model, $p \in (0, 1]$ is the probability of an entry of the linear coefficient vector being nonzero. The data are generated with the reference dictionary $\mathbf{D}_0 \in \mathbb{R}^{10\times 10}$ (i.e., $K = 10$) satisfying $\mathbf{D}_0^T\mathbf{D}_0 = \mu\mathbf{1}\mathbf{1}^T + (1-\mu)\mathbf{I}$ for $\mu \in [0, 1)$, see Example 5 for details. For each $(\mu, \frac{s}{K})$ or (μ, p) tuple, ten batches of $N = 2000$ signals $\{\mathbf{x}_i\}_{i=1}^{2000}$ are generated according to the noiseless linear model $\mathbf{x}_i = \mathbf{D}_0\boldsymbol{\alpha}_i$, with $\{\boldsymbol{\alpha}_i\}_{i=1}^{2000}$ drawn $i.i.d.$ from the s-sparse Gaussian model or $i.i.d.$ from the Bernoulli(p)-Gaussian model. For each batch, the dictionary is estimated through an alternating minimization algorithm in the `SPAMS` package (Mairal et al., 2010), with initial dictionary set to be $\mathbf{D}_0$. The grayscale intensity in the figure corresponds to the Frobenius error of the difference between the estimated dictionary and the reference dictionary $\mathbf{D}_0$, averaged for the ten batches. The "phase boundary" curve corresponds to the theoretical boundary that separates the region of local identifiability (below the curve) and the region of local non-identifiability (above the curve) according to Theorem 1 of this paper. The "Sufficient condition (Corollary 1)" and "Necessary condition (Corollary 1)" curves are the lower and upper bounds given by Corollary 1 to approximate the exact phase boundary. Finally, the "Sufficient condition (GS)" curve corresponds to the lower bound by GS. Note that for the s-sparse Gaussian model, the "Sufficient condition (Corollary 1)" and "Necessary condition (Corollary 1)" curves coincide with the phase boundary. See also Appendix Figures B.1 and B.2 for additional simulation results with $K = 20$ and $K = 50$.

recovers a complete reference dictionary for sparsity level up to $O(K)$. While in this paper we do not provide an algorithm, our identifiability conditions suggest theoretical limits of dictionary recovery for all algorithms attempting to solve the optimization problem (2). In particular, in the regime where the reference dictionary is not identifiable, no algorithm can simultaneously solve (2) and recover the ground truth reference dictionary.

Other related works include generalization bounds for signal reconstruction errors under the learned dictionary (Maurer and Pontil, 2010; Vainsencher et al., 2011; Mehta and Gray, 2013; Gribonval et al., 2013), dictionary identifiability through combinatorial matrix theory (Hillar and Sommer, 2015), as well as algorithms and theories for the closely related independent component analysis (Comon, 1994; Arora et al., 2012b) and nonnegative matrix factorization (Arora et al., 2012a; Recht et al., 2012).

The rest of the paper is organized as follows: In Section 2, we give basic assumptions and describe the two probabilistic models for signal generation. Section 3 develops sufficient and almost necessary local identifiability conditions under both models for the population problem, and establishes lower and upper approximating bounds. In Section 4, we will present local identifiability results for the finite sample problem. Detailed proofs for the theoretical results can be found in the Appendix.

2. Preliminaries

In this section, we will introduce notations and basic assumptions for our analysis.

2.1 Notations

For a positive integer m, define $[\![m]\!]$ to be the set of the first m positive integers, $\{1, \ldots, m\}$. The notation $\mathbf{x}[i]$ denotes the i-th entry of the vector $\mathbf{x} \in \mathbb{R}^m$. For a non-empty index set $S \subset [\![m]\!]$, we denote by $|S|$ the set cardinality and $\mathbf{x}[S] \in \mathbb{R}^{|S|}$ the sub-vector indexed by S. We define $\mathbf{x}[-j] \coloneqq (\mathbf{x}[1], \ldots, \mathbf{x}[j-1], \mathbf{x}[j+1], \ldots, \mathbf{x}[m]) \in \mathbb{R}^{m-1}$ to be the vector $\mathbf{x}$ without its j-th entry.

For a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, we denote by $\mathbf{A}[i, j]$ its (i, j)-th entry. For non-empty sets $S \subset [\![m]\!]$ and $T \subset [\![n]\!]$, denote by $\mathbf{A}[S, T]$ the sub-matrix of $\mathbf{A}$ with the rows indexed by S and columns indexed by T. Denote by $\mathbf{A}[i,]$ and $\mathbf{A}[, j]$ the i-th row and the j-th column of $\mathbf{A}$ respectively. Similar to the vector case, the notation $\mathbf{A}[-i, j] \in \mathbb{R}^{m-1}$ denotes the j-th column of $\mathbf{A}$ without its i-th entry.

For $p \geq 1$, the ℓ_p-norm of a vector $\mathbf{x} \in \mathbb{R}^m$ is defined as $\|\mathbf{x}\|_p = (\sum_{i=1}^m |\mathbf{x}[i]|^p)^{1/p}$, with the convention that $\|\mathbf{x}\|_0 = |\{i : \mathbf{x}[i] \neq 0\}|$ and $\|\mathbf{x}\|_\infty = \max_i |\mathbf{x}[i]|$. For any norm $\|.\|$ on $\mathbb{R}^m$, the dual norm of $\|.\|$ is defined as $\|\mathbf{x}\|^* = \sup_{\mathbf{y} \neq 0} \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{y}\|}$.

For two sequences of real numbers $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, we denote by $a_n = O(b_n)$ if there is a constant $C > 0$ such that $a_n \leq C b_n$ for all $n \geq 1$. For $a \in \mathbb{R}$, denote by $\lfloor a \rfloor$ the integer part of a and $\lceil a \rceil$ the smallest integer greater than or equal to a. Throughout this paper, we shall agree that $\frac{0}{0} = 0$.

2.2 Basic Assumptions

We denote by $\mathcal{D} \subset \mathbb{R}^{d \times K}$ the constraint set of dictionaries for the optimization problem (2). In this paper, we consider square dictionaries, i.e., $d = K \geq 2$. As in GS, we choose $\mathcal{D}$ to

be the *oblique manifold* (Absil et al., 2008):

$$\mathcal{D} = \left\{ \mathbf{D} \in \mathbb{R}^{K \times K} : \ \|\mathbf{D}[,k]\|_2 = 1 \text{ for all } k = 1, \ldots, K \right\}.$$

We also call a dictionary column $\mathbf{D}[,k]$ an *atom* of the dictionary. Denote by $\mathbf{D}_0 \in \mathcal{D}$ the *reference dictionary*, i.e., the ground truth dictionary that generates the signals. With these notations, we now give a formal definition for local identifiability:

Definition 1 *(Local identifiability) Let $L(\mathbf{D}) : \mathcal{D} \to \mathbb{R}$ be an objective function. We say that the reference dictionary $\mathbf{D}_0$ is locally identifiable with respect to $L(\mathbf{D})$ if $\mathbf{D}_0$ is a strict local minimum of $L(\mathbf{D})$.*

Sign-permutation ambiguity. As noted by previous works GS and Geng et al. (2011), the ℓ_1-norm objective function $L(\mathbf{D}) = L_N(\mathbf{D})$ of (2) has an intrinsic sign-permutation ambiguity. Let $\mathbf{D}' = \mathbf{D}\mathbf{P}\mathbf{\Lambda}$ for some permutation matrix $\mathbf{P}$ and diagonal matrix $\mathbf{\Lambda}$ with ± 1 diagonal entries. It is easy to see that $\mathbf{D}'$ and $\mathbf{D}$ have the same objective value. Thus, the objective function $L_N(\mathbf{D})$ has at least $2^K K!$ local minima. We can only recover $\mathbf{D}_0$ up to column permutation and sign changes.

Note that if the dictionary atoms are linearly dependent, the effective dimension is strictly less than K and the problem essentially becomes over-complete. Since dealing with over-complete dictionaries is beyond the scope of this paper, we make the following assumption:

Assumption I *(Complete dictionaries). The reference dictionary $\mathbf{D}_0 \in \mathcal{D} \subset \mathbb{R}^{K \times K}$ is full rank.*

Let $\mathbf{M}_0 = \mathbf{D}_0^T \mathbf{D}_0$ be the *dictionary atom collinearity matrix* containing the inner-products between dictionary atoms. Since each dictionary atom has unit ℓ_2-norm, $\mathbf{M}_0[i,i] = 1$ for all $i \in [\![K]\!]$. In addition, as $\mathbf{D}_0$ is full rank, $\mathbf{M}_0$ is positive definite and $|\mathbf{M}_0[i,j]| < 1$ for all $i \neq j$.

We assume that a signal is generated as a random linear combination of the dictionary atoms. In this paper, we consider the following two probabilistic models for the random linear coefficients:

Probabilistic models for sparse coefficients. Denote by $\mathbf{z} \in \mathbb{R}^K$ a random vector from the K-dimensional standard normal distribution.

Model 1—*$SG(s)$.* Let $\mathbf{S}$ be a size-s subset uniformly drawn from all size-s subsets of $[\![K]\!]$. Define $\boldsymbol{\xi} \in \{0,1\}^K$ by setting $\boldsymbol{\xi}[j] = I\{j \in \mathbf{S}\}$ for $j \in [\![K]\!]$, where $I\{.\}$ is the indicator function. Let $\boldsymbol{\alpha} \in \mathbb{R}^K$ be such that $\boldsymbol{\alpha}[j] = \boldsymbol{\xi}[j]\mathbf{z}[j]$. Then we say $\boldsymbol{\alpha}$ is drawn from the *s-sparse Gaussian model*, or $SG(s)$.

Model 2—*$BG(p)$.* For $j \in [\![K]\!]$, let $\boldsymbol{\xi}[j]$'s be *i.i.d.* Bernoulli random variable with success probability $p \in (0,1]$. Let $\boldsymbol{\alpha} \in \mathbb{R}^K$ be such that $\boldsymbol{\alpha}[j] = \boldsymbol{\xi}[j]\mathbf{z}[j]$. Then we say $\boldsymbol{\alpha}$ is drawn from the *Bernoulli(p)-Gaussian model*, or $BG(p)$.

With the above two models, we state the following assumption for random signal generation:

Assumption II *(Signal generation). For $i \in [\![N]\!]$, let $\boldsymbol{\alpha}_i \in \mathbb{R}^K$ be either i.i.d. s-sparse Gaussian vectors or i.i.d. Bernoulli(p)-Gaussian vectors. The signals $\mathbf{x}_i \in \mathbb{R}^K$ are generated according to the noiseless linear model:*

$$\mathbf{x}_i = \mathbf{D}_0 \boldsymbol{\alpha}_i.$$

Remarks:
(1) The above two models and their variants were studied in a number of prior theoretical works, including Gribonval and Schnass (2010); Geng et al. (2011); Agarwal et al. (2014b); Sun et al. (2017).
(2) By construction, a random vector generated from the s-sparse model has exactly s nonzero entries. The data points $\mathbf{x}_i$'s therefore lie within the union of the linear spans of s dictionary atoms (Figure 2 Left). The Bernoulli(p)-Gaussian model, on the other hand, allows the random coefficient vector to have any number of nonzero entries ranging from 0 to K with a mean of pK. As a result, some data points, called "non-sparse outliers" in GS, can be outside of any sparse linear span of the dictionary atoms (Figure 2 Right and Figure 1 of GS). We refer readers to the remarks following Example 5 in Section 3 for a discussion of the effect of non-sparse outliers on local identifiability.
(3) Gribonval et al. (2015) assumed a more general distribution for the sparse coefficients. While our local identifiability results can potentially be extended to their model, such an extension would require significantly more complicated notations and make the corresponding results less interpretable. For the sake of accessibility and interpretability, we focus only on the two probabilistic models above.

In this paper, we study the problem of dictionary identifiability with respect to the population objective function $\mathbb{E}\, L_N(\mathbf{D})$ (Section 3) and the finite sample objective function $L_N(\mathbf{D})$ (Section 4). In order to analyze these objective functions, it is convenient to define the following "group LASSO"-type norms:

Definition 2 *Let $m \geq 1$ be an integer and $\mathbf{w} \in \mathbb{R}^m$.*

1. *For $k \in [\![m]\!]$, define*
$$|||\mathbf{w}|||_k = \frac{\sum_{|S|=k} \|\mathbf{w}[S]\|_2}{\binom{m-1}{k-1}}.$$

2. *For $p \in (0,1)$, define*
$$|||\mathbf{w}|||_p = \sum_{k=0}^{m-1} \texttt{pbinom}(k; m-1, p) |||\mathbf{w}|||_{k+1},$$
where `pbinom` *is the probability mass function of the binomial distribution:*
$$\texttt{pbinom}(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Remarks:
(1) Note that the above norms $|||\mathbf{w}|||_k$ and $|||\mathbf{w}|||_p$ are in fact the expected values of $|\mathbf{w}^T \boldsymbol{\alpha}|$

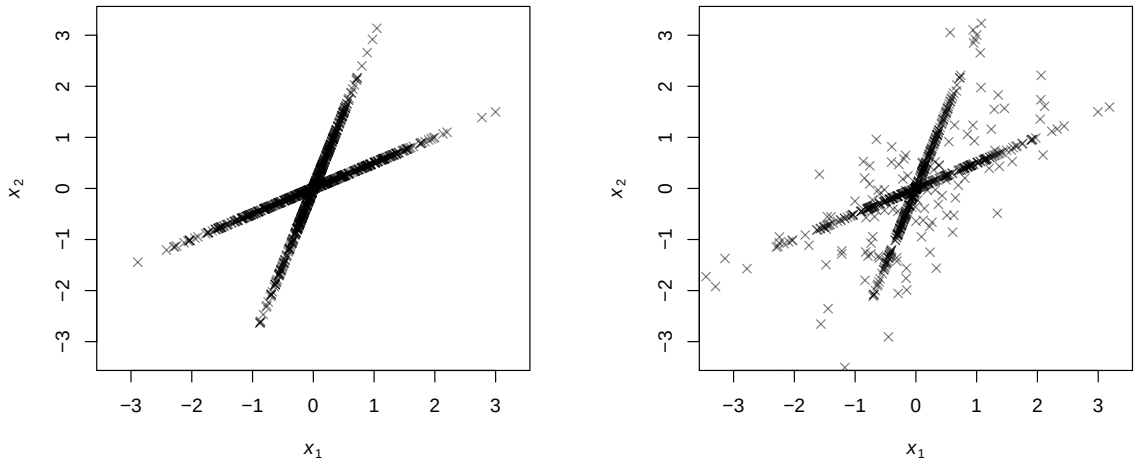

Figure 2: Data generation for $K = 2$. Left: the s-sparse Gaussian model with $s = 1$; Right: the Bernoulli(p)-Gaussian model with $p = 0.2$. The inner product between the two dictionary atoms is 0.7. A sample of $N = 1000$ data points are generated for both models. For the s-sparse model, all data points are perfectly aligned with the two lines corresponding to the two dictionary atoms. For the Bernoulli(p)-Gaussian model, a number of data points fall outside the two lines. According to our Theorem 1 and 3, despite those outliers and the high collinearity between the two atoms, the reference dictionary is still locally identifiable for $N = \infty$ and with high probability for finite samples.

with the random vector $\boldsymbol{\alpha}$ drawn from $SG(s)$ and $BG(p)$ models respectively. For invertible $\mathbf{D} \in \mathcal{D}$, it can be shown that the objective function for one signal $\mathbf{x} = \mathbf{D}_0\boldsymbol{\alpha}$ is

$$l(\mathbf{x}, \mathbf{D}) = \|\mathbf{H}\boldsymbol{\alpha}\|_1 = \sum_{j=1}^{K} |\mathbf{H}[j,]\boldsymbol{\alpha}|,$$

where $\mathbf{H} = \mathbf{D}^{-1}\mathbf{D}_0$. Thus, taking the expectation of the objective function with respect to $\mathbf{x}$, we end up with a quantity involving either $\sum_{j=1}^{K} |||\mathbf{H}[j,]|||_s$ or $\sum_{j=1}^{K} |||\mathbf{H}[j,]|||_p$. This is the motivation of defining these norms.

(2) In particular, $|||\mathbf{w}|||_1 = \|\mathbf{w}\|_1$ and $|||\mathbf{w}|||_m = \|\mathbf{w}\|_2$.

(3) The norms defined above are special cases of the group LASSO penalty by Yuan and Lin (2006). For $|||\mathbf{w}|||_k$, the summation covers all size-k subsets of $[\![m]\!]$. The normalization factor is the number of times $\mathbf{w}[i]$ appears in the numerator. Thus, $|||\mathbf{w}|||_k$ is essentially the average of the ℓ_2-norms of all size-k sub-vectors of $\mathbf{w}$. On the other hand, $|||\mathbf{w}|||_p$ is a weighted average of $|||\mathbf{w}|||_k$'s with binomial probabilities.

3. Population Analysis

In this section, we establish local identifiability results for the case where infinitely many signals are observed. Denote by $\mathbb{E}\, l(\mathbf{x}_1, \mathbf{D})$ the expectation of the objective function $l(\mathbf{x}_1, \mathbf{D})$ of (1) with respect to the random signal $\mathbf{x}_1$. By the strong law of large numbers, as the number of signals N tends to infinity, the empirical objective function $L_N(\mathbf{D}) = \frac{1}{N}\sum_{i=1}^{N} l(\mathbf{x}_i, \mathbf{D})$ converges almost surely to its population mean $\mathbb{E}\, l(\mathbf{x}_1, \mathbf{D})$ for each fixed $\mathbf{D} \in \mathcal{D}$. Therefore the population version of the optimization problem (2) is

$$\min_{\mathbf{D}\in\mathcal{D}} \mathbb{E}\, l(\mathbf{x}_1, \mathbf{D}) \tag{3}$$

Since by assumption the reference dictionary $\mathbf{D}_0$ is full rank, we only need to work with $\mathbf{D} \in \mathcal{D}$ that is also full rank. Indeed, if the linear span of columns $\text{span}(\mathbf{D}) \neq \mathbb{R}^K$, then $\mathbf{D}_0\boldsymbol{\alpha}_1 \notin \text{span}(\mathbf{D})$ with nonzero probability. Thus $\mathbf{D}$ is infeasible with nonzero probability and so $\mathbb{E}\, l(\mathbf{x}_1, \mathbf{D}) = +\infty$. For a full rank dictionary $\mathbf{D}$, the following lemma gives the closed-form expression for the expected objective function $\mathbb{E}\, l(\mathbf{x}_1, \mathbf{D})$.

Lemma 1 *(Closed-form objective functions) Let $\mathbf{D}$ be a full rank dictionary in $\mathcal{D}$ and $\mathbf{x}_1 = \mathbf{D}_0\boldsymbol{\alpha}_1$, where $\boldsymbol{\alpha}_1 \in \mathbb{R}^K$ is a random vector. For notational convenience, let $\mathbf{H} = \mathbf{D}^{-1}\mathbf{D}_0$.*

1. *If $\boldsymbol{\alpha}_1$ is generated according to the $SG(s)$ model with $s \in [\![K-1]\!]$,*

$$L_{SG(s)}(\mathbf{D}) \coloneqq \mathbb{E}\, l(\mathbf{x}_1, \mathbf{D}) = \sqrt{\frac{2}{\pi}}\frac{s}{K}\sum_{j=1}^{K} |||\mathbf{H}[j,]|||_s. \tag{4}$$

2. *If $\boldsymbol{\alpha}_1$ is generated according to the $BG(p)$ model with $p \in (0,1)$,*

$$L_{BG(p)}(\mathbf{D}) \coloneqq \mathbb{E}\, l(\mathbf{x}_1, \mathbf{D}) = \sqrt{\frac{2}{\pi}}p\sum_{j=1}^{K} |||\mathbf{H}[j,]|||_p. \tag{5}$$

For the non-sparse cases where $s = K$ and $p = 1$, we have

$$L_{SG(s)}(\mathbf{D}) = L_{BG(p)}(\mathbf{D}) = \sqrt{\frac{2}{\pi}}\sum_{j=1}^{K} \|\mathbf{H}[j,]\|_2.$$

Remarks: It can be seen from the above closed-form expressions that the two models are closely related. First of all, it is natural to identify p with $\frac{s}{K}$, the fraction of expected number of nonzero entries in $\boldsymbol{\alpha}_1$. Next, by definition, $|||.|||_p$ is a binomial average of $|||.|||_k$. Therefore, the Bernoulli-Gaussian objective function $L_{BG(p)}(\mathbf{D})$ can be treated as a binomial average of the s-sparse objective function $L_{SG(s)}(\mathbf{D})$.

By analyzing the above closed-form expressions of the ℓ_1-norm objective function, we establish the following sufficient and almost necessary conditions for population local identifiability:

Theorem 1 *(Population local identifiability) Recall that* $\mathbf{M}_0 = \mathbf{D}_0^T\mathbf{D}_0$ *and* $\mathbf{M}_0[-j,j]$ *denotes the* j*-th column of the off-diagonal part of* $\mathbf{M}_0$*. Let* $|||.|||_s^*$ *and* $|||.|||_p^*$ *be the dual norm of* $|||.|||_s$ *and* $|||.|||_p$ *respectively.*

1. *(*$SG(s)$ *models) For* $K \geq 2$ *and* $s \in [\![K-1]\!]$*, if*

$$\max_{j\in[\![K]\!]} |||\mathbf{M}_0[-j,j]|||_s^* < 1 - \frac{s-1}{K-1}.$$

   then $\mathbf{D}_0$ *is locally identifiable with respect to* $L_{SG(s)}$*.*

2. *(*$BG(p)$ *models) For* $K \geq 2$ *and* $p \in (0,1)$*, if*

$$\max_{j\in[\![K]\!]} |||\mathbf{M}_0[-j,j]|||_p^* < 1 - p.$$

   then $\mathbf{D}_0$ *is locally identifiable with respect to* $L_{BG(p)}$*.*

Moreover, the above conditions are almost necessary in the sense that if the reversed strict inequalities hold, then $\mathbf{D}_0$ *is not locally identifiable.*

On the other hand, if $s = K$ *or* $p = 1$*, then* $\mathbf{D}_0$ *is not locally identifiable with respect to* $L_{SG(s)}$ *or* $L_{BG(p)}$*.*

Proof sketch. Let $\{\mathbf{D}_t\}_{t\in\mathbb{R}}$ be a collection of dictionaries $\mathbf{D}_t \in \mathcal{D}$ indexed by $t \in \mathbb{R}$ and $L(\mathbf{D}) = \mathbb{E}\, l(\mathbf{x}_1, \mathbf{D})$ be the population objective function. The reference dictionary $\mathbf{D}_0$ is a strict local minimum of $L(\mathbf{D})$ on the manifold $\mathcal{D}$ if and only if the following statement holds: for any $\{\mathbf{D}_t\}_{t\in\mathbb{R}}$ that is a smooth function of t with non-vanishing derivative at $t = 0$, $L(\mathbf{D}_t)$ has a strict local minimum at $t = 0$. For a fixed $\{\mathbf{D}_t\}_{t\in\mathbb{R}}$, to ensure that $L(\mathbf{D}_t)$ achieves a strict local minimum at $t = 0$, it suffices to have the following one-sided derivative inequalities:

$$\lim_{t\downarrow 0^+} \frac{L(\mathbf{D}_t) - L(\mathbf{D}_0)}{t} > 0 \text{ and } \lim_{t\uparrow 0^-} \frac{L(\mathbf{D}_t) - L(\mathbf{D}_0)}{t} < 0.$$

It can be shown that the above inequalities are equivalent to:

$$\max_{j\in[\![K]\!]} \left|\mathbf{M}_0[-j,j]^T\mathbf{w}\right| < \begin{cases} 1 - \frac{s-1}{K-1} & \text{for } SG(s) \\ 1 - p & \text{for } BG(p) \end{cases}$$

where $\mathbf{w} \in \mathbb{R}^{K-1}$ is a unit vector under norm $|||.|||_s$ or $|||.|||_p$ and corresponds to the direction in which $\mathbf{D}_t$ approaches $\mathbf{D}_0$ as t tends to zero. Since $t = 0$ has to be a strict local minimum for all smooth $\{\mathbf{D}_t\}_{t\in\mathbb{R}}$ or approaching directions, by taking the supremum over all such unit vectors the LHS of the above inequality becomes the dual norm of $|||.|||_s$ or $|||.|||_p$. On the other hand, $\mathbf{D}_0$ is not a local minimum if $\lim_{t\downarrow 0^+}(L(\mathbf{D}_t) - L(\mathbf{D}_0))/t < 0$ or $\lim_{t\uparrow 0^-}(L(\mathbf{D}_t) - L(\mathbf{D}_0))/t > 0$ for some $\{\mathbf{D}_t\}_{t\in\mathbb{R}}$. Thus the sufficient condition is also almost necessary. We refer readers to Section A.1.2 for the detailed proof.

Local identifiability phase boundary. Theorem 1 indicates that population local identifiability undergoes a phase transition. The following equations

$$\max_{j\in[\![K]\!]} |||\mathbf{M}_0[-j,j]|||_s^* = 1 - \frac{s-1}{K-1} \text{ and } \max_{j\in[\![K]\!]} |||\mathbf{M}_0[-j,j]|||_p^* = 1 - p \tag{6}$$

define the phase boundaries which separate the regions of local identifiability and non-identifiability under respective models. It is unclear whether $\mathbf{D}_0$ is locally identifiable on the phase boundary. If either equality in (6) holds, the directional derivative of the objective function at $\mathbf{D}_0$ become zero in certain directions. Hence analyzing local identifiability in this case requires higher order derivative computations that quickly become complicated.

Collinearity and sparsity. These are the two factors that determine local identifiability. Intuitively, for $\mathbf{D}_0$ to be locally identifiable, neither can the atoms of $\mathbf{D}_0$ be too linearly dependent, nor can the random linear coefficients be too dense. For the s-sparse Gaussian model, the quantity $\max_{j\in[\![K]\!]} |||\mathbf{M}_0[-j,j]|||_s^*$ measures the size of the off-diagonal entries of $\mathbf{M}_0$ and hence the collinearity of the dictionary atoms. In addition, that quantity depends on the sparsity parameter s. By Lemma 7 in the Appendix, $\max_{j\in[\![K]\!]} |||\mathbf{M}_0[-j,j]|||_s^*$ is increasing with respect to s. Similar conclusion holds for the Bernoulli-Gaussian model. Therefore, sparser linear coefficients will lead to less restrictive requirement on dictionary atom collinearity. See, for example, the phase boundaries in Figure 1.

Next, we will present a few examples to gain more intuition for the local identifiability conditions.

Example 1 *(1-sparse Gaussian model) A full rank $\mathbf{D}_0$ is always locally identifiable at the population level under a 1-sparse Gaussian model. Indeed, by Corollary 7 in the Appendix, $|||\mathbf{M}_0[-j,j]|||_1^* = \max_{i\neq j} |\mathbf{M}_0[i,j]| < 1$ for all $j \in [\![K]\!]$. Thus, a full rank dictionary $\mathbf{D}_0$ always satisfies the sufficient condition.*

Example 2 *($(K-1)$-sparse Gaussian model) For $j \in [\![K]\!]$, by Corollary 7,*

$$|||\mathbf{M}_0[-j,j]|||_{K-1}^* = \|\mathbf{M}_0[-j,j]\|_2.$$

Therefore the phase boundary under the $(K-1)$-sparse model is

$$\max_{j\in[\![K]\!]} \|\mathbf{M}_0[-j,j]\|_2 = \frac{1}{K-1}.$$

Example 3 *(Orthogonal dictionaries) If $\mathbf{M}_0 = \mathbf{I}$, then*

$$\max_{j\in[\![K]\!]} |||\mathbf{M}_0[-j,j]|||_s^* = \max_{j\in[\![K]\!]} |||\mathbf{M}_0[-j,j]|||_p^* = 0.$$

Therefore orthogonal dictionaries are always locally identifiable if $s < K$ or $p < 1$.

Example 4 *(Minimally dependent dictionary atoms) Let $\mu \in (-1,1)$. Consider a dictionary atom collinearity matrix $\mathbf{M}_0$ such that $\mathbf{M}_0[1,2] = \mathbf{M}_0[2,1] = \mu$ and $\mathbf{M}_0[i,j] = 0$ for all other $i \neq j$. By Corollary 8,*

$$\max_{j\in[\![K]\!]} |||\mathbf{M}_0[-j,j]|||_s^* = \max_{j\in[\![K]\!]} |||\mathbf{M}_0[-j,j]|||_p^* = |\mu|.$$

Thus the phase boundaries under respective models are

$$|\mu| = 1 - \frac{s-1}{K-1} \text{ and } |\mu| = 1 - p.$$

Notice that for the Bernoulli-Gaussian model with $K = 2$, the phase boundary agrees with the empirical phase boundary in Figure 3 of GS.

Example 5 *(Constant inner-product dictionaries) Let* $\mathbf{M}_0 = \mu \mathbf{1}\mathbf{1}^T + (1-\mu)\mathbf{I}$*, i.e.,* $\mathbf{D}_0[,i]^T\mathbf{D}_0[,j] = \mu$ *for* $1 \le i < j \le K$*. Note that* $\mathbf{M}_0$ *is positive definite if and only if* $\mu \in (-\frac{1}{K-1}, 1)$*. By Corollary 9, we have*

$$|||\mathbf{M}_0[-j,j]|||_s^* = \sqrt{s}|\mu|.$$

Thus for the s-sparse model, the phase boundary is

$$\sqrt{s}|\mu| = 1 - \frac{s-1}{K-1}.$$

Similarly for the Bernoulli(p)-Gaussian model, we have

$$|||\mathbf{M}_0[-j,j]|||_p^* = |\mu|p(K-1)\Big(\sum_{k=0}^{K-1} \texttt{pbinom}(k, K-1, p)\sqrt{k}\Big)^{-1}.$$

Thus the phase boundary is

$$|\mu| = \frac{1-p}{p(K-1)} \sum_{k=0}^{K-1} \texttt{pbinom}(k, K-1, p)\sqrt{k}.$$

Figures 3 shows the phase boundaries for different dictionary sizes under the two models. As K increases, the phase boundary moves toward the lower left of the region. This observation indicates that recovering the reference dictionary locally becomes increasingly difficult for larger dictionary size. See also Appendix Figures B.1 and B.2 for simulation results with larger K's.

The effect of non-sparse outliers. Example 5 demonstrates how the presence of non-sparse outliers in the Bernoulli-Gaussian model (Figure 2 Right) affects the requirements for local identifiability. Set $p = \frac{s}{K}$ in order to have the same level of sparsity with the $SG(s)$ model. Applying Jensen's inequality, one can show that

$$\frac{1-p}{p(K-1)} \sum_{k=0}^{K-1} \texttt{pbinom}(k, K-1, p)\sqrt{k} < \frac{1}{\sqrt{s}}(1 - \frac{s-1}{K-1}),$$

which indicates that the phase boundary of the s-sparse models is always above that of the Bernoulli-Gaussian model with the same level of sparsity. The gap between the two phase boundaries is the extra cost in terms of the collinearity parameter μ for locally recovering the dictionary in the presence of non-sparse outliers. One extreme example is the case where $s = 1$ and correspondingly $p = \frac{1}{K}$. By Example 1, under a 1-sparse model the reference dictionary $\mathbf{D}_0$ is always locally identifiable if $|\mu| < 1$. But for the $BG(\frac{1}{K})$ model, by the remarks under Corollary 1, $\mathbf{D}_0$ is not locally identifiable if $|\mu| > 1 - \frac{1}{K}$. Hence, the requirement for μ in the presence of outliers is at least $\frac{1}{K}$ more stringent than that in the case of no outliers.

However, such a difference diminishes as the number of dictionary atoms K increases. Indeed, by Lemma 2, one can establish the following lower bound for the phase boundary under the $BG(p)$ model

$$\frac{1-p}{p(K-1)} \sum_{k=0}^{K-1} \texttt{pbinom}(k, K-1, p)\sqrt{k} \ge \frac{1-p}{\sqrt{p(K-1)+1}} \approx \frac{1}{\sqrt{s}}(1 - \frac{s-1}{K-1}),$$

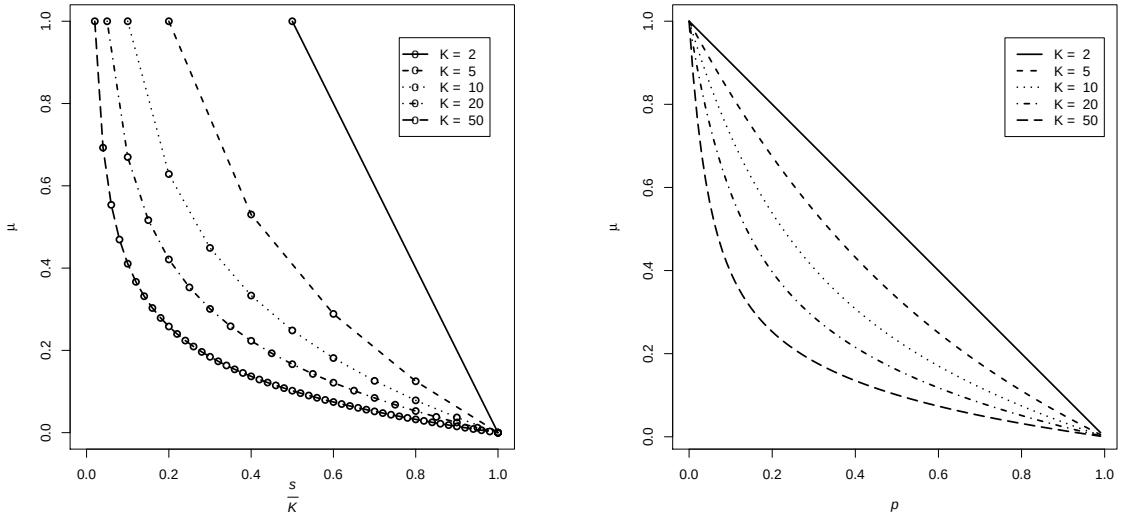

Figure 3: Local identifiability phase boundaries for constant inner-product dictionaries, under Left: the s-sparse Gaussian model; Right: the Bernoulli(p)-Gaussian model. For each model, phase boundaries for different dictionary sizes K are shown. Note that $\frac{s}{K} \in \{\frac{1}{K}, \frac{2}{K}, \ldots, 1\}$ and $p \in (0, 1]$. The area under the curves is the region where the reference dictionaries are locally identifiable at the population level. Due to symmetry, we only plot the portion of the phase boundaries for $\mu > 0$.

for fixed sparsity level $p = \frac{s}{K}$ and large K.

In general, the dual norms $|||.|||_s^*$ and $|||.|||_p^*$ have no closed-form expressions. According to Corollary 6 in the Appendix, computing those quantities involves solving a second order cone problem (SOCP) with exponentially many constraints. The following Lemma 2, on the other hand, gives computationally inexpensive approximation bounds.

Definition 3 *(Hyper-geometric distribution related quantities) Let m be a positive integer and $d, k \in \{0\} \cup [\![m]\!]$. Denote by $L_m(d, k)$ the hypergeometric random variable with parameter m, d and k, i.e., the number of 1's after drawing without replacement k elements from d 1's and $m - d$ 0's. Now for each $d \in \{0\} \cup [\![m]\!]$, define the function $\tau_m(d, .)$ with domain on $[0, m]$ as follows: set $\tau_m(d, 0) = 0$. For $a \in (k - 1, k]$ where $k \in [\![m]\!]$, define*

$$\tau_m(d, a) = \mathbb{E}\sqrt{L_m(d, k-1)} + (\mathbb{E}\sqrt{L_m(d, k)} - \mathbb{E}\sqrt{L_m(d, k-1)})(a - (k-1)).$$

Lemma 2 *(Lower and upper bounds for $|||.|||_s^*$ and $|||.|||_p^*$) Let m be a positive integer and $\mathbf{z} \in \mathbb{R}^m$.*

1. *For $s \in [\![m]\!]$,*

$$\max\left(\|\mathbf{z}\|_\infty, \sqrt{\frac{s}{m}} \max_{T \subset [\![m]\!]} \frac{\|\mathbf{z}[T]\|_1}{\sqrt{|T|}}\right) \leq \frac{s}{m} \max_{T \subset [\![m]\!]} \frac{\|\mathbf{z}[T]\|_1}{\tau_m(|T|, s)} \leq |||\mathbf{z}|||_s^* \leq \max_{S \subset [\![m]\!], |S| = s} \|\mathbf{z}[S]\|_2.$$

2. For $p \in (0,1)$,

$$\max\left(\|\mathbf{z}\|_\infty, \sqrt{p}\max_{T\subset[\![m]\!]}\frac{\|\mathbf{z}[T]\|_1}{\sqrt{|T|}}\right) \le p\max_{T\subset[\![m]\!]}\frac{\|\mathbf{z}[T]\|_1}{\tau_m(|T|,pm)} \le |||\mathbf{z}|||_p^* \le \max_{S\subset[\![m]\!],|S|=k}\|\mathbf{z}[S]\|_2.$$

where $k = \lceil p(m-1)+1\rceil$.

Remarks:
(1) We refer readers to Lemma 10 and 11 for the detailed version of the above results.
(2) Since we agree that $\frac{0}{0} = 0$, the case where $T = \emptyset$ does not affect taking the maximum of all subsets.
(3) Consider a sparse vector $\mathbf{z} = (z, 0, \ldots, 0)^T \in \mathbb{R}^m$. By Corollary 8,

$$|||\mathbf{z}|||_s^* = |||\mathbf{z}|||_p^* = |z| = \|\mathbf{z}\|_\infty = \max_{S\subset[\![m]\!],|S|=1}\|\mathbf{z}[S]\|_2.$$

So the all the bounds are achievable by a sparse vector.
(4) Now consider a dense vector $\mathbf{z} = (z, \ldots, z)^T \in \mathbb{R}^m$. By Corollary 9,

$$|||\mathbf{z}|||_s^* = \sqrt{s}|z| = \sqrt{\frac{s}{m}}\max_{T\subset[\![m]\!]}\frac{\|\mathbf{z}[T]\|_1}{\sqrt{|T|}} = \max_{S\subset[\![m]\!],|S|=s}\|\mathbf{z}[S]\|_2.$$

Thus the bounds for $|||\mathbf{z}|||_s^*$ can also be achieved by a dense vector. Similarly, by the upper-bound for $|||\mathbf{z}|||_p^*$,

$$|||\mathbf{z}|||_p^* \le \sqrt{pm+1}|z|.$$

On the other hand,

$$|||\mathbf{z}|||_p^* \ge \sqrt{p}\max_{T\subset[\![m]\!]}\frac{\|\mathbf{z}\|_1}{\sqrt{|T|}} = \sqrt{p}|z|\max_{T\subset[\![m]\!]}\sqrt{|T|} = \sqrt{pm}|z|.$$

Thus both bounds for $|||\mathbf{z}|||_p^*$ are basically the same for large pm.
(5) **Computation.** To compute the lower and upper bounds efficiently, we first sort the elements in $|\mathbf{z}|$ in descending order. Without loss of generality, we can assume that $|\mathbf{z}[1]| \ge |\mathbf{z}[2]| \ge \ldots \ge |\mathbf{z}[m]|$. Thus the upper-bound quantity becomes

$$\max_{S\subset[\![m]\!],|S|=k}\|\mathbf{z}[S]\|_2 = (\sum_{i=1}^k \mathbf{z}[i]^2)^{1/2}.$$

For the lower-bound quantities, note that

$$\max_{T\subset[\![m]\!]}\frac{\|\mathbf{z}[T]\|_1}{\tau_m(|T|,k)} = \max_{d\in[\![m]\!]}\max_{T\subset[\![m]\!],|T|=d}\frac{\|\mathbf{z}[T]\|_1}{\tau_m(d,k)} = \max_{d\in[\![m]\!]}\frac{\sum_{i=1}^d|\mathbf{z}[i]|}{\tau_m(d,k)}.$$

Thus, the major computation burden now is $\tau_m(d,k) = \mathbb{E}\sqrt{L_m(d,k)}$, for all $d \in [\![m]\!]$. We do not know a closed-form formula for $\mathbb{E}\sqrt{L_m(d,k)}$ except for $d = 1$ or $d = m$. In practice, we compute $\mathbb{E}\sqrt{L_m(d,k)}$ using its definition formula. On an OS X laptop with 1.8 GHz Intel Core i7 processor and 4GB of memory, the function `dhyper` in the statistics software

R can compute $\mathbb{E}\sqrt{L_{2000}(d,1000)}$ for all $d \in [\![2000]\!]$ within 0.635 second. Note that the number of dictionary atoms in most applications is typically smaller than 2000.

When m is too large, the LHS lower bounds can be used. Note that

$$\max_{T\subset[\![m]\!]} \frac{\|\mathbf{z}[T]\|_1}{\sqrt{|T|}} = \max_{d\in[\![m]\!]} \frac{\sum_{i=1}^{d}|\mathbf{z}[i]|}{\sqrt{d}},$$

which can be computed easily.

For notational simplicity, we will define the following quantities:

Definition 4 *For $a \in (0,K)$, define*

$$\nu_a(\mathbf{M}_0) = \max_{1\le j\le K} \max_{S\subset[\![K]\!], j\notin S} \frac{\|\mathbf{M}_0[S,j]\|_1}{\tau_{K-1}(|S|,a)}.$$

Definition 5 *(Cumulative coherence) For $k \in [\![K-1]\!]$, define the k-th cumulative coherence of a reference dictionary $\mathbf{D}_0$ as*

$$\mu_k(\mathbf{M}_0) = \max_{1\le j\le K} \max_{S\subset[\![K]\!], |S|=k, j\notin S} \|\mathbf{M}_0[S,j]\|_2.$$

Remarks: The above quantity is actually the ℓ_2 analog of the ℓ_1 k-th cumulative coherence defined in Gribonval et al. (2015). Also, notice that $\mu_1(\mathbf{M}_0) = \max_{l\neq j}|\mathbf{M}_0[l,j]|$ which is the plain mutual coherence of the reference dictionary.

With the above definitions and as a direct consequence of Lemma 2, we obtain a sufficient condition and a necessary condition for population local identifiability:

Corollary 1 *Under the notations of Theorem 1, we have*

1. *Let $K \ge 2$ and $s \in [\![K-1]\!]$.*
   - *If $\mu_s(\mathbf{M}_0) < 1 - \frac{s-1}{K-1}$, then $\mathbf{D}_0$ is locally identifiable with respect to $L_{SG(s)}$;*
   - *If $\frac{s}{K-1}\nu_s(\mathbf{M}_0) > 1 - \frac{s-1}{K-1}$, then $\mathbf{D}_0$ is not locally identifiable with respect to $L_{SG(s)}$.*
2. *Let $K \ge 2$ and $p \in (0,1)$.*
   - *If $\mu_k(\mathbf{M}_0) < 1-p$, where $k = \lceil p(K-2)+1\rceil$, then $\mathbf{D}_0$ is locally identifiable with respect to $L_{BG(p)}$;*
   - *If $p\nu_k(\mathbf{M}_0) > 1-p$, where $k = p(K-1)$, then $\mathbf{D}_0$ is not locally identifiable with respect to $L_{BG(p)}$.*

Remarks:
(1) In particular, by Lemma 2, if $\mu_1(\mathbf{M}_0) > 1 - \frac{s-1}{K-1}$ or $\mu_1(\mathbf{M}_0) > 1-p$, then $\mathbf{D}_0$ is not locally identifiable.
(2) We can also replace $\frac{s}{K-1}\nu_s(\mathbf{M}_0)$ or $p\nu_k(\mathbf{M}_0)$ by the corresponding lower bound quantities in Lemma 2 which are easier to compute but give weaker necessary conditions.

Comparison with GS. Corollary 1 enables us to compare our local identifiability condition directly with that of GS. For the Bernoulli(p)-Gaussian model, the population version of the sufficient condition for local identifiability by GS is:

$$\mu_{K-1}(\mathbf{M}_0) = \max_{1\le j\le K} \|\mathbf{M}_0[-j,j]\|_2 < 1-p. \tag{7}$$

Note that $\mu_{K-1}(\mathbf{M}_0) \ge \mu_k(\mathbf{M}_0)$ for $k \le K-1$.

Thus, our local identifiability result implies that of GS. Moreover, the quantity $\|\mathbf{M}_0[-j,j]\|_2$ in inequality (7) computes the ℓ_2-norm of the entire $\mathbf{M}_0[-j,j]$ vector and is independent of the sparsity parameter p. On the other hand, in our sufficient condition, $\max_{|S|=k, j\notin S} \|\mathbf{M}_0[S,j]\|_2$ computes the largest ℓ_2-norm of all size-k sub-vectors of $\mathbf{M}_0[-j,j]$. Since $k = \lceil p(K-2)+1\rceil$ is essentially pK, when the random linear coefficients are sparse, the sufficient bound by GS is much more conservative compared to ours.

More concretely, let us consider constant inner-product dictionaries with parameter $\mu > 0$ as in Example 5. The sufficient conditions by GS and by our Corollary 1 are respectively

$$\sqrt{K}\mu \le 1-p \text{ and } \sqrt{pK+1}\mu \le 1-p,$$

showing that the sufficient condition by GS is much more conservative for small value of p. See Figure 1, Appendix Figures B.1 and B.2 for a graphical comparison of the bounds for $K = 10, 20$ and 50.

Local identifiability for sparsity level $O(\mu^{-2})$. For notational convenience, let $\mu = \mu_1(\mathbf{M}_0)$ be the mutual coherence of the reference dictionary. For the s-sparse model, by Lemma 2, $\mu_s(\mathbf{M}_0) \le \sqrt{s}\mu$. Thus the first part of the corollary implies a simpler sufficient condition:

$$\sqrt{s}\mu < 1 - \frac{s-1}{K-1}.$$

From the above inequality, it can be seen that if $1 - \frac{s-1}{K-1} > \delta$ for some $\delta > 0$, the reference dictionary is locally identifiable for sparsity level s up to the order $O(\mu^{-2})$.

Similarly for the Bernoulli(p)-Gaussian model, since for $k = \lceil p(K-2)+1\rceil$,

$$\mu_k(\mathbf{M}_0) \le \sqrt{pK+1}\mu,$$

we have the following sufficient condition for local identifiability:

$$\sqrt{pK+1}\mu \le 1-p.$$

As before, if $1-p > \delta$ for some $\delta > 0$, the reference dictionary is locally identifiable for sparsity level pK up to the order $O(\mu^{-2})$. On the other hand, the same arguments for the condition by GS leads to $K = O(\mu^{-2})$, which, does not take advantage of sparsity.

In addition, by Example 5 and Remark (4) under Lemma 2, we also know that the sparsity requirement $O(\mu^{-2})$ cannot be improved in general.

Our result seems to be the first to demonstrate that $O(\mu^{-2})$ is the optimal order of sparsity level for exact local recovery of a reference dictionary. For a predefined over-complete dictionary, classical results such as Donoho and Elad (2003) and Fuchs (2004) show that basis pursuit recovers an s-sparse linear coefficient vector with sparsity level s

up to the order $O(\mu^{-1})$. For over-complete dictionary learning, Geng et al. (2011) showed that exact local recovery is also possible for s-sparse model with s up to $O(\mu^{-1})$. While our results are only for complete dictionaries, we conjecture that $O(\mu^{-2})$ is also the optimal order of sparsity level for over-complete dictionaries. In fact, Schnass (2015) proved that the response maximization criterion—an alternative formulation of dictionary learning—can approximately recover the over-complete reference dictionary locally with sparsity level s up to $O(\mu^{-2})$. It will be of interest to investigate whether the same sparsity requirement hold for the ℓ_1-minimization dictionary learning (2) in the case of exact local recovery and over-complete dictionaries.

A note on global identifiability. With the notations in Definition 1, we say that the reference dictionary $\mathbf{D}_0$ is *globally identifiable* with respect to $L(\mathbf{D})$ if (1) $\mathbf{D}_0$ is a global minimum of $L(\mathbf{D})$, and (2) for any other dictionary $\mathbf{D}'$ that cannot be transformed to $\mathbf{D}_0$ under any column permutation and sign changes, $L(\mathbf{D}_0) < L(\mathbf{D}')$. It is easy to see that global identifiability implies local identifiability, and so any necessary conditions for local identifiability are also necessary for global identifiability. Sufficient conditions for the global case are much harder to find. However, for dictionaries with orthogonal columns, we can show the following:

Corollary 2 *Suppose that the reference dictionary $\mathbf{D}_0$ is orthogonal, i.e., $\mathbf{M}_0 = \mathbf{I}$.*

1. *$\mathbf{D}_0$ is globally identifiable with respect to $L_{SG(s)}$ if and only if $\frac{s}{K} < 1$.*
2. *$\mathbf{D}_0$ is globally identifiable with respect to $L_{BG(p)}$ if and only if $p < 1$.*

The above result indicates that for orthogonal dictionaries, we have global identifiability as long as the linear coefficients $\boldsymbol{\alpha}_i$'s are not entirely dense. Note that these conditions are exactly the same as in the local identifiability case, see Example 3. Naturally, it is of greater interest to derive global identifiability condition for non-orthogonal dictionaries. Simulation results in Figure 3 of GS demonstrate that for $K = 2$, global identifiability seems to share the same phase transition boundary with local identifiability, i.e., $\mu + p = 1$. Furthermore, the surface plot of the objective function in Figure 2 of GS shows no other spurious local minima. To establish global identifiability for non-orthogonal dictionaries, it is unavoidable to characterize global optima of the non-convex objective function. An on-going work of Y. Wang and the authors (Wang et al.) might help shed light on this challenging problem.

4. Finite Sample Analysis

In this section, we will present finite sample results for local dictionary identifiability. For notational convenience, we first define the following quantities:

$$\begin{aligned}
\mathcal{P}_1(\epsilon, N; \mu, K) &= 2\exp\left(-\frac{N\epsilon^2}{108K\mu}\right), \\
\mathcal{P}_2(\epsilon, N; p, K) &= 2\exp\left(-p\frac{N\epsilon^2}{18p^2K + 9\sqrt{2pK}}\right), \\
\mathcal{P}_3(\epsilon, N; p, K) &= 3\left(\frac{24}{\epsilon p} + 1\right)^K \exp\left(-p\frac{N\epsilon^2}{360}\right).
\end{aligned}$$

Recall that $\mathbf{M}_0 = \mathbf{D}_0^T \mathbf{D}_0$ and $\mu_1(\mathbf{M}_0)$ is the mutual coherence of the reference dictionary $\mathbf{D}_0$. The following two theorems give local identifiability conditions under the s-sparse Gaussian model and the Bernoulli-Gaussian model:

Theorem 2 *(Finite sample local identifiability for SG(s)) Let $\boldsymbol{\alpha}_i \in \mathbb{R}^K$, $i \in [\![N]\!]$, be i.i.d. SG(s) random vectors with $s \in [\![K-1]\!]$. The signals $\mathbf{x}_i$'s are generated as $\mathbf{x}_i = \mathbf{D}_0 \boldsymbol{\alpha}_i$. Assume $0 < \epsilon \leq \frac{1}{2}$.*

1. *If*

$$\max_{j \in [\![K]\!]} |||\mathbf{M}_0[-j, j]|||_s^* \leq 1 - \frac{s-1}{K-1} - \sqrt{\frac{\pi}{2}}\epsilon,$$

then $\mathbf{D}_0$ is locally identifiable with respect to $L_N(\mathbf{D})$ with probability exceeding

$$1 - K^2 \left(\mathcal{P}_1(\epsilon, N; \mu_1(\mathbf{M}_0), K) + \mathcal{P}_2(\epsilon, N; \frac{s}{K}, K) + \mathcal{P}_3(\epsilon, N; \frac{s}{K}, K) \right).$$

2. *If*

$$\max_{j \in [\![K]\!]} |||\mathbf{M}_0[-j, j]|||_s^* \geq 1 - \frac{s-1}{K-1} + \sqrt{\frac{\pi}{2}}\epsilon,$$

then $\mathbf{D}_0$ is not locally identifiable with respect to $L_N(\mathbf{D})$ with probability exceeding

$$1 - K \left(\mathcal{P}_1(\epsilon, N; \mu_1(\mathbf{M}_0), K) + \mathcal{P}_2(\epsilon, N; \frac{s}{K}, K) + \mathcal{P}_3(\epsilon, N; \frac{s}{K}, K) \right).$$

Theorem 3 *(Finite sample local identifiability for BG(p)) Let $\boldsymbol{\alpha}_i \in \mathbb{R}^K$, $i \in [\![N]\!]$, be i.i.d. BG(p) random vectors with $p \in (0, 1)$. The signals $\mathbf{x}_i$'s are generated as $\mathbf{x}_i = \mathbf{D}_0 \boldsymbol{\alpha}_i$. Let $K_p = K + 2p^{-1}$ and assume $0 < \epsilon \leq \frac{1}{2}$.*

1. *If*

$$\max_{j \in [\![K]\!]} |||\mathbf{M}_0[-j, j]|||_p^* \leq 1 - p - \sqrt{\frac{\pi}{2}}\epsilon,$$

then $\mathbf{D}_0$ is locally identifiable with respect to $L_N(\mathbf{D})$ with probability exceeding

$$1 - K^2 \left(\mathcal{P}_1(\epsilon, N; \mu_1(\mathbf{M}_0), K_p) + \mathcal{P}_2(\epsilon, N; p, K_p) + \mathcal{P}_3(\epsilon, N; p, K) \right).$$

2. *If*

$$\max_{j \in [\![K]\!]} |||\mathbf{M}_0[-j, j]|||_p^* \geq 1 - p + \sqrt{\frac{\pi}{2}}\epsilon,$$

then $\mathbf{D}_0$ is not locally identifiable with respect to $L_N(\mathbf{D})$ with probability exceeding

$$1 - K \left(\mathcal{P}_1(\epsilon, N; \mu_1(\mathbf{M}_0), K_p) + \mathcal{P}_2(\epsilon, N; p, K_p) + \mathcal{P}_3(\epsilon, N; p, K) \right).$$

Remarks: The conditions for finite sample local identifiability are essentially identical to their population counterparts. The main difference is an margin of $\sqrt{\frac{\pi}{2}}\epsilon$ on the RHS of the inequalities. Such a margin appears as a result of our proof techniques: we show that the derivative of L_N is within $O(\epsilon)$ of its expectation and then apply local identifiability results for the population case.

Sample size requirement. The theorems indicate that if the number of signals is a multiple of the following quantity,

$$\text{For } SG(s)\text{: } \frac{1}{\epsilon^2}\max\left\{\mu_1(\mathbf{M}_0)K\log K,\ s\log K,\ \frac{K}{s}K\log\left(\frac{K}{\epsilon s}\right)\right\}$$

$$\text{For } BG(p)\text{: } \frac{1}{\epsilon^2}\max\left\{\mu_1(\mathbf{M}_0)K\log K,\ pK\log K,\ \frac{1}{p}K\log\left(\frac{1}{\epsilon p}\right)\right\}$$

then with high probability we can determine the conditions for local identifiability. Thus, in the worst case, the sample size requirements for the two models are respectively

$$O(\frac{K\log K}{\frac{s}{K}}) \text{ and } O(\frac{K\log K}{p}).$$

Our sample size requirement is similar to that of GS, who shows that $O(\frac{K\log K}{p(1-p)})$ signals is enough for locally recovering an incoherent reference dictionary. Our result indicates the $1-p$ factor in their denominator is not necessary.

The following two corollaries are the finite sample counterparts of Corollary 1.

Corollary 3 *Under the same assumptions of Theorem 2,*

1. *(Sufficient condition for $SG(s)$) If*

$$\mu_s(\mathbf{M}_0) \le 1 - \frac{s-1}{K-1} - \sqrt{\frac{\pi}{2}}\epsilon,$$

then $\mathbf{D}_0$ is locally identifiable with respect to $L_N(\mathbf{D})$, with the same probability bound in the first part of Theorem 2.

2. *(Necessary condition for $SG(s)$) If*

$$\frac{s}{K-1}\nu_s(\mathbf{M}_0) \ge 1 - \frac{s-1}{K-1} + \sqrt{\frac{\pi}{2}}\epsilon,$$

then $\mathbf{D}_0$ is not locally identifiable with respect to $L_N(\mathbf{D})$, with the same probability bound in the second part of Theorem 2.

Corollary 4 *Under the same assumptions of Theorem 3,*

1. *(Sufficient condition for $BG(p)$) Let $k = \lceil p(K-1)+1\rceil$. If*

$$\mu_k(\mathbf{M}_0) \le 1 - p - \sqrt{\frac{\pi}{2}}\epsilon,$$

then $\mathbf{D}_0$ is locally identifiable with respect to $L_N(\mathbf{D})$, with the same probability bound in the first part of Theorem 3.

2. *(Necessary condition for BG(p)) Let $k = p(K-1)$. If*

$$p\nu_k(\mathbf{M}_0) \geq 1 - p + \sqrt{\frac{\pi}{2}}\epsilon,$$

then $\mathbf{D}_0$ is not locally identifiable with respect to $L_N(\mathbf{D})$, with the same probability bound in the second part of Theorem 3.

Remarks: As before, denote by $\mu \in [0, 1)$ the coherence of the reference dictionary. The above two corollaries indicate that the reference dictionary is locally identifiable with high probability for sparsity level s or pK up to the order $O(\mu^{-2})$.

Proof sketch for Theorem 2 and 3. Similar to the population case, by taking one-sided derivatives of $L_N(\mathbf{D}_t)$ with respect to t at $t = 0$ for all smooth $\{\mathbf{D}_t\}_{t\in\mathbb{R}}$, we can derive a sufficient and almost necessary algebraic condition for the reference dictionary $\mathbf{D}_0$ to be a strict local minimum of $L_N(\mathbf{D})$. Using the concentration inequalities in Lemma 3—5, we show that the stochastic terms in the algebraic condition are close to their expectations with high probability. The population results for local identifiability can then be applied. The proofs for the two signal generation models are conceptually the same after establishing Lemma 8 to relate the $|||.|||_p^*$ norm to the $|||.|||_s^*$ norm. The detailed proof can be found in Section A.2.

Comparison with the proof by GS. The key difference between our analysis and that of GS is that we use an alternative but equivalent formulation of dictionary learning. Instead of (2), GS studied the following problem:

$$\min_{\mathbf{D}\in\mathcal{D},\boldsymbol{\alpha}_i} \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{\alpha}_i\|_1 \tag{8}$$
$$\text{subject to } \mathbf{x}_i = \mathbf{D}\boldsymbol{\alpha}_i \text{ for all } i \in [\![N]\!].$$

Note that the above formulation optimizes jointly over $\mathbf{D}$ and $\boldsymbol{\alpha}_i$ for $i \in [\![N]\!]$, as opposed to optimizing with respect to the only parameter $\mathbf{D}$ in our case. For complete dictionaries, this formulation is equivalent to the formulation in (2) in the sense that $\hat{\mathbf{D}}$ is a local minimum of (2) if and only if $(\hat{\mathbf{D}}, \hat{\mathbf{D}}^{-1}[\mathbf{x}_1, \ldots, \mathbf{x}_N])$ is a local minimum of (8), see Remark 3.1 of GS. The number of parameters to be estimated in (8) is $O(K^2 + KN)$, compared to $O(K^2)$ free parameters in (2). The growing number of parameters make the GS formulation less tractable to analyze under a signal generation model.

GS did not directly study the population case. They first obtained an algebraic condition for local identifiability that is sufficient and almost necessary. However, the condition is convoluted and hard to interpret due to its direct dependence on the signals $\mathbf{x}_i$'s. In order to determine the number of signals required for successful local recovery, they further investigated their condition under the Bernoulli-Gaussian model. During the probabilistic analysis, the sharp algebraic condition was weakened, resulting in a sufficient condition that is far from being necessary.

In contrast, we start with probabilistic generative models. The number of parameters remains constant as N increases. This allows us to study the population problem directly

and to apply concentration inequalities for the finite sample problem. Therefore, studying the optimization problem (2) instead of (8) is the key to establishing an interpretable sufficient and almost necessary local identifiability condition.

5. Conclusions and Future Work

We have established sufficient and almost necessary conditions for local dictionary identifiability under both the s-sparse Gaussian model and the Bernoulli-Gaussian model in the case of noiseless signals and complete dictionaries. For finite samples with a fixed sparsity level, we have shown that as long as the number of $i.i.d.$ signals scales as $O(K \log K)$, with high probability we can determine the local identifiability conditions of a reference dictionary.

There are several directions for future research. In this paper, we focused mainly on the local behaviors of the ℓ_1-norm objective function. As we previously discussed, investigating global identifiability conditions is a natural next step. Simulations in GS suggest a close connection between local and global identifiability. To understand this problem further, we need to characterize global optima of the non-convex objective function. In an on-going work of Y. Wang and the authors, we are developing promising techniques to analyze global properties of ℓ_1-minimization dictionary learning in the non-orthogonal case.

Moreover, one can extend our results to a wider class of sub-Gaussian distributions other than the standard Gaussian distribution considered in this paper. We foresee little technical difficulty for this extension. However, it should be noted that the quantities involved in our local identifiability conditions, i.e., the $|||.|||_s^*$ and $|||.|||_p^*$ norms, are consequences of the standard Gaussian assumption. Under a different distribution, it can be even more computationally challenging to verify the resulting local identifiability conditions.

Finally, it would be also desirable to improve the sufficient condition by Geng et al. (2011) and Gribonval et al. (2015) for over-complete dictionaries and noisy signals. One interesting implication of our results is that local recovery is possible for sparsity level up to the order $O(\mu^{-2})$ for a μ-coherent reference dictionary. We conjecture the same sparsity requirement for the over-complete and/or noisy signal cases. In either scenario, the closed-form expression for the objective function is no longer available. A full characterization of local dictionary identifiability demands novel techniques for analyzing the local behaviors of the objective function.

Acknowledgments

This research is supported in part by the Citadel Fellowship at the Department of Statistics of UCB, NHGRI grant U01HG007031, NSF grants DMS-1107000, CDS&E-MSS 1228246, DMS-1160319 (FRG), ARO grant W911NF-11-1-0114, AFOSR grant FA9550-14-1-0016,and the Center for Science of Information (CSoI), a US NSF Science and Technology Center, under grant agreement CCF-0939370. The authors would like to thank Sivaraman Balakrishnan and Yu Wang for helpful comments on the manuscript.

Appendix A. Proofs

Let $L(\mathbf{D})$ be the dictionary learning objective function and $\{\mathbf{D}_t\}_{t\in\mathbb{R}}$ be the collection of dictionaries $\mathbf{D}_t \in \mathcal{D}$ parameterized by $t \in \mathbb{R}$. By definition, $\{\mathbf{D}_t\}_{t\in\mathbb{R}}$ passes through the reference dictionary $\mathbf{D}_0$ at $t = 0$. Similar to Gribonval and Schnass (2010), to ensure that $\mathbf{D}_0$ is a strict local minimum of $L(\mathbf{D})$, it suffices to have

$$\lim_{t\downarrow 0^+} \frac{L(\mathbf{D}_t) - L(\mathbf{D}_0)}{t} > 0 \text{ and } \lim_{t\uparrow 0^-} \frac{L(\mathbf{D}_t) - L(\mathbf{D}_0)}{t} < 0,$$

for all $\{\mathbf{D}_t\}_{t\in\mathbb{R}}$ that is a smooth function of t with rate of approaching $\mathbf{D}_0$ bounded away from zero. On the other hand, if either of the above strict inequalities holds in the reversed direction for some smooth $\{\mathbf{D}_t\}_{t\in\mathbb{R}}$, then $\mathbf{D}_0$ is not a local minimum of $L(\mathbf{D})$.

Since $\mathbf{D}_0$ is full rank by assumption, the minimum eigenvalue of $\mathbf{M}_0 = \mathbf{D}_0^T\mathbf{D}_0$ is strictly greater than zero. By continuity of the minimum eigenvalue of $\mathbf{D}_t^T\mathbf{D}_t$ (see e.g., the Bauer-Fike Theorem), $\mathbf{D}_t$ should also be full rank when t is sufficiently small. Thus without loss of generality we can only consider full rank dictionaries $\mathbf{D}_t$. For any full rank $\mathbf{D} \in \mathcal{D}$, there is an invertible matrix $\mathbf{A} \in \mathbb{R}^{K\times K}$ such that $\mathbf{D} = \mathbf{D}_0\mathbf{A}$. For any $k \in [\![K]\!]$, by the constraint $\|\mathbf{D}[,k]\|_2 = 1$, $\mathbf{A}[,k]^T\mathbf{M}_0\mathbf{A}[,k] = 1$. Define the set for all such $\mathbf{A}$'s as

$$\mathcal{A} = \{\mathbf{A} \in \mathbb{R}^{K\times K} : \mathbf{A} \text{ is invertible and } \mathbf{A}[,k]^T\mathbf{M}_0\mathbf{A}[,k] = 1 \text{ for all } k \in [\![K]\!]\}. \tag{9}$$

It follows immediately that the set $\{\mathbf{D}_0\mathbf{A} : \mathbf{A} \in \mathcal{A}\}$ is the collection of $\mathbf{D} \in \mathcal{D}$ such that $\mathbf{D}$ is full rank. Thus, to ensure that $\mathbf{D}_0$ is a strict local minimum of $L(\mathbf{D})$, it suffices to show

$$\Delta^+(L, \{\mathbf{A}_t\}_t) \coloneqq \lim_{t\downarrow 0^+} \frac{L(\mathbf{D}_0\mathbf{A}_t) - L(\mathbf{D}_0)}{t} > 0, \tag{10}$$

$$\Delta^-(L, \{\mathbf{A}_t\}_t) \coloneqq \lim_{t\uparrow 0^-} \frac{L(\mathbf{D}_0\mathbf{A}_t) - L(\mathbf{D}_0)}{t} < 0, \tag{11}$$

for all smooth functions $\{\mathbf{A}_t\}_{t\in\mathbb{R}}$ with $\mathbf{A}_t \in \mathcal{A}$, $\mathbf{A}_0 = \mathbf{I}$ and nonzero derivative at $t = 0$. In addition, to demonstrate that $\mathbf{D}_0$ is not a local minimum of $L(\mathbf{D})$, it suffices to have (10) or (11) to hold in the reversed direction for some $\{\mathbf{A}_t\}_t$ with the aforementioned properties. We will be using this characterization of local minimum to prove local identifiability results for both the population case and the finite sample case.

A.1 Proofs of the Population Results

A.1.1 Proof of Lemma 1

Proof Since $\mathbb{E}||\mathbf{H}\boldsymbol{\alpha}_1||_1 = \sum_{j=1}^K \mathbb{E}|\mathbf{H}[j,]\boldsymbol{\alpha}_1|$, it suffices to compute $\mathbb{E}|\mathbf{H}[j,]\boldsymbol{\alpha}_1|$. Let S be any nonempty subset of $[\![K]\!]$. Recall that the random variable $\mathbf{S}_1 \subset [\![K]\!]$ denotes the support of random coefficient $\boldsymbol{\alpha}_1$. Conditioning on the event $\{\mathbf{S}_1 = S\}$, the random variable $\mathbf{H}[j,]\boldsymbol{\alpha}_1$ follows a normal distribution with mean 0 and standard deviation $\|\mathbf{H}[j,S]\|_2$. Hence

$$\mathbb{E}\left|\mathbf{H}[j,]\boldsymbol{\alpha}_1\right| = \mathbb{E}[\mathbb{E}[|\mathbf{H}[j,]\boldsymbol{\alpha}_1| \,|\mathbf{S}_1]] = \sqrt{\frac{2}{\pi}}\mathbb{E}\|\mathbf{H}[j,\mathbf{S}_1]\|_2.$$

(1) Under the s-sparse Gaussian model, $\mathbb{P}(\mathbf{S}_1 = S) = \binom{K}{s}^{-1}$ for any $|S| = s$. Thus we have

$$\mathbb{E}\|\mathbf{H}[j,\mathbf{S}_1]\|_2 = \binom{K}{s}^{-1} \sum_{S:|S|=s} \|\mathbf{H}[j,S]\|_2 = \frac{s}{K}|||\mathbf{H}[j,]|||_s.$$

Hence the objective function for the s-sparse Gaussian model is

$$L_{SG(s)}(\mathbf{D}) = \sum_{j=1}^{K} \mathbb{E}\,|\mathbf{H}[j,]\boldsymbol{\alpha}_1| = \sqrt{\frac{2}{\pi}}\frac{s}{K}\sum_{j=1}^{K}|||\mathbf{H}[j,]|||_s.$$

In particular, for $s = K$, $|||\mathbf{H}[j,]|||_K = \|\mathbf{H}[j,]\|_2$ and so

$$L_{SG(s)}(\mathbf{D}) = \sqrt{\frac{2}{\pi}}\sum_{j=1}^{K}\|\mathbf{H}[j,]\|_2.$$

(2) Under the Bernoulli(p)-Gaussian model, $\mathbb{P}(\mathbf{S}_1 = S) = p^{|S|}(1-p)^{K-|S|}$. So we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{H}[j,\mathbf{S}_1]\|_2] &= \sum_{k=1}^{K}\sum_{S:|S|=k} p^k(1-p)^{K-k}\|\mathbf{H}[j,S]\|_2 \\ &= p\sum_{k=0}^{K-1}\mathtt{pbinom}(k;K-1,p)|||\mathbf{H}[j,]|||_{k+1}.\end{aligned}$$

Therefore for $p \in (0,1)$, the objective function under the Bernoulli-Gaussian model is

$$L_{BG(p)}(\mathbf{D}) = \sum_{j=1}^{K}\mathbb{E}\,|\mathbf{H}[j,]\boldsymbol{\alpha}_1| = \sqrt{\frac{2}{\pi}}p\sum_{j=1}^{K}|||\mathbf{H}[j,]|||_p.$$

Finally, if $p = 1$, we have

$$L_{BG(p)}(\mathbf{D}) = \sqrt{\frac{2}{\pi}}\sum_{j=1}^{K}\|\mathbf{H}[j,]\|_2.$$

■

A.1.2 Proof of Theorem 1

Proof (1) Let us first consider the s-sparse Gaussian model. By (10) and (11), to ensure that $\mathbf{D}_0$ is a local minimum of $L_{SG(s)}(\mathbf{D})$, it suffices to show

$$\Delta^+(L_{SG(s)},\{\mathbf{A}_t\}_t) > 0 \text{ and } \Delta^-(L_{SG(s)},\{\mathbf{A}_t\}_t) < 0, \tag{12}$$

for all smooth functions $\{\mathbf{A}_t\}_t$ with $\mathbf{A}_t \in \mathcal{A}$, $\mathbf{A}_0 = \mathbf{I}$ and nonzero derivative at $t = 0$. Note that by Lemma 1,

$$\Delta^+(L_{SG(s)},\{\mathbf{A}_t\}_t) = \sqrt{\frac{2}{\pi}}\frac{s}{K}\sum_{j=1}^{K}\lim_{t\downarrow 0^+}\frac{1}{t}\left(\left|\left|\left|\mathbf{A}_t^{-1}[j,]\right|\right|\right|_s - |||\mathbf{I}[j,]|||_s\right). \tag{13}$$

For each $j \in [\![K]\!]$, we have

$$\binom{K-1}{s-1} \left|\!\left|\!\left| \mathbf{A}_t^{-1}[j,] \right|\!\right|\!\right|_s = \sum_{S:|S|=s, j\in S} \|\mathbf{A}_t^{-1}[j,S]\|_2 + \sum_{S:|S|=s, j\notin S} \|\mathbf{A}_t^{-1}[j,S]\|_2 \tag{14}$$

Denote by $\dot{\mathbf{A}}_0 \in \mathbb{R}^{K\times K}$ the derivative of $\{\mathbf{A}_t\}_t$ at $t=0$. Since $\mathbf{A}_t \in \mathcal{A}$ for all $t \in \mathbb{R}$, it can be shown that

$$\mathbf{M}_0[,k]^T \dot{\mathbf{A}}_0[,k] = 0 \ \text{ for all } k \in [\![K]\!]. \tag{15}$$

By (15), we have

$$\dot{\mathbf{A}}_0[j,j] = -\sum_{i\neq j} \mathbf{M}_0[i,j]\dot{\mathbf{A}}_0[i,j] \text{ for all } j \in [\![K]\!]. \tag{16}$$

Now notice that

$$\left.\frac{d\mathbf{A}_t^{-1}}{dt}\right|_{t=0} = -\mathbf{A}_0^{-1}\dot{\mathbf{A}}_0\mathbf{A}_0^{-1} = -\dot{\mathbf{A}}_0. \tag{17}$$

Combining the above equality with Lemma 14 and 15, we have

$$\lim_{t\downarrow 0^+} \frac{1}{t}\left(\|\mathbf{A}_t^{-1}[j,S]\|_2 - \|\mathbf{I}[j,S]\|_2\right) = \begin{cases} -\dot{\mathbf{A}}_0[j,j] & \text{if } j \in S \\ \|\dot{\mathbf{A}}_0[j,S]\|_2 & \text{if } j \notin S \end{cases}$$

Therefore

$$\lim_{t\downarrow 0^+} \frac{1}{t}\left(\left|\!\left|\!\left| \mathbf{A}_t^{-1}[j,] \right|\!\right|\!\right|_s - \left|\!\left|\!\left| \mathbf{I}[j,] \right|\!\right|\!\right|_s\right) = -\dot{\mathbf{A}}_0[j,j] + \binom{K-1}{s-1}^{-1} \sum_{S:|S|=s, j\notin S} \|\dot{\mathbf{A}}_0[j,S]\|_2. \tag{18}$$

Combining (13), (14), (16) and (18), we have

$$\begin{aligned}\sqrt{\frac{\pi}{2}}\frac{K}{s}\Delta^+(L_{SG(s)}, \{\mathbf{A}_t\}_t) &= -\sum_{j=1}^K \dot{\mathbf{A}}_0[j,j] + \binom{K-1}{s-1}^{-1} \sum_j \sum_{S:|S|=s, j\notin S} ||\dot{\mathbf{A}}_0[j,S]||_2 \\ &= \sum_{j=1}^K \Big(\sum_{i\neq j} \mathbf{M}_0[i,j]\dot{\mathbf{A}}_0[j,i] + \binom{K-1}{s-1}^{-1} \sum_{S:|S|=s, j\notin S} \|\dot{\mathbf{A}}_0[j,S]\|_2 \Big).\end{aligned}$$

Similarly, one can show

$$\sqrt{\frac{\pi}{2}}\frac{K}{s}\Delta^-(L_{SG(s)}, \{\mathbf{A}_t\}_t) = \sum_{j=1}^K \Big(\sum_{i\neq j} \mathbf{M}_0[i,j]\dot{\mathbf{A}}_0[j,i] - \binom{K-1}{s-1}^{-1} \sum_{S:|S|=s, j\notin S} \|\dot{\mathbf{A}}_0[j,S]\|_2 \Big).$$

Thus for $s \in [\![K-1]\!]$, to establish (12) it suffices to require for each $j \in [\![K]\!]$,

$$\Big| \sum_{i\neq j} \mathbf{M}_0[i,j]\dot{\mathbf{A}}_0[j,i] \Big| < \frac{K-s}{K-1}\binom{K-2}{s-1}^{-1} \sum_{S:|S|=s, j\notin S} \|\dot{\mathbf{A}}_0[j,S]\|_2 = \frac{K-s}{K-1} \left|\!\left|\!\left| \dot{\mathbf{A}}_0[j,-j] \right|\!\right|\!\right|_s. \tag{19}$$

for any $\dot{\mathbf{A}}_0$ such that $\dot{\mathbf{A}}_0[j,-j] \neq 0$. Since $\dot{\mathbf{A}}_0[j,i]$ is a free variable for $i \neq j$, (19) is equivalent to

$$\left|\mathbf{M}_0[-j,j]^T\mathbf{w}\right| < \frac{K-s}{K-1},$$

for all $\mathbf{w} \in \mathbb{R}^{K-1}$ such that $\|\mathbf{w}\|_s = 1$. Thus by the definition of the dual norm, it suffices to have

$$\|\mathbf{M}_0[-j,j]\|_s^* = \sup_{\|\mathbf{w}\|_s=1} \left|\mathbf{M}_0[-j,j]^T\mathbf{w}\right| < \frac{K-s}{K-1}.$$

Therefore, the condition

$$\max_{1\leq j\leq K} \|\mathbf{M}_0[-j,j]\|_s^* < \frac{K-s}{K-1} = 1 - \frac{s-1}{K-1}. \tag{20}$$

is sufficient for $\mathbf{D}_0$ to be locally identifiable with respect to the objective function $L_{SG(s)}$.

Similarly, one can check that if the reversed strict inequality in (20) holds, $\mathbf{D}_0$ is not a local minimum of $L_{SG(s)}(\mathbf{D})$. Thus we complete the proof for the s-sparse model.

(2) Now consider the Bernoulli(p)-Gaussian model for $p \in (0,1)$. First of all, note that we have

$$\begin{aligned}
&\sqrt{\frac{\pi}{2}}\frac{1}{p}\Delta^{\pm}(L_{BG(p)}, \{\mathbf{A}_t\}_t) = \sum_{j=1}^{K} \lim_{t\to 0^{\pm}} \frac{1}{t}\left(\left\|\left\|\mathbf{A}_t^{-1}[j,]\right\|\right\|_p - \|\|\mathbf{I}[j,]\|\|_p\right)\\
=&\sum_{j=1}^{K}\left(\sum_{i\neq j}\mathbf{M}_0[i,j]\dot{\mathbf{A}}_0[j,i] \pm (1-p)\sum_{k=0}^{K-2} p^k(1-p)^{K-2-k}\sum_{S:|S|=k+1, j\notin S}\|\dot{\mathbf{A}}_0[j,S]\|_2\right)\\
=&\sum_{j=1}^{K}\left(\dot{\mathbf{A}}_0[j,-j]^T\mathbf{M}_0[-j,j] \pm (1-p)\sum_{k=0}^{K-2}\texttt{pbinom}(k;K-2,p)\left\|\left\|\dot{\mathbf{A}}_0[j,-j]\right\|\right\|_{k+1}\right)\\
=&\sum_{j=1}^{K}\left(\dot{\mathbf{A}}_0[j,-j]^T\mathbf{M}_0[-j,j] \pm (1-p)\left\|\left\|\dot{\mathbf{A}}_0[j,-j]\right\|\right\|_p\right).
\end{aligned}$$

Thus, similar to the s-sparse Gaussian case, it can be shown that a sufficient condition for local identifiability is

$$\left|\mathbf{M}_0[-j,j]^T\mathbf{w}\right| < 1-p,$$

for all $j \in [\![K]\!]$ and all $\mathbf{w} \in \mathbb{R}^{K-1}$ such that $\|\mathbf{w}\|_p = 1$. The above condition is equivalent to

$$\max_{1\leq j\leq K} \|\mathbf{M}_0[-j,j]\|_p^* < 1-p.$$

The rest of the proof can be proceeded as in the case of the s-sparse Gaussian model.

(3) Let us consider the non-sparse case where $s = K$ or $p = 1$. In this case, since the objective functions are the same under both models (see Theorem 1), we only need to consider the s-sparse Gaussian model. If $s = K$, the RHS quantity in Inequality (19) is zero. Thus, the reference dictionary is not locally identifiable if

$$\left|\mathbf{M}_0[-j,j]^T\mathbf{w}\right| > 0,$$

for some $j \in [\![K]\!]$ and $\mathbf{w} \in \mathbb{R}^{K-1}$. Thus, if $\mathbf{M}_0$ is not the identity matrix, or equivalently, if the reference dictionary $\mathbf{D}_0$ is not orthogonal, $\mathbf{D}_0$ is not locally identifiable.

Next, let us deal with the case where $\mathbf{D}_0$ is orthogonal. Let $\mathbf{D} \in \mathcal{D}$ be a full rank dictionary and $\mathbf{W} = \mathbf{D}^{-1}$. Since $\mathbf{D}_0$ is orthogonal, $\|\mathbf{W}[j,]\mathbf{D}_0\|_2 = \|\mathbf{W}[j,]\|_2$. By the fact that $\mathbf{WD} = \mathbf{I}$ and $\|\mathbf{D}[,j]\|_2 = 1$, we have $1 = \mathbf{W}[j,]\mathbf{D}[,j] \le \|\mathbf{W}[j,]\|_2\|\mathbf{D}[,j]\|_2 = \|\mathbf{W}[j,]\|_2$, where the equality holds if and only if $\mathbf{W}[j,]^T = \pm\mathbf{D}[,j]$.

Under the K-sparse Gaussian model,

$$L_{SG(K)}(\mathbf{D}) = \sqrt{\frac{2}{\pi}} \sum_{j=1}^{K} \|\mathbf{W}[j,]\mathbf{D}_0\|_2 = \sqrt{\frac{2}{\pi}} \sum_{j=1}^{K} \|\mathbf{W}[j,]\|_2 \ge \sqrt{\frac{2}{\pi}} K = L_{SG(K)}(\mathbf{D}_0),$$

where the equality holds for any $\mathbf{D}$ such that $\mathbf{D}^T\mathbf{D} = \mathbf{I}$. Thus, $L_{SG(K)}(\mathbf{D}_0) = L_{SG(K)}(\mathbf{D}_0\mathbf{U})$ for any orthogonal matrix $\mathbf{U} \in \mathbb{R}^{K\times K}$, i.e., the objective function remains the same as we rotate $\mathbf{D}_0$. Therefore, $\mathbf{D}_0$ is not a strict local minimum of $L_{SG(K)}$.

In conclusion, $\mathbf{D}_0$ is not locally identifiable when $s = K$ or $p = 1$.

■

A.1.3 Proof of Corollary 2

Proof Let $\mathbf{D} \in \mathcal{D}$ be a full rank dictionary and $\mathbf{W} = \mathbf{D}^{-1}$. Under the K-sparse Gaussian model, $\mathbf{D}_0$ is not locally identifiable by Theorem 1. Hence it is not a strict local minimum of $L_{SG(K)}(\mathbf{D})$ and cannot be a strict global minimum.

Now suppose $s < K$, by Lemma 1 and 7,

$$L_{SG(s)}(\mathbf{D}) = \sqrt{\frac{2}{\pi}} \frac{s}{K} \sum_{j=1}^{K} |||\mathbf{W}[j,]\mathbf{D}_0|||_s \ge \sqrt{\frac{2}{\pi}} \frac{s}{K} \sum_{j=1}^{K} \|\mathbf{W}[j,]\mathbf{D}_0\|_2 \ge \sqrt{\frac{2}{\pi}} s = L_{SG(s)}(\mathbf{D}_0).$$

So $\mathbf{D}_0$ is a global minimum. Next we will show $\mathbf{D}_0$ is the only strict global minimum up to column permutation and sign changes. In the above formula, for the first equality to hold, we have $|||\mathbf{W}[j,]\mathbf{D}_0|||_s = \|\mathbf{W}[j,]\mathbf{D}_0\|_2$ for all j, implying that the vector $\mathbf{W}[j,]\mathbf{D}_0$ has at most one nonzero entry, see Lemma 7. For the second inequality to hold, $\mathbf{W}[j,]^T = \pm\mathbf{D}[,j]$, see arguments in Part (3) of the Theorem 1 proof. Combining these two conditions, $\|\mathbf{W}[j,]\mathbf{D}_0\|_2 = \|\mathbf{D}[j,]\|_2 = 1$ and so the nonzero entry can only be ± 1. Therefore, $\mathbf{WD}_0$ is an identity matrix after proper column permutation and sign changes.

The proof for the Bernoulli-Gaussian model is similar and hence omitted. ■

A.2 Proofs of the Finite Sample Results: Theorem 2 and Theorem 3

Proof We will first recall the signal generation procedure in Section 2. Let $\mathbf{z}$ be a K-dimensional standard Gaussian vector, and $\boldsymbol{\xi} \in \{0,1\}^K$ be either an s-sparse random vector or a Bernoulli random vector with probability p. Let $\mathbf{z}_1, \ldots, \mathbf{z}_N$ and $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N$ be identical and independent copies of $\mathbf{z}$ and $\boldsymbol{\xi}$ respectively. For each $i \in [\![N]\!]$ and $j \in [\![K]\!]$, define

$\boldsymbol{\alpha}_i[j] = \mathbf{z}_i[j]\boldsymbol{\xi}_i[j]$. For $S \subset [\![K]\!]$, define

$$\chi_i(S) = \begin{cases} 1 & \text{if } \boldsymbol{\xi}_i[k] = 1 \text{ for all } k \in S \text{ and } \boldsymbol{\xi}_i[k] = 0 \text{ for all } k \in S^c, \\ 0 & \text{otherwise.} \end{cases}$$

As in the population case, in the following analysis we will work with full rank dictionaries. First of all, notice that

$$l(\mathbf{D}, \mathbf{x}_i) = \|\mathbf{D}^{-1}\mathbf{x}_i\|_1 = \|\mathbf{D}^{-1}\mathbf{D}_0\boldsymbol{\alpha}_i\|_1 = \sum_{j=1}^{K} \left|\mathbf{A}^{-1}[j,]\boldsymbol{\alpha}_i\right| = \sum_{j=1}^{K}\sum_{k=1}^{K}\left(\sum_{S:|S|=k} |\mathbf{A}^{-1}[j,S]\mathbf{z}_i[S]|\chi_i(S)\right).$$

Next, we have

$$\begin{aligned}
\Delta^+(l(.,\mathbf{x}_i), \{\mathbf{A}_t\}_t) &= \lim_{t\downarrow 0^+} \frac{1}{t}\left(l(\mathbf{D}_0\mathbf{A}_t, \mathbf{x}_i) - l(\mathbf{D}_0, \mathbf{x}_i)\right) \\
&= \sum_{j=1}^{K}\Bigg(-\sum_{k=1}^{K}\sum_{S:j\in S,|S|=k} \dot{\mathbf{A}}_0[j,j]|\mathbf{z}_i[j]|\chi_i(S) \\
&\quad - \mathbf{sgn}(\mathbf{z}_i[j])\sum_{k=2}^{K}\sum_{S:j\in S,|S|=k}\sum_{l\in S,l\neq j} \dot{\mathbf{A}}_0[j,l]\mathbf{z}_i[l]\chi_i(S) \\
&\quad + \sum_{k=1}^{K-1}\sum_{S:j\notin S,|S|=k} |\dot{\mathbf{A}}_0[j,S]\mathbf{z}_i[S]|\chi_i(S)\Bigg). \qquad (21)
\end{aligned}$$

Here $\mathbf{sgn}(x)$ is the sign function of $x \in \mathbb{R}$ such that $\mathbf{sgn}(x) = 1$ for $x > 0$, $\mathbf{sgn}(x) = -1$ for $x < 0$ and $\mathbf{sgn}(x) = 0$ for $x = 0$. By (16), the first term in (21) can be rearranged as follows

$$\begin{aligned}
-\sum_{j=1}^{K}|\mathbf{z}_i[j]|\sum_{k=1}^{K}\sum_{S:j\in S,|S|=k}\dot{\mathbf{A}}_0[j,j]\chi_i(S) &= \sum_{j=1}^{K}|\mathbf{z}_i[j]|\sum_{k=1}^{K}\sum_{S:j\in S,|S|=k}\sum_{l\neq j}\mathbf{M}_0[l,j]\dot{\mathbf{A}}_0[l,j]\chi_i(S) \\
&= \sum_{j=1}^{K}\sum_{l\neq j}\mathbf{M}_0[j,l]\dot{\mathbf{A}}_0[j,l]\left(|\mathbf{z}_i[l]|\sum_{k=1}^{K}\sum_{S:l\in S,|S|=k}\chi_i(S)\right).
\end{aligned}$$

The second term in (21) can be rewritten as

$$-\sum_{j=1}^{K}\mathbf{sgn}(\mathbf{z}_i[j]) \times \sum_{l\neq j}(\dot{\mathbf{A}}_0[j,l]\mathbf{z}_i[l]) \times \sum_{k=2}^{K}\sum_{S:\{j,l\}\in S,|S|=k}\chi_i(S).$$

For $j, l \in [\![K]\!]$ such that $j \neq l$, define the following quantities

$$\mathbf{F}_i[l,j] = \mathbf{M}_0[j,l]|\mathbf{z}_i[l]|\sum_{k=1}^{K}\sum_{S:l\in S,|S|=k}\chi_i(S), \qquad (22)$$

$$\mathbf{G}_i[l,j] = \mathbf{sgn}(\mathbf{z}_i[j])\mathbf{z}_i[l]\sum_{k=2}^{K}\sum_{S:\{j,l\}\in S,|S|=s}\chi_i(S), \qquad (23)$$

whereas $\mathbf{F}[j,j] = \mathbf{G}[j,j] = 0$. For each $j \in [\![K]\!]$, also define

$$\mathbf{t}_i[j](\mathbf{w}) = \sum_{k=1}^{K-1} \sum_{S: j \notin S, |S|=k} |\mathbf{w}[S]^T \mathbf{z}_i[S]| \chi_i(S). \tag{24}$$

Let $\bar{\mathbf{F}}$, $\bar{\mathbf{G}}$ and $\bar{\mathbf{t}}$ be the sample average of $\mathbf{F}_i$, $\mathbf{G}_i$ and $\mathbf{t}_i$ respectively. With the definitions (22)—(24), we have

$$\begin{aligned}
\Delta^+(L_N, \{\mathbf{A}_t\}_t) &= \frac{1}{N} \sum_{i=1}^{N} \Delta^+(l(., \mathbf{x}_i), \{\mathbf{A}_t\}_t) \\
&= \sum_{j=1}^{K} \frac{1}{N} \sum_{i=1}^{N} \left(\dot{\mathbf{A}}_0[j,] \mathbf{F}_i[,j] + \dot{\mathbf{A}}_0[j,] \mathbf{G}_i[,j] + \mathbf{t}_i[j](\dot{\mathbf{A}}_0[j,]) \right) \\
&= \sum_{j=1}^{K} \left(\dot{\mathbf{A}}_0[j,] \bar{\mathbf{F}}[,j] - \dot{\mathbf{A}}_0[j,] \bar{\mathbf{G}}[,j] + \bar{\mathbf{t}}[j](\dot{\mathbf{A}}_0[j,]) \right)
\end{aligned}$$

On the other hand,

$$\Delta^-(L_N, \{\mathbf{A}_t\}_t) = \sum_{j=1}^{K} \left(\dot{\mathbf{A}}_0[j,] \bar{\mathbf{F}}[,j] - \dot{\mathbf{A}}_0[j,] \bar{\mathbf{G}}[,j] - \bar{\mathbf{t}}[j](\dot{\mathbf{A}}_0[j,]) \right).$$

Now for $j \in [\![K]\!]$, $s \in [\![K-1]\!]$ and $p \in (0,1)$, define

$$\mathcal{E}_j(s) = \{\mathbf{w} \in \mathbb{R}^K, |||\mathbf{w}[-j]|||_s = 1, \mathbf{w}[j] = 0\},$$
$$\mathcal{F}_j(p) = \{\mathbf{w} \in \mathbb{R}^K, |||\mathbf{w}[-j]|||_p = 1, \mathbf{w}[j] = 0\}.$$

Thus to ensure that $\mathbf{D}_0$ is a local minimum, it suffices to have for each $j \in [\![K]\!]$,

$$H_j(\mathbf{w}) \coloneqq \left| \mathbf{w}^T \bar{\mathbf{F}}[,j] - \mathbf{w}^T \bar{\mathbf{G}}[,j] \right| - \bar{\mathbf{t}}[j](\mathbf{w}) < 0,$$

for all $\mathbf{w} \in \mathcal{E}_j(s)$ for the s-sparse Gaussian model or all $\mathbf{w} \in \mathcal{F}_j(p)$ for the Bernoulli(p)-Gaussian model.

(1) For the s-sparse Gaussian model, let $j \in [\![K]\!]$ and define

$$h_j(\mathbf{w}) = \sqrt{\frac{2}{\pi}} \frac{s}{K} \left(\left| \mathbf{w}^T \mathbf{M}_0[,j] \right| - \frac{K-s}{K-1} \right),$$

which can be thought of as the expected value of $H_j(\mathbf{w})$. Note that by triangle inequality,

$$\begin{aligned}
&\sup_{\mathbf{w} \in \mathcal{E}_j(s)} |H_j(\mathbf{w}) - h_j(\mathbf{w})| \\
\leq &\sup_{\mathbf{w} \in \mathcal{E}_j(s)} \left| \mathbf{w}^T \left(\bar{\mathbf{F}}[,j] - \sqrt{\frac{2}{\pi}} \frac{s}{K} \mathbf{M}_0[,j] \right) \right| + \sup_{\mathbf{w} \in \mathcal{E}_j(s)} \left| \mathbf{w}^T \bar{\mathbf{G}}[,j] \right| + \sup_{\mathbf{w} \in \mathcal{E}_j(s)} \left| \bar{\mathbf{t}}[j](\mathbf{w}) - \sqrt{\frac{2}{\pi}} \frac{s}{K} \frac{K-s}{K-1} \right| \\
= &\left| \left| \left| \bar{\mathbf{F}}[-j,j] - \sqrt{\frac{2}{\pi}} \frac{s}{K} \mathbf{M}_0[-j,j] \right| \right| \right|_s^* + |||\bar{\mathbf{G}}[-j,j]|||_s^* + \sup_{\mathbf{w} \in \mathcal{E}_j(s)} \left| \bar{\mathbf{t}}[j](\mathbf{w}) - \sqrt{\frac{2}{\pi}} \frac{s}{K} \frac{K-s}{K-1} \right|.
\end{aligned} \tag{25}$$

Thus, $\sup_{\mathbf{w}\in\mathcal{E}_j(s)} |H_j(\mathbf{w}) - h_j(\mathbf{w})| > \frac{s}{K}\epsilon$ implies at least one of the three terms on the RHS is greater than $\frac{s}{K}\frac{\epsilon}{3}$. Using a union bound and by Lemma 3—5, we have

$$\begin{aligned}\mathbb{P}\left\{\sup_{\mathbf{w}\in\mathcal{E}_j(s)} |H_j(\mathbf{w}) - h_j(\mathbf{w})| > \frac{s}{K}\epsilon\right\} &\le 2K\exp\left(-\frac{N\epsilon^2}{108K\|\mathbf{M}_0[-j,j]\|_\infty}\right)\\ &\quad + 2K\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{18(s/K)s+9\sqrt{2s}}\right)\\ &\quad + 3\left(\frac{24K}{\epsilon s}+1\right)^K\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{360}\right). \end{aligned} \tag{26}$$

It is easy to see that the event $\left\{\sup_{\mathbf{w}\in\mathcal{E}_j(s)} |H_j(\mathbf{w}) - h_j(\mathbf{w})| \le \frac{s}{K}\epsilon\right\}$ implies

$$\sup_{\mathbf{w}\in\mathcal{E}_j(s)} h_j(\mathbf{w}) - \frac{s}{K}\epsilon \le \sup_{\mathbf{w}\in\mathcal{E}_j(s)} H_j(\mathbf{w}) \le \sup_{\mathbf{w}\in\mathcal{E}_j(s)} h_j(\mathbf{w}) + \frac{s}{K}\epsilon. \tag{27}$$

On the other hand,

$$\sup_{\mathbf{w}\in\mathcal{E}_j(s)} h_j(\mathbf{w}) = \sqrt{\frac{2}{\pi}}\frac{s}{K}\left(|||\mathbf{M}_0[-j,j]|||_s^* - \frac{K-s}{K-1}\right).$$

Thus, if $|||\mathbf{M}_0[-j,j]|||_s^* < \frac{K-s}{K-1} - \sqrt{\frac{\pi}{2}}\epsilon$, $\sup_{\mathbf{w}\in\mathcal{E}_j(s)} H_j(\mathbf{w}) < 0$ except with probability at most the bound in (26). To ensure $\mathbf{D}_0$ to be a local minimum, it suffices to have $\sup_{\mathbf{w}\in\mathcal{E}_j(s)} H_j(\mathbf{w}) < 0$ for all $j \in [\![K]\!]$. Thus, if $|||\mathbf{M}_0[-j,j]|||_s^* < \frac{K-s}{K-1} - \sqrt{\frac{\pi}{2}}\epsilon$ for all $j \in [\![K]\!]$, we have

$$\begin{aligned}\mathbb{P}\left\{\mathbf{D}_0 \text{ is locally identifiable}\right\} &\ge \mathbb{P}\left\{\max_j \sup_{\mathbf{w}\in\mathcal{E}_j(s)} H_j(\mathbf{w}) < 0\right\}\\ &\ge 1 - \mathbb{P}\left\{\max_j \sup_{\mathbf{w}\in\mathcal{E}_j(s)} H_j(\mathbf{w}) \ge 0\right\}\\ &\ge 1 - \sum_{j=1}^K \mathbb{P}\left\{\sup_{\mathbf{w}\in\mathcal{E}_j(s)} H_j(\mathbf{w}) \ge 0\right\}\\ &\ge 1 - \sum_{j=1}^K \mathbb{P}\left\{\sup_{\mathbf{w}\in\mathcal{E}_j(s)} |H_j(\mathbf{w}) - h_j(\mathbf{w})| > \frac{s}{K}\epsilon\right\}\\ &\ge 1 - 2K^2\exp\left(-\frac{N\epsilon^2}{108K\max_{l\neq j}|\mathbf{M}_0[l,j]|}\right)\\ &\quad - 2K^2\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{18(s/K)s+9\sqrt{2s}}\right)\\ &\quad - 3K\left(\frac{24K}{\epsilon s}+1\right)^K\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{360}\right). \end{aligned}$$

On the other hand, to ensure $\mathbf{D}_0$ is not locally identifiable with high probability, it suffices to have $|||\mathbf{M}_0[-j,j]|||_s^* > \frac{K-s}{K-1} + \sqrt{\frac{\pi}{2}}\epsilon$ for some $j \in [\![K]\!]$. Indeed, under that condition, the

LHS inequality in (27) implies $\sup_{\mathbf{w}\in\mathcal{E}_j(s)} H_j(\mathbf{w}) > 0$. Therefore

$$\begin{aligned}
\mathbb{P}\left\{\mathbf{D}_0 \text{ is not locally identifiable}\right\} &\geq \mathbb{P}\left\{\sup_{\mathbf{w}\in\mathcal{E}_j(s)} H_j(\mathbf{w}) > 0\right\} \\
&\geq 1 - \mathbb{P}\left\{\sup_{\mathbf{w}\in\mathcal{E}_j(s)} H_j(\mathbf{w}) \leq 0\right\} \\
&\geq 1 - \mathbb{P}\left\{\sup_{\mathbf{w}\in\mathcal{E}_j(s)} |H_j(\mathbf{w}) - h_j(\mathbf{w})| > \frac{s}{K}\epsilon\right\} \\
&\geq 1 - 2K\exp\left(-\frac{N\epsilon^2}{108K\|\mathbf{M}_0[-j,j]\|_\infty}\right) \\
&\quad - 2K\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{18(s/K)s + 9\sqrt{2s}}\right) \\
&\quad - 3\left(\frac{24K}{\epsilon s} + 1\right)^K \exp\left(-\frac{s}{K}\frac{N\epsilon^2}{360}\right).
\end{aligned}$$

(2) For the Bernoulli(p)-Gaussian model, define

$$\nu_j(\mathbf{w}) = \sqrt{\frac{2}{\pi}} p\left(\left|\mathbf{w}^T\mathbf{M}_0[,j]\right| - (1-p)\right).$$

Similar to (25), by triangle inequality,

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{F}_j(p)} |H_j(\mathbf{w}) - \nu_j(\mathbf{w})| \\
\leq &\left|\left|\left|\bar{\mathbf{F}}[-j,j] - \sqrt{\frac{2}{\pi}} p\mathbf{M}_0[-j,j]\right|\right|\right|_p^* + \left|\left|\left|\bar{\mathbf{G}}[-j,j]\right|\right|\right|_p^* + \sup_{\mathbf{w}\in\mathcal{F}_j(p)} \left|\bar{\mathbf{t}}[j](\mathbf{w}) - \sqrt{\frac{2}{\pi}} p(1-p)\right|.
\end{aligned}$$

Then the analysis can be carried out in a similar manner using the parallel version of the concentration inequalities, i.e., Part 2 of Lemma 3—5. ■

A.3 Concentration Inequalities

We will make frequent use of the following version of Bernstein's inequality. The proof of the inequality can be found in, e.g., Chapter 14 of Bühlmann and van de Geer (2011).

Theorem 4 *(Bernstein's inequality) Let $Y_1, \ldots, Y_N$ be independent random variables that satisfy the moment condition*

$$\mathbb{E}Y_i^m \leq \frac{1}{2} \times V \times m! \times B^{m-2},$$

for integers $m \geq 2$. Then

$$\mathbb{P}\left\{\frac{1}{N}\left|\sum_{i=1}^N Y_i - \mathbb{E}Y_i\right| > \epsilon\right\} \leq 2\exp\left(-\frac{N\epsilon^2}{2V + 2B\epsilon}\right).$$

Lemma 3 *(Uniform concentration of $\bar{\mathbf{F}}[-j,j]$) For $i \in [\![N]\!]$, let $\mathbf{F}_i \in \mathbb{R}^{K\times K}$ be defined as in (22) and $\bar{\mathbf{F}} = (1/N)\sum_{i=1}^N \mathbf{F}_i$.*

1. *Under the s-sparse Gaussian model with $s \in [\![K-1]\!]$,*

$$\mathbb{P}\left\{\left|\left|\left|\bar{\mathbf{F}}[-j,j] - \sqrt{\frac{2}{\pi}}\frac{s}{K}\mathbf{M}_0[-j,j]\right|\right|\right|_s^* > \frac{s}{K}\epsilon\right\} \le 2K\exp\left(-\frac{N\epsilon^2}{12K\|\mathbf{M}_0[-j,j]\|_\infty}\right),$$

   for $0 < \epsilon \le 1$.

2. *Under the Bernoulli-Gaussian model with parameter $p \in (0,1)$,*

$$\mathbb{P}\left\{\left|\left|\left|\bar{\mathbf{F}}[-j,j] - \sqrt{\frac{2}{\pi}}p\mathbf{M}_0[-j,j]\right|\right|\right|_p^* > p\epsilon\right\} \le 2K\exp\left(-\frac{N\epsilon^2}{12(K+2p^{-1})\|\mathbf{M}_0[-j,j]\|_\infty}\right),$$

   for $0 < \epsilon \le 1$.

In particular, if $\|\mathbf{M}_0[-j,j]\|_\infty = 0$, then the RHS bound is trivially zero.

Proof (1) First of all, we will prove the inequality for the s-sparse model. Notice that by Lemma 2, we have

$$\begin{aligned}\left|\left|\left|\bar{\mathbf{F}}[-j,j] - \sqrt{\frac{2}{\pi}}\frac{s}{K}\mathbf{M}_0[-j,j]\right|\right|\right|_s^* &\le \max_{|S|=s, j\notin S}\|\bar{\mathbf{F}}[S,j] - \sqrt{\frac{2}{\pi}}\frac{s}{K}\mathbf{M}_0[S,j]\|_2 \\ &\le \sqrt{s}\max_{l\neq j}|\bar{\mathbf{F}}[l,j] - \sqrt{\frac{2}{\pi}}\frac{s}{K}\mathbf{M}_0[l,j]|.\end{aligned}$$

For convenience, define

$$\mathbf{v}_i[l] = |\mathbf{z}_i[l]|\sum_{k=1}^K \sum_{|S|=k, l\in S}\chi_i(S) - \sqrt{\frac{2}{\pi}}\frac{s}{K}.$$

for $i \in [\![N]\!]$ and $l \in [\![K]\!]$. Note that $\sum_{k=1}^K \sum_{l\in S,|S|=k}\chi_i(S) = 1$ with probability $\binom{K}{s}^{-1}\binom{K-1}{s-1} = \frac{s}{K}$. Thus

$$\mathbb{E}\left(\sum_{k=1}^K \sum_{|S|=k,l\in S}\chi_i(S)\right)^m = \frac{s}{K}.$$

For $m \ge 1$, by Jensen's inequality $|\frac{a+b}{2}|^m \le \frac{1}{2}(|a|^m + |b|^m)$ and $\mathbb{E}|Z|^m \ge (\mathbb{E}|Z|)^m = \left(\frac{2}{\pi}\right)^{\frac{m}{2}}$, where Z is a standard Gaussian variable. In addition, $\mathbb{E}|Z|^m \le (m-1)!! \le 2^{-\frac{m}{2}}m!$. Hence

$$\begin{aligned}\mathbb{E}|\mathbf{v}_i[l]|^m &\le 2^{m-1}\left(\mathbb{E}|\mathbf{z}_i[l]|^m + \left(\frac{2}{\pi}\right)^{\frac{m}{2}}\left(\frac{s}{K}\right)^m\right) \\ &\le 2\times\mathbb{E}|Z|^m\times 2^{m-1} \\ &\le 2\times\left(\frac{1}{2}\right)^{\frac{m}{2}} m!\times 2^{m-1} \\ &= \frac{1}{2}\times\frac{4s}{K}\times m!\times(\sqrt{2})^{m-2}.\end{aligned}$$

Thus by Bernstein's inequality, we have

$$\mathbb{P}\left\{\left|\frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i[l]\right|>\epsilon\right\}\leq 2\exp\left(-\frac{N\epsilon^2}{2(4\frac{s}{K}+\sqrt{2}\epsilon)}\right).$$

Therefore,

$$\begin{aligned}\mathbb{P}\left\{\left|\mathbf{M}_0[j,l]\frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i[l]\right|>\frac{s}{K}\epsilon\right\}&\leq 2\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{2(4\mathbf{M}_0[j,l]^2+\sqrt{2}\left|\mathbf{M}_0[j,l]\right|\epsilon)}\right)\\&\leq 2\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{2\left|\mathbf{M}_0[j,l]\right|(4+\sqrt{2}\epsilon)}\right)\\&\leq 2\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{12\left|\mathbf{M}_0[j,l]\right|}\right).\end{aligned}$$

for $\epsilon\leq 1$. Notice that if $\mathbf{M}_0[j,l]=0$ the LHS probability is trivially zero. Using a union bound, we have

$$\begin{aligned}\mathbb{P}\left\{\|\bar{\mathbf{F}}[-j,j]-\sqrt{\frac{2}{\pi}}\frac{s}{K}\mathbf{M}_0[-j,j]\|_\infty>\frac{s}{K}\epsilon\right\}&=\mathbb{P}\left\{\max_{l\neq j}\left|\mathbf{M}_0[j,l]\frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i[l]\right|>\epsilon\right\}\\&\leq 2K\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{12\|\mathbf{M}_0[-j,j]\|_\infty}\right).\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{P}\left\{\left|\left|\left|\bar{\mathbf{F}}[-j,j]-\sqrt{\frac{2}{\pi}}\frac{s}{K}\mathbf{M}_0[-j,j]\right|\right|\right|_s^*>\frac{s}{K}\epsilon\right\}&\leq\mathbb{P}\left\{\sqrt{s}\|\bar{\mathbf{F}}[,j]-\sqrt{\frac{2}{\pi}}\frac{s}{K}\mathbf{M}_0[,j]\|_\infty>\frac{s}{K}\epsilon\right\}\\&\leq 2K\exp\left(-\frac{N\epsilon^2}{12K\|\mathbf{M}_0[-j,j]\|_\infty}\right).\end{aligned}$$

(2) Now let us consider the Bernoulli-Gaussian model. Notice that by Lemma 8, for $\frac{s-1}{K-1}\geq p$, we have

$$\begin{aligned}\left|\left|\left|\bar{\mathbf{F}}[-j,j]-\sqrt{\frac{2}{\pi}}p\mathbf{M}_0[-j,j]\right|\right|\right|_p^*&\leq\left|\left|\left|\bar{\mathbf{F}}[-j,j]-\sqrt{\frac{2}{\pi}}p\mathbf{M}_0[-j,j]\right|\right|\right|_s^*\\&\leq\sqrt{s}\left\|\bar{\mathbf{F}}[-j,j]-\sqrt{\frac{2}{\pi}}p\mathbf{M}_0[-j,j]\right\|_\infty.\end{aligned}$$

Now let $s=\lceil pK-p+1\rceil\leq pK+2$. For $i\in[\![N]\!]$ and $l\in[\![K]\!]$, define

$$\mathbf{u}_i[l]=|\mathbf{z}_i[l]|\sum_{k=1}^{K}\sum_{|S|=s,l\in S}\chi_i(S)-\sqrt{\frac{2}{\pi}}p.$$

Note that the event $\left\{\sum_{k=1}^{K}\sum_{|S|=k,l\in S}\chi_i(S)=1\right\}$ is the same as the event that $\{\boldsymbol{\alpha}_i[l]=1\}$, which, happens with probability p. Thus

$$\mathbb{E}\left(\sum_{k=1}^{K}\sum_{|S|=k,l\in S}\chi_i(S)\right)^m = p.$$

Similar to the case of s-sparse model,

$$\mathbb{E}\,|\mathbf{u}_i[l]|^m \le \frac{1}{2}\times 4p\times m!\times\left(\sqrt{2}\right)^{m-2}.$$

By Bernstein's inequality, we have

$$\mathbb{P}\left\{\left|\frac{1}{N}\sum_{i=1}^{N}\mathbf{u}_i[l]\right|>\epsilon\right\}\le 2\exp\left(-\frac{N\epsilon^2}{2(4p+\sqrt{2}\epsilon)}\right).$$

Therefore

$$\begin{aligned}\mathbb{P}\left\{\left|\left|\left|\bar{\mathbf{F}}[-j,j]-\sqrt{\frac{2}{\pi}}p\mathbf{M}_0[-j,j]\right|\right|\right|_p^*>p\epsilon\right\} &\le \mathbb{P}\left\{\sqrt{s}\|\bar{\mathbf{F}}[,j]-\sqrt{\frac{2}{\pi}}p\mathbf{M}_0[,j]\|_\infty>p\epsilon\right\}\\ &\le 2K\exp\left(-\frac{p}{s}\frac{N\epsilon^2}{2\|\mathbf{M}_0[-j,j]\|_\infty(4+\sqrt{2}\epsilon)}\right)\\ &\le 2K\exp\left(-\frac{N\epsilon^2}{12(K+2p^{-1})\|\mathbf{M}_0[-j,j]\|_\infty}\right),\end{aligned}$$

for $\epsilon\le 1$. ■

Lemma 4 *(Uniform concentration of $\bar{\mathbf{G}}[-j,j]$) For $i\in[\![N]\!]$, let $\mathbf{G}_i\in\mathbb{R}^{K\times K}$ be defined as in (23) and $\bar{\mathbf{G}}=(1/N)\sum_{i=1}^{N}\mathbf{G}_i$.*

1. *Under the s-sparse Gaussian model with $s\in[\![K-1]\!]$,*

$$\mathbb{P}\left\{\left|\left|\left|\bar{\mathbf{G}}[-j,j]\right|\right|\right|_s^*>\frac{s}{K}\epsilon\right\}\le 2K\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{2(s/K)s+\sqrt{2s}}\right),$$

   for $0<\epsilon\le 1$.

2. *Under the Bernoulli-Gaussian model with parameter $p\in(0,1)$,*

$$\mathbb{P}\left\{\left|\left|\left|\bar{\mathbf{G}}[-j,j]\right|\right|\right|_p^*>p\epsilon\right\}\le 2K\exp\left(-p\frac{N\epsilon^2}{p(pK+2)+\sqrt{2(pK+2)}}\right),$$

   for $0<\epsilon\le 1$.

Proof The proof is highly similar to that of Lemma 3 and so we will omit some common steps.
(1) We first prove the concentration inequality for the s-sparse model. Notice that

$$\left|\left|\left|\bar{\mathbf{G}}[-j,j]\right|\right|\right|_s^* \leq \sqrt{s}\max_{l\neq j}|\bar{\mathbf{G}}[l,j]|.$$

In addition,

$$\begin{aligned}\mathbb{E}\left(\sum_{k=2}^{K}\sum_{\{j,l\}\in S,|S|=k}\chi_i(S)\right)^m &= \mathbb{E}\left(\sum_{k=2}^{K}\sum_{|S|=k,\{j,l\}\in S}\chi_i(S)\right)\\ &= \binom{K}{s}^{-1}\binom{K-2}{s-2} = \frac{s(s-1)}{K(K-1)} \leq (\frac{s}{K})^2.\end{aligned}$$

Thus

$$\mathbb{E}|\mathbf{G}_i[l,j]|^m \leq 2^{-m/2}m!\times(\frac{s}{K})^2 = \frac{1}{2}\times(\frac{s}{K})^2\times m!\times(\frac{1}{\sqrt{2}})^{m-2}.$$

By Bernstein inequality:

$$\mathbb{P}\left\{\left|\frac{1}{N}\sum_{i=1}^{N}\mathbf{G}_i[l,j]\right| > \epsilon\right\} \leq 2\exp\left(-\frac{N\epsilon^2}{2(s/K)^2+\sqrt{2}\epsilon)}\right).$$

Thus we have

$$\begin{aligned}\mathbb{P}\left\{\left|\left|\left|\bar{\mathbf{G}}[-j,j]\right|\right|\right|_s^* > \frac{s}{K}\epsilon\right\} &\leq \mathbb{P}\left\{\sqrt{s}\max_{l\neq j}|\bar{\mathbf{G}}[l,j]| > \frac{s}{K}\epsilon\right\}\\ &\leq 2K\exp\left(-\frac{(s/K)^2N(\epsilon^2/s)}{2(s/K)^2+\sqrt{2}(s/K)(\epsilon/\sqrt{s}))}\right)\\ &\leq 2K\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{2(s/K)s+\sqrt{2s}\epsilon}\right)\\ &\leq 2K\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{2(s/K)s+\sqrt{2s}}\right),\end{aligned}$$

for $\epsilon \leq 1$.
(2) For Bernoulli-Gaussian model, notice that

$$\left|\left|\left|\bar{\mathbf{G}}[-j,j]\right|\right|\right|_p^* \leq \left|\left|\left|\bar{\mathbf{G}}[-j,j]\right|\right|\right|_s^* \leq \sqrt{s}\max_{l\neq j}|\bar{\mathbf{G}}[l,j]|,$$

for $s = \lceil pK-p+1\rceil \leq pK+2$. Also,

$$\mathbb{E}|\mathbf{G}_i[l,j]|^m \leq 2^{-m/2}m!\times p^2 = \frac{1}{2}\times p^2\times m!\times(\frac{1}{\sqrt{2}})^{m-2}.$$

Thus

$$\begin{aligned}\mathbb{P}\left\{\left|\left|\left|\bar{\mathbf{G}}[-j,j]\right|\right|\right|_s^* > p\epsilon\right\} &\leq \mathbb{P}\left\{\sqrt{s}\max_{l\neq j}|\bar{\mathbf{G}}[l,j]| > p\epsilon\right\}\\ &\leq 2K\exp(-p\frac{N(\epsilon^2)}{2ps+\sqrt{2s}})\\ &\leq 2K\exp\left(-p\frac{N\epsilon^2}{p(pK+2)+\sqrt{2(pK+2)}}\right),\end{aligned}$$

for $\epsilon \leq 1$. ■

Lemma 5 *(Uniform concentration of $\bar{\mathbf{t}}[j](\mathbf{w})$) For $i \in [\![N]\!]$, let $\mathbf{t}_i$ be the function from $\mathbb{R}^K$ to $\mathbb{R}^K$ defined as in (24) and $\bar{\mathbf{t}} = (1/N)\sum_{i=1}^N \mathbf{t}_i$. Recall that for $j \in [\![K]\!]$, $s \in [\![K-1]\!]$ and $p \in (0,1)$,*

$$\mathcal{E}_j(s) = \{\mathbf{w} \in \mathbb{R}^K, |||\mathbf{w}[-j]|||_s = 1, \mathbf{w}[j] = 0\},$$

$$\mathcal{F}_j(p) = \{\mathbf{w} \in \mathbb{R}^K, |||\mathbf{w}[-j]|||_p = 1, \mathbf{w}[j] = 0\}.$$

1. *Under the s-sparse Gaussian model with $s \in [\![K-1]\!]$,*

$$\mathbb{P}\left\{\sup_{w \in \mathcal{E}_j(s)} \left|\bar{\mathbf{t}}[j](\mathbf{w}) - \sqrt{\frac{2}{\pi}}\frac{s}{K}\frac{K-s}{K-1}\right| > \frac{s}{K}\epsilon\right\} \leq 3\left(\frac{8K}{\epsilon s}+1\right)^K \exp\left(-\frac{s}{K}\frac{N\epsilon^2}{40}\right),$$

   for $0 < \epsilon \leq \frac{1}{2}$.

2. *Under the Bernoulli-Gaussian model with parameter $p \in (0,1)$,*

$$\mathbb{P}\left\{\sup_{w \in \mathcal{F}_j(p)} \left|\bar{\mathbf{t}}[j](\mathbf{w}) - \sqrt{\frac{2}{\pi}}p(1-p)\right| > p\epsilon\right\} \leq 3\left(\frac{8}{\epsilon p}+1\right)^K \exp\left(-p\frac{N\epsilon^2}{40}\right),$$

   for $0 < \epsilon \leq \frac{1}{2}$.

Proof (1) Under the s-sparse model, we have

$$\begin{aligned}
\mathbb{E}|\mathbf{t}_i[j](\mathbf{w})|^m &= \mathbb{E}\left(\sum_{|S|=s, j\notin S} |\mathbf{w}[S]^T\mathbf{z}_i[S]|\chi_i(S)\right)^m \\
&= \sum_{|S|=s, j\notin S} \mathbb{E}|\mathbf{w}[S]^T\mathbf{z}_i[S]|^m \mathbb{E}\chi_i(S) \\
&= \binom{K}{s}^{-1} \sum_{|S|=s, j\notin S} \mathbb{E}|\mathbf{w}[S]^T\mathbf{z}_i[S]|^m.
\end{aligned}$$

Notice that we have used the facts that the events $\chi_i(S)$'s are mutually exclusive and that $\mathbf{z}_i[S]$ and $\chi_i(S)$ are independent. Since the random variable $\mathbf{w}[S]^T\mathbf{z}_i[S]$ has distribution $N(0, \|\mathbf{w}[S]\|_2)$, $\mathbb{E}|\mathbf{w}[S]^T\mathbf{z}_i[S]|^m = \|\mathbf{w}[S]\|_2^m \mathbb{E}|Z|^m \leq 2^{-\frac{m}{2}} m!$. Therefore

$$\mathbb{E}|\mathbf{t}_i[j](\mathbf{w})|^m \leq 2^{-\frac{m}{2}} m! \binom{K}{s}^{-1} \sum_{j\notin S, |S|=s} \|\mathbf{w}[S]\|_2^m.$$

Note that by Lemma 7, $|||\mathbf{w}[-j]|||_s \geq \|\mathbf{w}[-j]\|_2 \geq \|\mathbf{w}[S]\|_2$ for all S such that $j \notin S$. For $\mathbf{w} \in \mathcal{E}_j(s)$, $|||\mathbf{w}|||_s = 1$ and so $\|\mathbf{w}_S\|_2 \leq 1$, which, further implies that $\|\mathbf{w}[S]\|_2^m \leq \|\mathbf{w}[S]\|_2$.

Thus we have

$$\begin{aligned}\mathbb{E}|\mathbf{t}_i[j](\mathbf{w})|^m &\le 2^{-\frac{m}{2}} m! \binom{K}{s}^{-1} \sum_{j\notin S,|S|=s} \|\mathbf{w}[S]\|_2 \\ &\le 2^{-\frac{m}{2}} m! \frac{s(K-s)}{K(K-1)} |||\mathbf{w}[-j]|||_s \\ &= 2^{-\frac{m}{2}} m! \frac{s(K-s)}{K(K-1)}\end{aligned}$$

For a fixed j, define

$$U_i(\mathbf{w}) = \mathbf{t}_i[j](\mathbf{w}) - \sqrt{\frac{2}{\pi}} \frac{s}{K} \frac{K-s}{K-1}.$$

Notice that $\mathbb{E}U_i(\mathbf{w}) = 0$. In addition,

$$\mathbb{E}|U_i(\mathbf{w})|^m \le 2^m \mathbb{E}|\mathbf{t}_i[j](\mathbf{w})|^m \le \frac{1}{2} \times 4 \frac{s}{K} \frac{K-s}{K-1} \times m! \times (\sqrt{2})^{m-2}.$$

By Bernstein's inequality

$$\mathbb{P}\left\{ \frac{1}{N} \left| \sum_{i=1}^N U_i(\mathbf{w}) \right| > \frac{s}{K}\epsilon \right\} \le 2\exp\left(-\frac{s}{K} \frac{N\epsilon^2}{2(4\frac{K-s}{K-1} + \sqrt{2}\epsilon)} \right) \le 2\exp\left(-\frac{s}{K} \frac{N\epsilon^2}{10} \right),$$

for $0 < \epsilon \le 1/2$. Now let $\{\mathbf{w}_i\}$ be an δ-cover of $\mathcal{E}_j(s)$. Since $\mathcal{E}_j(s)$ is contained in the unit ball $\{\mathbf{w} \in \mathbb{R}^{K-1} : \|\mathbf{w}\|_2 \le 1\}$, there exists a cover such that $|\{\mathbf{w}_l\}| \le \left(\frac{2}{\delta} + 1\right)^{K-1}$. For any $\mathbf{w}, \mathbf{w}' \in \mathcal{E}_j(s)$, we have

$$\left| U_i(\mathbf{w}) - U_i(\mathbf{w}') \right| \le \sum_{j\notin S,|S|=s} \left| (\mathbf{w}[S] - \mathbf{w}'[S])^T \mathbf{z}_i[S] \right| \chi_i(S).$$

Let Z be a standard Gaussian variable. We have

$$\begin{aligned}\mathbb{P}\left\{ \sum_{|S|=s, j\notin S} \left| \mathbf{w}[S]^T \mathbf{z}_i[S] \right| \chi_i(S) > \epsilon \right\} &= \binom{K-1}{s}^{-1} \sum_{|S|=s, j\notin S} \mathbb{P}\left\{ \left| \mathbf{w}[S]^T \mathbf{z}_i[S] \right| > \epsilon \right\} \\ &= \binom{K-1}{s}^{-1} \sum_{|S|=s, j\notin S} \mathbb{P}\left\{ \|\mathbf{w}[S]\|_2 |Z| > \epsilon \right\} \\ &\le \mathbb{P}\left\{ \|\mathbf{w}\|_2 |Z| > \epsilon \right\}.\end{aligned}$$

Let Z_i, $i = 1, \ldots, N$, be *i.i.d.* standard Gaussian variables. By the one-sided Bernstein's inequality,

$$\mathbb{P}\left\{ \frac{1}{N} \sum_{i=1}^N |Z_i| \ge 2 \right\} \le \exp\left(-\frac{N(2 - \sqrt{2/\pi})^2}{2(4 + \sqrt{2}(2 - \sqrt{2/\pi}))} \right) \le \exp\left(-\frac{N}{8} \right).$$

Now let $\delta = \frac{s}{K}\frac{\epsilon}{4}$. Thus

$$\begin{aligned}\mathbb{P}\left\{\sup_{\|\mathbf{w}-\mathbf{w}'\|_2\le\delta}\left|\frac{1}{N}\sum_{i=1}^{N}(U_i(\mathbf{w})-U_i(\mathbf{w}'))\right| > \frac{s}{K}\frac{\epsilon}{2}\right\} &\le \mathbb{P}\left\{\sup_{\|\mathbf{w}'-\mathbf{w}\|_2\le\delta}\frac{1}{N}\sum_{i=1}^{N}\left|U_i(\mathbf{w})-U_i(\mathbf{w}')\right| > \frac{s}{K}\frac{\epsilon}{2}\right\}\\ &\le \mathbb{P}\left\{\sup_{\|\mathbf{w}'-\mathbf{w}\|_2\le\delta}\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{w}-\mathbf{w}'\|_2|Z_i| > \frac{s}{K}\frac{\epsilon}{2}\right\}\\ &\le \mathbb{P}\left\{\delta\frac{1}{N}\sum_{i=1}^{N}|Z_i| > \frac{s}{K}\frac{\epsilon}{2}\right\} \le \mathbb{P}\left\{\frac{1}{N}\sum_{i=1}^{N}|Z_i| > 2\right\}\\ &\le \exp\left(-\frac{N}{8}\right).\end{aligned}$$

By triangle inequality

$$\sup_{\|\mathbf{w}'-\mathbf{w}\|_2\le\delta}\left|\frac{1}{N}\sum_{i=1}^{N}U_i(\mathbf{w}')\right| \le \sup_{\|\mathbf{w}'-\mathbf{w}\|_2\le\delta}\left|\frac{1}{N}\sum_{i=1}^{N}(U_i(\mathbf{w})-U_i(\mathbf{w}'))\right| + \left|\frac{1}{N}\sum_{i=1}^{N}U_i(\mathbf{w})\right|.$$

Using a union bound, we have

$$\begin{aligned}\mathbb{P}\{\sup_{\|\mathbf{w}'-\mathbf{w}\|_2\le\delta}\left|\frac{1}{N}\sum_{i=1}^{N}U_i(\mathbf{w}')\right| > \frac{s}{K}\epsilon\} &\le \mathbb{P}\left\{\sup_{\|\mathbf{w}-\mathbf{w}'\|_2\le\delta}\left|\frac{1}{N}\sum_{i=1}^{N}(U_i(\mathbf{w})-U_i(\mathbf{w}'))\right| > \frac{s}{K}\frac{\epsilon}{2}\right\}\\ &\quad + \mathbb{P}\left\{\left|\frac{1}{N}\sum_{i=1}^{N}U_i(\mathbf{w})\right| > \frac{s}{K}\frac{\epsilon}{2}\right\}\\ &\le \exp\left(-\frac{N}{8}\right) + 2\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{40}\right)\\ &\le 3\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{40}\right),\end{aligned}$$

for $0 < \epsilon \le 1$. Now apply union bound again,

$$\begin{aligned}\mathbb{P}\left\{\sup_{\mathbf{w}\in\mathcal{E}_j(s)}\frac{1}{N}\left|\sum_{i=1}^{N}U_i(\mathbf{w})\right| > \frac{s}{K}\epsilon\right\} &\le \mathbb{P}\left\{\max_{l}\sup_{||\mathbf{w}-\mathbf{w}_l||_2\le\delta}\frac{1}{N}\left|\sum_{i=1}^{N}U_i(\mathbf{w})\right| > \frac{s}{K}\epsilon\right\}\\ &\le 3\left(\frac{8K}{\epsilon s}+1\right)^{K}\exp\left(-\frac{s}{K}\frac{N\epsilon^2}{40}\right).\end{aligned}$$

(2) For $\mathbf{w} \in \mathcal{F}_j(p)$, under the Bernoulli-Gaussian model,

$$\begin{aligned}
\mathbb{E}\left|\mathbf{t}_i[j](\mathbf{w})\right|^m &= \mathbb{E}|Z|^m \sum_{k=1}^{K-1} \sum_{|S|=k, j\notin S} \|\mathbf{w}[S]\|_2^m \times p^k (1-p)^{K-k} \\
&\leq \mathbb{E}|Z|^m p \sum_{k=1}^{K-1} \sum_{|S|=k, j\notin S} \|\mathbf{w}[S]\|_2 \times p^{k-1} (1-p)^{K-k} \\
&= \mathbb{E}|Z|^m p(1-p) \sum_{k=0}^{K-2} \sum_{|S|=k+1, j\notin S} \|\mathbf{w}[S]\|_2 \times p^k (1-p)^{K-2-k} \\
&= \mathbb{E}|Z|^m p(1-p) \left|\!\left|\!\left|\mathbf{w}[-j]\right|\!\right|\!\right|_p = \mathbb{E}|Z|^m p(1-p) \\
&\leq 2^{-m/2} m! p(1-p).
\end{aligned}$$

Notice that we have used the fact that $\|\mathbf{w}[S]\|_2 \leq \|\mathbf{w}[-j]\|_2 \leq \left|\!\left|\!\left|\mathbf{w}[-j]\right|\!\right|\!\right|_p = 1$ for all S such that $j \notin S$. For each fixed $\mathbf{w}$, define

$$V_i(\mathbf{w}) = \mathbf{t}_i[j](\mathbf{w}) - \sqrt{\frac{2}{\pi}}(1-p)p.$$

Now we have

$$\mathbb{E}|V_i(\mathbf{w})|^m \leq 2^m \mathbb{E}|\mathbf{t}_i[j](\mathbf{w})|^m \leq \frac{1}{2} \times 4p(1-p) \times m! \times (\sqrt{2})^{m-2}.$$

The remaining parts of the proof can be proceeded exactly as in the case of the s-sparse model, noticing that we only need to replace $\frac{s}{K}$ by p, and $\frac{K-s}{K-1}$ by $1-p$.

■

A.4 Dual Analysis of $\left|\!\left|\!\left|.\right|\!\right|\!\right|_s$ and $\left|\!\left|\!\left|.\right|\!\right|\!\right|_p$

In this section, we will characterize the dual norms $\left|\!\left|\!\left|.\right|\!\right|\!\right|_s^*$ and $\left|\!\left|\!\left|.\right|\!\right|\!\right|_p^*$ by second order cone programs (SOCP). The characterization is helpful for deriving bounds for these special norms in the next section.

Lemma 6 *For $i \in [\![M]\!]$, let $\mathbf{A}_i$ be an $k_i \times K$ with rank k_i. For $\mathbf{z} \in \mathbb{R}^K$, define*

$$\|\mathbf{z}\|_{\mathbf{A}} = \sum_{i=1}^{M} \|\mathbf{A}_i \mathbf{z}\|_2.$$

Then the dual norm of $\|.\|_{\mathbf{A}}$ is

$$\|\mathbf{v}\|_{\mathbf{A}}^* = \inf \left\{ \max_i \|\mathbf{y}_i\|_2, \mathbf{y}_i \in \mathbb{R}^{k_i}, \sum_{i=1}^{M} \mathbf{A}_i^T \mathbf{y}_i = \mathbf{v} \right\}.$$

Proof

$$\|\mathbf{v}\|_{\mathbf{A}}^* = \sup_{\mathbf{z}\neq 0} \frac{\mathbf{v}^T\mathbf{z}}{\|\mathbf{z}\|_{\mathbf{A}}} = \sup\left\{\mathbf{v}^T\mathbf{z} : \|\mathbf{z}\|_{\mathbf{A}} \leq 1\right\}.$$

Introducing Lagrange multiplier $\lambda \geq 0$ for the inequality constraint, the above problem is equivalent to the following

$$\begin{aligned}\|\mathbf{v}\|_{\mathbf{A}}^* &= \sup_{\mathbf{z}}\left\{\inf_{\lambda\geq 0}\left\{\mathbf{v}^T\mathbf{z} + \lambda(1 - \|\mathbf{z}\|_{\mathbf{A}})\right\}\right\} \\ &= \sup_{\mathbf{z}}\left\{\inf_{\lambda\geq 0}\left\{\mathbf{v}^T\mathbf{z} + \lambda(1 - \sum_{i=1}^{M}\|\mathbf{A}_i\mathbf{z}\|_2)\right\}\right\}.\end{aligned}$$

The dual problem is

$$d = \inf_{\lambda\geq 0}\left\{\sup_{\mathbf{z}}\left\{\mathbf{v}^T\mathbf{z} + \lambda(1 - \sum_{i=1}^{M}\|\mathbf{A}_i\mathbf{z}\|_2)\right\}\right\}.$$

Notice that $\|\mathbf{A}_i\mathbf{z}\|_2 = \sup\{\mathbf{z}^T\mathbf{A}_i^T\mathbf{u}_i : \|\mathbf{u}_i\|_2 \leq 1\}$. Hence

$$d = \inf_{\lambda\geq 0}\left\{\lambda + \sup_{\mathbf{z},\mathbf{u}}\left\{\mathbf{z}^T(\mathbf{v} - \lambda\sum_{i=1}^{M}\mathbf{A}_i^T\mathbf{u}_i) : \|\mathbf{u}_i\|_2 \leq 1\right\}\right\}.$$

Since the vector $\mathbf{z}$ can be arbitrary, in order to have a finite value, we must have $\lambda\sum_{i=1}^{M}\mathbf{A}_i^T\mathbf{u}_i = \mathbf{v}$. Now let $\mathbf{y}_i = \lambda\mathbf{u}_i$, the problem becomes

$$d = \inf_{\lambda\geq 0}\left\{\lambda : \sum_{i=1}^{M}\mathbf{A}_i^T\mathbf{y}_i = \mathbf{v}, \|\mathbf{y}_i\|_2 \leq \lambda\right\}.$$

The above problem is exactly equivalent to

$$\inf\left\{\max_i \|\mathbf{y}_i\|_2, \mathbf{y}_i \in \mathbb{R}^{k_i}, \sum_{i=1}^{M}\mathbf{A}_i^T\mathbf{y}_i = \mathbf{v}\right\}.$$

Finally, notice that the original problem is convex and strictly feasible. Thus Slater's condition holds and the duality gap is zero. Hence

$$||\mathbf{v}||_{\mathbf{A}}^* = \inf\left\{\max_i \|\mathbf{y}_i\|_2, \mathbf{y}_i \in \mathbb{R}^{k_i}, \sum_{i=1}^{M}\mathbf{A}_i^T\mathbf{y}_i = \mathbf{v}\right\}.$$

■

The following corollary gives an alternative characterization of $|||.|||_s$ and $|||.|||_p$:

Corollary 5 *Denote by $\mathbf{y}_S \in \mathbb{R}^{|S|}$ a variable vector indexed by the set S (as opposed to being a sub-vector of $\mathbf{y}$). For $\mathbf{z} \in \mathbb{R}^m$, we have*

$$|||\mathbf{z}|||_s^* = \inf \left\{ \max_{|S|=s} \|\mathbf{y}_S\|_2 : \mathbf{y}_S \in \mathbb{R}^s, \sum_{|S|=s} \mathbf{E}_S^T \mathbf{y}_S = \mathbf{z} \right\},$$

and

$$|||\mathbf{z}|||_p^* = \inf \left\{ \max_{S} \|\mathbf{y}_S\|_2 : \mathbf{y}_S \in \mathbb{R}^{|S|}, \sum_{k=0}^{m-1} \texttt{pbinom}(k; m-1, p) \sum_{|S|=k+1} \mathbf{E}_S^T \mathbf{y}_S = \mathbf{z} \right\},$$

where $\mathbf{E}_S = \mathbf{I}[S,]/\binom{m-1}{|S|-1}$ and $\mathbf{I} \in \mathbb{R}^{m \times m}$ is the identity matrix.

Proof This is simply a direct application of Lemma 6. ■

Corollary 6 *The dual norms $|||.|||_s^*$ and $|||.|||_p^*$ can be computed via a Second Order Cone Program (SOCP).*

Proof Introducing additional variable $t \geq 0$, the problem of computing $|||\mathbf{z}|||_s^*$ is equivalent to the following formulation

$$\begin{aligned} \inf_{t, \mathbf{y}_S} \quad & t \;\; s.t. \; \|\mathbf{y}_S\|_2 \leq t \text{ for all } S \text{ such that } |S| = s \\ & \text{and } \sum_{|S|=s} \mathbf{E}_S^T \mathbf{y}_S = \mathbf{z}. \end{aligned}$$

Notice that the above program is already in the standard form of SOCP. The case of $|||.|||_p^*$ can be handled in a similar manner. ■

A.5 Inequalities of $|||.|||_s$ and $|||.|||_p$ and Their Duals

As demonstrated in the last section, it is in general expensive to compute $|||.|||_s^*$ and $|||.|||_p^*$. In this section, we will derive sharp and easy-to-compute lower and upper bounds to approximate these quantities.

Lemma 7 *(Monotonicity of $|||\mathbf{z}|||_s$ and $|||\mathbf{z}|||_p$) Let $\mathbf{z} \in \mathbb{R}^m$. $|||\mathbf{z}|||_1 = \|\mathbf{z}\|_1$ and $|||\mathbf{z}|||_m = \|\mathbf{z}\|_2$. For $1 \leq l < k \leq m$, we have $|||\mathbf{z}|||_l \geq |||\mathbf{z}|||_k$; similarly for $0 < p < q < 1$, $|||\mathbf{z}|||_p \geq |||\mathbf{z}|||_q$. Furthermore, the equalities hold iff the vector $\mathbf{z}$ contains at most one non-zero entry.*

Proof By definition, we have

$$|||\mathbf{w}|||_1 = \frac{\sum_{|S|=1} \|\mathbf{w}[S]\|_2}{\binom{m-1}{1-1}} = \|\mathbf{w}\|_1.$$

Similarly,

$$|||\mathbf{w}|||_m = \frac{\sum_{|S|=m} \|\mathbf{w}[S]\|_2}{\binom{m-1}{m-1}} = \|\mathbf{w}\|_2.$$

For $1 \le k \le m-1$, let S' be a subset of $[\![m]\!]$ such that $|S'| = k+1$. By triangle inequality

$$\sum_{|S|=k, S\subset S'} \|\mathbf{z}[S]\|_2 \ge k\|\mathbf{z}[S']\|_2,$$

where the equality holds iff $\|\mathbf{z}[S']\|_0 \le 1$. Thus

$$\sum_{|S'|=k+1} \sum_{|S|=k, S\subset S'} \|\mathbf{z}[S]\|_2 \ge k \sum_{|S'|=k+1} \|\mathbf{z}[S']\|_2,$$

and the equality holds iff $\|\mathbf{z}\|_0 \le 1$. Notice that the LHS of the above inequality is simply $(m-k)\sum_{|S|=k} \|\mathbf{z}[S]\|_2$. Therefore

$$|||\mathbf{z}|||_k = \binom{m-1}{k-1}^{-1} \sum_{|S|=k} \|\mathbf{z}[S]\|_2 \ge \binom{m-1}{k}^{-1} \sum_{|S|=k+1} \|\mathbf{z}[S]\|_2 = |||\mathbf{z}|||_{k+1},$$

and so the inequality holds.

For $|||.|||_p$, let Y be a random variable that follows the binomial distribution with parameters $m-1$ and p. Observe that $|||\mathbf{z}|||_p = \mathbb{E}|||\mathbf{z}|||_{Y+1}$, where the expectation is taken with respect to Y. If $\|\mathbf{z}\|_0 > 1$, $|||\mathbf{z}|||_k$ is strictly decreasing in k by the first part. Hence, $|||\mathbf{z}|||_p$ as a function of p is also strictly decreasing on $(0,1)$. Indeed, it can be shown that

$$\frac{d}{dp}|||\mathbf{z}|||_p = \sum_{k=0}^{m-1} \texttt{pbinom}(k; K-1, p)\left(|||\mathbf{z}|||_{k+1} - |||\mathbf{z}|||_k\right) < 0.$$

If $\|\mathbf{z}\|_0 \le 1$, then $|||\mathbf{z}|||_1 = |||\mathbf{z}|||_m$ and so $\frac{d}{dp}|||\mathbf{z}|||_p = 0$. Therefore $|||\mathbf{z}|||_p = |||\mathbf{z}|||_1$ is a constant in p. On the other hand, if $|||\mathbf{z}|||_p = |||\mathbf{z}|||_q$ for $0 < p < q < 1$, by the fact that $\frac{d}{dp}|||\mathbf{z}|||_p \le 0$, we must have $\frac{d}{dp}|||\mathbf{z}|||_p = 0$ and so $|||\mathbf{z}|||_{k-1} = |||\mathbf{z}|||_k$ for all $k \in [\![m]\!]$. Thus $\|\mathbf{z}\|_0 \le 1$. ■

Corollary 7 *(Monotonicity of $|||\mathbf{z}|||_s^*$ and $|||\mathbf{z}|||_p^*$) Let $\mathbf{z} \in \mathbb{R}^m$. $|||\mathbf{z}|||_1^* = \|\mathbf{z}\|_\infty$ and $|||\mathbf{z}|||_m^* = \|\mathbf{z}\|_2$. For $1 \le i < j \le m$, we have $|||\mathbf{z}|||_i^* \le |||\mathbf{z}|||_j^*$; similarly for $0 < p < q < 1$, $|||\mathbf{z}|||_p^* \le |||\mathbf{z}|||_q^*$. Furthermore, the equalities hold iff the vector $\mathbf{z}$ contains at most one non-zero entry.*

Proof This is a direct consequence of Lemma 7 and the dual norm definition $|||\mathbf{z}|||^* = \sup_{\mathbf{y}\neq 0} \frac{\mathbf{z}^T\mathbf{y}}{|||\mathbf{y}|||}$. ■

Lemma 8 *Let $p \in (0,1)$ and $k = \lceil (m-1)p + 1 \rceil$. For any $\mathbf{z} \in \mathbb{R}^m$, we have*

1. $|||\mathbf{z}|||_p \ge |||\mathbf{z}|||_k$.

2. $\|\mathbf{z}\|_p^* \leq \|\mathbf{z}\|_k^*$.

Proof Define the function f with domain on $[1, m]$ as follows: let $f(1) = \|\mathbf{z}\|_1 = \|\mathbf{z}\|_1$; for $i \in [\![m-1]\!]$ and $a \in (i, i+1]$, define

$$f(a) = \|\mathbf{z}\|_i + (\|\mathbf{z}\|_{i+1} - \|\mathbf{z}\|_i)(a - i).$$

It is clear that f is piecewise linear by construction. In addition, by Lemma 12, f is also convex. Notice that $\|\mathbf{z}\|_p = \mathbb{E}\|\mathbf{z}\|_{Y+1} = \mathbb{E}f(Y+1)$, where Y is a random variable from the binomial distribution with parameters $m-1$ and p. By Jensen's inequality,

$$\mathbb{E}f(Y+1) \geq f(\mathbb{E}Y + 1) = f((m-1)p + 1).$$

Thus by Lemma 7, $\|\mathbf{z}\|_p \geq \|\mathbf{z}\|_k$ for all $k \geq (m-1)p + 1$. So the first part follows.

To upperbound $\|\mathbf{z}\|_p^*$, notice that if $k \geq (m-1)p + 1$,

$$\|\mathbf{z}\|_p^* = \sup_{\mathbf{w} \neq 0} \frac{\mathbf{w}^T\mathbf{z}}{\|\mathbf{w}\|_p} \leq \sup_{\mathbf{w} \neq 0} \frac{\mathbf{w}^T\mathbf{z}}{\|\mathbf{w}\|_k} = \|\mathbf{z}\|_k^*.$$

■

For the following lemmas, the quantities $\tau_m(d, a)$ and $L_m(d, k)$ are defined as in Definition 3.

Lemma 9 *(Approximating $\tau_m(d, a)$) For $d \in [\![m]\!]$ and $a \in (0, m]$,*

$$\tau_m(d, a) \leq \sqrt{\frac{da}{m}}.$$

Proof For $k \in [\![m]\!]$, by Jensen's inequality,

$$\mathbb{E}\sqrt{L_m(d, k)} \leq \sqrt{\mathbb{E}L_m(d, k)} = \sqrt{\frac{dk}{m}}.$$

Note that the last equality follows from the expectation of a hypergeometric random variable. Now suppose $a \in (k-1, k]$. By the above inequality and apply Jensen's inequality one more time, we have

$$\begin{aligned}\tau_m(d, a) &= (k-a)\mathbb{E}\sqrt{L_m(d, k-1)} + (1-(k-a))\mathbb{E}\sqrt{L_m(d, k)} \\ &\leq (k-a)\sqrt{\frac{d(k-1)}{m}} + (1-(k-a))\sqrt{\frac{dk}{m}} = \sqrt{\frac{da}{m}}.\end{aligned}$$

■

Lemma 10 *(Lower bounds for $\|\mathbf{z}\|_s^*$ and $\|\mathbf{z}\|_p^*$) Let $\mathbf{z} \in \mathbb{R}^m$. We have*

1. *For $s \in [\![m]\!]$,*

$$|||\mathbf{z}|||_s^* \geq \frac{s}{m} \max_{T \subset [\![m]\!]} \frac{\|\mathbf{z}[T]\|_1}{\tau_m(|T|, s)} \geq \max\left(\|\mathbf{z}\|_\infty, \sqrt{\frac{s}{m}} \max_{T \subset [\![m]\!]} \frac{\|\mathbf{z}[T]\|_1}{\sqrt{|T|}} \right).$$

2. *For $p \in (0,1)$,*

$$\begin{aligned} |||\mathbf{z}|||_p^* &\geq p \max_{T \subset [\![m]\!]} \left\{ \Big(\sum_{k=0}^{m} \mathtt{pbinom}(k, m, p) \tau_m(|T|, k) \Big)^{-1} \|\mathbf{z}[T]\|_1 \right\} \\ &\geq p \max_{T \subset [\![m]\!]} \frac{\|\mathbf{z}[T]\|_1}{\tau_m(|T|, pm)} = \max\left(\|\mathbf{z}\|_\infty, \sqrt{p} \max_{T \subset [\![m]\!]} \frac{\|\mathbf{z}[T]\|_1}{\sqrt{|T|}} \right). \end{aligned}$$

Proof (1) Note that by definition,

$$|||\mathbf{z}|||_s^* = \sup_{\mathbf{w}} \frac{\mathbf{z}^T \mathbf{w}}{|||\mathbf{w}|||_s}$$

Let $d \in [\![m]\!]$ and $T \subset [\![m]\!]$ such that $|T| = d$. Define $\mathbf{w} \in \mathbb{R}^m$ such that $\mathbf{w}[i] = 1$ for $i \in T$ and $\mathbf{w}[i] = 0$ for $i \in T^c$. We have:

$$\begin{aligned} |||\mathbf{w}|||_s = \binom{m-1}{s-1}^{-1} \sum_{|S|=s} \|\mathbf{w}[S]\|_2 &= \binom{m-1}{s-1}^{-1} \sum_{l=\max(0,s+d-m)}^{\min(s,d)} \sum_{|S|=s, |S\cap T|=l} \|\mathbf{w}[S]\|_2 \\ &= \binom{m-1}{s-1}^{-1} \sum_{l=\max(0,s+d-m)}^{\min(s,d)} \sum_{|S|=s, |S\cap T|=l} \sqrt{l} \\ &= \binom{m-1}{s-1}^{-1} \sum_{l=\max(0,s+d-m)}^{\min(s,d)} \binom{d}{l} \binom{m-d}{s-l} \sqrt{l} \\ &= \frac{m}{s} \mathbb{E}\sqrt{L_m(d,s)} = \frac{m}{s} \tau_m(d,s). \end{aligned}$$

Thus for all $d \in [\![m]\!]$ and any subset T such that $|T| = d$, we have shown

$$|||\mathbf{z}|||_s^* \geq \frac{s}{m} \frac{\|\mathbf{z}[T]\|_1}{\tau_m(d,s)}.$$

Note that if $d = 1$, $\mathbb{E}\sqrt{L_m(d,s)} = \frac{s}{m}$. Therefore

$$\frac{s}{m} \frac{\|\mathbf{z}[T]\|_1}{\tau_m(d,s)} \geq \|\mathbf{z}\|_\infty,$$

Moreover, by Lemma 9,

$$\tau_m(d,s) \leq \sqrt{\frac{ds}{m}}.$$

Hence we have

$$\frac{s}{m}\frac{\|\mathbf{z}[T]\|_1}{\tau_m(d,s)} \geq \sqrt{\frac{s}{m}}\frac{\|\mathbf{z}[T]\|_1}{\sqrt{d}},$$

and the first part of the claim follows.
(2) For the same $\mathbf{w} \in \mathbb{R}^m$ defined previously,

$$\begin{aligned}
|||\mathbf{w}|||_p &= \sum_{k=0}^{m-1} \mathtt{pbinom}(k, m-1, p)|||\mathbf{w}|||_{k+1} \\
&= m\sum_{k=0}^{m-1} \mathtt{pbinom}(k, m-1, p)\frac{\tau_m(d,k+1)}{k+1} \\
&= m\sum_{k=0}^{m-1} \binom{m-1}{k} p^k(1-p)^{m-k-1}\frac{1}{k+1}\tau_m(d,k+1) \\
&= \frac{1}{p}\sum_{k=0}^{m-1} \binom{m}{k+1} p^{k+1}(1-p)^{m-(k+1)}\tau_m(d,k+1) \\
&= \frac{1}{p}\sum_{k=0}^{m} \binom{m}{k} p^{k}(1-p)^{m-k}\tau_m(d,k).
\end{aligned}$$

Thus for all $d \in [\![m]\!]$ and any subset T such that $|T| = d$, we have shown

$$|||\mathbf{z}|||_p^* \geq p\Big(\sum_{k=0}^{m} \mathtt{pbinom}(k, m, p)\tau_m(d,k)\Big)^{-1}\|\mathbf{z}[T]\|_1.$$

Next, we will show

$$\sum_{k=0}^{m} \mathtt{pbinom}(k, m, p)\tau_m(d,k) \leq \tau_m(d, pm).$$

To this end, let us first notice that the LHS quantity is a binomial average of $\tau_m(d,k)$ with respect to k. By construction, $\tau_m(d,.)$ is piecewise linear. Furthermore, $\tau_m(d,.)$ is also concave by Lemma 13. Now let Y be a random variable having the binomial distribution with parameters m and p. By Jensen's inequality,

$$\sum_{k=0}^{m} \mathtt{pbinom}(k, m, p)\tau_m(d,k) = \mathbb{E}\tau_m(d,Y) \leq \tau_m(d, \mathbb{E}Y) = \tau_m(d, mp).$$

In particular, if $d = 1$, it is easy to see that $\tau_m(d, mp) = p$. So

$$p\left(\max_{T\subset[\![m]\!],|T|=1}\Big(\sum_{k=0}^{m} \mathtt{pbinom}(k, m, p)\tau_m(|T|,k)\Big)^{-1}\|\mathbf{z}[T]\|_1\right) \geq \|\mathbf{z}\|_\infty.$$

On the other hand, by Lemma 9,

$$\tau_m(d, pm) \leq \sqrt{\frac{d}{m}}\sqrt{pm} = \sqrt{pd}.$$

Therefore

$$p\Big(\sum_{k=0}^{m} \texttt{pbinom}(k,m,p)\tau_m(d,k)\Big)^{-1}\|\mathbf{z}[T]\|_1 \geq \sqrt{p}\frac{\|\mathbf{z}[T]\|_1}{\sqrt{d}},$$

and the proof is complete. ■

Lemma 11 *(Upper bounds for $\|\!|\mathbf{z}|\!\|_s^*$ and $\|\!|\mathbf{z}|\!\|_p^*$) Let $\mathbf{z} \in \mathbb{R}^m$.*

1. *For $s \in [\![m]\!]$,*
$$\|\!|\mathbf{z}|\!\|_s^* \leq \max_{|S|=s} \|\mathbf{z}[S]\|_2.$$
2. *For $p \in (0,1)$,*
$$\|\!|\mathbf{z}|\!\|_p^* \leq \max_{|S|=k} \|\mathbf{z}[S]\|_2,$$
where $k = \lceil p(m-1)+1 \rceil$.

Proof To establish the upper bound, we will use the equivalent formulation of $\|\!|.|\!\|_s^*$ in Corollary 5. For $S \subset [\![m]\!]$ of size s, as in Corollary 5, let $\mathbf{E}_S = \mathbf{I}[S,]/\binom{m-1}{s-1}$ where $\mathbf{I} \in \mathbb{R}^{m\times m}$ is the identity matrix. If we set $\mathbf{y}_S = \mathbf{z}[S]$, then $\sum_{|S|=s} \mathbf{E}_S^T \mathbf{y}_S = \mathbf{z}$ and so $\{\mathbf{y}_S\}$ is feasible. Therefore

$$\|\!|\mathbf{z}|\!\|_s^* \leq \max_{|S|=s} \|\mathbf{z}[S]\|_2.$$

The upperbound of $\|\!|\mathbf{z}|\!\|_p^*$ follows from the inequality $\|\!|\mathbf{z}|\!\|_p^* \leq \|\!|\mathbf{z}|\!\|_k^*$ for $k = \lceil p(m-1)+1 \rceil$ by the second part of Lemma 8. ■

Corollary 8 *(1-sparse vectors) Let $\mathbf{z} = (z, 0, \ldots, 0)^T \in \mathbb{R}^m$. We have*

$$\|\!|\mathbf{z}|\!\|_s^* = \|\!|\mathbf{z}|\!\|_p^* = |z|.$$

Proof These are direct consequences of Lemma 10 and Lemma 11. ■

Corollary 9 *(All-constant vectors) Let $\mathbf{z} \in \mathbb{R}^m$ be such that $\mathbf{z}[i] = z$ for all $i \in [\![m]\!]$. We have*

1. $\|\!|\mathbf{z}|\!\|_s^* = \sqrt{s}|z|$.
2. $\|\!|\mathbf{z}|\!\|_p^* = mp\Big(\sum_{k=0}^m \texttt{pbinom}(k,m,p)\sqrt{k}\Big)^{-1}|z|$.

Proof First of all, note that $L_m(m,k) = k$ and $\mathbb{E}\sqrt{L_m(m,k)} = \sqrt{k}$. Thus by Lemma 10 and 11, we have

$$\|\!|\mathbf{z}|\!\|_s^* = \sqrt{s}|z|.$$

So the first part of the claim is verified. Next, by Lemma 10,

$$|||\mathbf{z}|||_p^* \geq mp\Big(\sum_{k=0}^{m} \texttt{pbinom}(k,m,p)\sqrt{k}\Big)^{-1}|z|.$$

On the other hand, for S such that $|S| = s$, we can define

$$\mathbf{y}_S = \frac{mp}{\sqrt{s}}\Big(\sum_{k=0}^{m} \texttt{pbinom}(k,m,p)\sqrt{k}\Big)^{-1}(z,\ldots,z)^T \in \mathbb{R}^s,$$

For notation simplicity, let $c = \frac{1}{mp}\Big(\sum_{k=0}^{m} \texttt{pbinom}(k,m,p)\sqrt{k}\Big)$. As in Corollary 5, for $S \subset [\![m]\!]$, let $\mathbf{E}_S = \mathbf{I}[S,]/\binom{m-1}{|S|-1}$. For $i \in [\![m]\!]$, we have

$$\begin{aligned}\sum_{k=0}^{m-1} \texttt{pbinom}(k;m-1,p) \sum_{|S|=k+1} (\mathbf{E}_S^T \mathbf{y}_S)[i] &= c^{-1}\sum_{k=0}^{m-1} \texttt{pbinom}(k;m-1,p)\frac{1}{\sqrt{k+1}} \\ &= c^{-1}\frac{z}{mp}\sum_{k=0}^{m} \texttt{pbinom}(k;m,p)\sqrt{k} = z.\end{aligned}$$

Thus by Corollary 5,

$$|||\mathbf{z}|||_p^* \leq \max_S \|\mathbf{y}_S\|_2 = mp\Big(\sum_{k=0}^{m} \texttt{pbinom}(k,m,p)\sqrt{k}\Big)^{-1}|z|,$$

and the proof is complete. ■

Lemma 12 *(Convexity of $|||\mathbf{z}|||_k$) Let $\mathbf{z} \in \mathbb{R}^m$, where $m \geq 3$. For $k \in [\![m-2]\!]$, we have the following inequality*

$$|||\mathbf{z}|||_k + |||\mathbf{z}|||_{k+2} \geq 2|||\mathbf{z}|||_{k+1}. \tag{28}$$

Proof We will first show that the claim is true for $k = m-2$. Notice that in this case $|||\mathbf{z}|||_{k+2} = |||\mathbf{z}|||_m = \|\mathbf{z}\|_2$. If $\|\mathbf{z}\|_2 = 0$, the inequality (28) is trivially true. Now suppose $\|\mathbf{z}\|_2 > 0$, dividing both sides of the inequality by $\|\mathbf{z}\|_2$, we have

$$\binom{m-1}{m-3}^{-1}\sum_{|S|=m-2}\frac{\|\mathbf{z}[S]\|_2}{\|\mathbf{z}\|_2} + 1 \geq 2\binom{m-1}{m-2}^{-1}\sum_{|S|=m-1}\frac{\|\mathbf{z}[S]\|_2}{\|\mathbf{z}\|_2}.$$

Now let $\mathbf{x} = (x_1,\ldots,x_m)^T \in \mathbb{R}^m$ be such that $x_i = \mathbf{z}[i]^2/\|\mathbf{z}\|_2^2$. It suffices to show

$$\sum_{|S|=m-2}\left(\sum_{i\in S} x_i\right)^{1/2} + \frac{(m-1)(m-2)}{2} \geq (m-2)\sum_{i=1}^{m}\sqrt{1-x_i}, \tag{29}$$

for all $\mathbf{x} \geq 0$ entry-wise such that $\sum_i x_i = 1$. We will now prove the above inequality by induction on m. First of all, notice that for the base case where $m = 3$, we need to show:

$$\sqrt{x_1} + \sqrt{x_2} + \sqrt{x_3} + 1 \geq \sqrt{1 - x_1} + \sqrt{1 - x_2} + \sqrt{1 - x_3},$$

with the constraints $x_i \geq 0$ and $x_1 + x_2 + x_3 = 1$. For fixed x_3, let

$$f(x_1) = \sqrt{x_1} + \sqrt{1 - x_1 - x_3} + \sqrt{x_3} + 1 - \sqrt{x_1 + x_3} - \sqrt{1 - x_1} - \sqrt{1 - x_3}.$$

We will show that $f(x_1)$ is minimized at $x_1 = 0$ or $x_1 = 1 - x_3$. Suppose now $x_1 > 0$. Taking derivative with respect to x_1:

$$f'(x_1) = \frac{1}{2}\left(\frac{1}{\sqrt{x_1}} - \frac{1}{\sqrt{1 - x_1 - x_3}} - \frac{1}{\sqrt{x_1 + x_3}} + \frac{1}{\sqrt{1 - x_1}}\right).$$

Let $l(x_1) = \frac{1}{\sqrt{x_1}} - \frac{1}{\sqrt{x_1 + x_3}}$. Note that $f'(x_1) = \frac{1}{2} l(x_1) - \frac{1}{2} l(1 - x_3 - x_1)$. Now we have

$$l'(x_1) = \frac{1}{2}(x_1 + x_3)^{-3/2} - \frac{1}{2} {x_1}^{-3/2}.$$

So $l(x_1)$ is decreasing on $(0, 1 - x_3)$ and by symmetry the function $l(1 - x_3 - x_1)$ is increasing on $(0, 1 - x_3)$. On the other hand, since $\lim_{x_1 \downarrow 0^+} l(x_1) = +\infty$ and $\lim_{x_1 \downarrow 0^+} l(1 - x_3 - x_1) = -\infty$, we know that $f'(x_1) > 0$ on $(0, \frac{1 - x_3}{2})$ and < 0 on $(\frac{1 - x_3}{2}, 1 - x_3)$. Thus, the minimum of f can only be attained at the boundaries, i.e., $x_1 = 0$ or $x_1 = 1 - x_3$. In either case we have

$$\begin{aligned} &\sqrt{x_1} + \sqrt{x_2} + \sqrt{x_3} + 1 - \sqrt{1 - x_1} - \sqrt{1 - x_2} - \sqrt{1 - x_3} \\ \geq &\sqrt{x_2} + \sqrt{x_3} - \sqrt{1 - x_2} - \sqrt{1 - x_3} = 0, \end{aligned}$$

as $x_2 + x_3 = 1$. So we establish (29) for $m = 3$.

Suppose (29) is also true for $m = n - 1$. For $m = n$, similar to the $m = 3$ case, for fixed $x_3, \ldots, x_n$, define

$$f(x_1) = \sum_{|S| = n-2} \Big(\sum_{i \in S} x_i \Big)^{1/2} + \frac{(n-1)(n-2)}{2} - (n-2) \sum_{i=1}^{n} \sqrt{1 - x_i},$$

subject to $x_i \geq 0$ and $\sum_i x_i = 1$. Again, we will show f attains its minimum at either $x_1 = 0$ or $x_1 = 1 - \sum_{i=3}^n x_i$. Notice that

$$\begin{aligned} \sum_{|S| = n-2} \Big(\sum_{j \in S} x_j \Big)^{1/2} = &\sum_{|S| = n-3, 1, 2 \notin S} \Big(x_1 + \sum_{j \in S} x_j \Big)^{1/2} + \sum_{|S| = n-4, 1, 2 \notin S} \Big(x_1 + x_2 + \sum_{j \in S} x_j \Big)^{1/2} \\ &+ \sum_{|S| = n-3, 1, 2 \notin S} \Big(x_2 + \sum_{j \in S} x_j \Big)^{1/2} + \Big(\sum_{j=3}^{n} x_j \Big)^{1/2} \\ = &\sum_{i=3}^{n} \Big(x_1 + \sum_{j=3}^{n} x_j - x_i \Big)^{1/2} + \sum_{3 \leq i < j \leq n} (1 - x_i - x_j)^{1/2} \\ &+ \sum_{i=3}^{n} (1 - x_1 - x_i)^{1/2} + \Big(\sum_{j=3}^{n} x_j \Big)^{1/2}. \end{aligned}$$

In addition,

$$\sum_{i=1}^{n}(1-x_i)^{1/2} = (1-x_1)^{1/2} + \Big(x_1 + \sum_{j=3}^{n} x_j\Big)^{1/2} + \sum_{i=3}^{n}(1-x_i)^{1/2}.$$

Taking derivative with respect to x_1,

$$\begin{aligned} f'(x_1) =& \frac{1}{2}\Big(\sum_{i=3}^{n}\Big(x_1 + \sum_{j=3}^{n} x_j - x_i\Big)^{-1/2} - \sum_{i=3}^{n}(1-x_1-x_i)^{-1/2} \\ &+ (n-2)(1-x_1)^{-1/2} - (n-2)\Big(x_1 + \sum_{i=3}^{n} x_i\Big)^{-1/2}\Big). \end{aligned}$$

Now let

$$l(x_1) = \sum_{i=3}^{n}\Big(x_1 + \sum_{j=3}^{n} x_j - x_i\Big)^{-1/2} - (n-2)\Big(x_1 + \sum_{j=3}^{n} x_j\Big)^{-1/2}.$$

So $2f'(x_1) = l(x_1) - l(1 - \sum_{i=3}^{n} x_i - x_1)$. Again

$$l'(x_1) = -\frac{1}{2}\sum_{i=3}^{n}\Big(x_1 + \sum_{j=3}^{n} x_j - x_i\Big)^{-3/2} + \frac{n-2}{2}\Big(x_1 + \sum_{j=3}^{n} x_j\Big)^{-3/2}.$$

It is easy to see that $l'(x_1) < 0$ and so $l(x_1)$ is decreasing on $(0, 1 - \sum_{i=3}^{n} x_i - x_1)$. On the other hand $\lim_{x_1 \downarrow 0^+} l(x_1) = +\infty$. By symmetry $f'(x_1) > 0$ on $\big(0, \frac{1}{2}(1 - \sum_{i=3}^{n} x_i - x_1)\big)$ and < 0 on $\big(0, \frac{1}{2}(1 - \sum_{i=3}^{n} x_i - x_1)\big)$. So f attains its minimum at $x_1 = 0$ or $x_1 = 1 - \sum_{i=3}^{n} x_i$. Hence we have

$$\begin{aligned} &\sum_{|S|=n-2}\Big(\sum_{i\in S} x_i\Big)^{1/2} + \frac{(n-1)(n-2)}{2} - (n-2)\sum_{i=1}^{n}(1-x_i)^{1/2} \\ \geq& \Big(\sum_{|S|=n-3, 1\notin S} + \sum_{|S|=n-2, 1\notin S}\Big)\Big(\sum_{j\in S} x_j\Big)^{1/2} + \frac{(n-2)(n-3)}{2} - (n-2)\sum_{i=2}^{n}(1-x_i)^{1/2}. \quad (30) \end{aligned}$$

By the induction assumption that (29) holds when $m = n-1$, we have

$$\sum_{|S|=n-3, 1\notin S}\Big(\sum_{j\in S} x_j\Big)^{1/2} + \frac{(n-2)(n-3)}{2} \geq (n-3)\sum_{i=2}^{n}(1-x_i)^{1/2}.$$

Thus (30) is greater than or equal to

$$\begin{aligned} &-\frac{(n-2)(n-3)}{2} + \sum_{|S|=n-2, 1\notin S}\Big(\sum_{j\in S} x_j\Big)^{1/2} + \frac{(n-1)(n-2)}{2} - (n-2) - \sum_{i=2}^{n}(1-x_i)^{1/2} \\ =& \sum_{|S|=n-2, 1\notin S}\Big(\sum_{j\in S} x_j\Big)^{1/2} - \sum_{i=2}^{n}(1-x_i)^{1/2} = \sum_{i=2}^{n}(1-x_i)^{1/2} - \sum_{i=2}^{n}(1-x_i)^{1/2} = 0. \end{aligned}$$

Thus we have verified the claim that (29) and hence (28) holds for $k = m-2$ for all $m \geq 3$. To establish the case for general $1 \leq k \leq m-2$, we again perform induction on the (m,k)-tuple. Note that the base case $m = 3$ and $k = 1$ has been previously proved. Suppose (28) holds for $m = n-1$ and $1 \leq k \leq n-3$. Now consider $m = n$ and $1 \leq k < n-2$. Notice that

$$\begin{aligned}
|||\mathbf{z}|||_k =& \frac{1}{n-k}\binom{n-1}{k-1}^{-1} \sum_{|T|=n-1} \sum_{|S|=k, S\subset T} \|\mathbf{z}[S]\|_2 \\
=& (n-1)\binom{n-2}{k-1}^{-1} \sum_{|T|=n-1} \sum_{|S|=k, S\subset T} \|\mathbf{z}[S]\|_2 \\
=& (n-1) \sum_{|T|=n-1} |||\mathbf{z}[T]|||_k.
\end{aligned}$$

By the induction assumption, for all T such that $|T| = n-1$, we have:

$$|||\mathbf{z}[T]|||_k + |||\mathbf{z}[T]|||_{k+2} \geq 2|||\mathbf{z}[T]|||_{k+1}.$$

Therefore

$$|||\mathbf{z}|||_k + |||\mathbf{z}|||_{k+2} - 2|||\mathbf{z}|||_{k+1} = (n-1) \sum_{|T|=n-1} (|||\mathbf{z}[T]|||_k + |||\mathbf{z}[T]|||_{k+2} - 2|||\mathbf{z}[T]|||_{k+1}) \geq 0.$$

Thus the claim also holds for $m = n$ and $1 \leq k < n-2$, completing the proof. ■

A.6 Miscellaneous

Lemma 13 *(Concavity of $\mathbb{E}\sqrt{L_m(d,k)}$) Let $d \in [\![m]\!]$. For $k \in [\![m-2]\!]$, we have*

$$\mathbb{E}\sqrt{L_m(d,k)} + \mathbb{E}\sqrt{L_m(d,k+2)} \leq 2\mathbb{E}\sqrt{L_m(d,k+1)}. \tag{31}$$

where the geometric random variable $L_m(d,k)$ is defined as in Definition 3.

Proof Suppose we are now sampling without replacement from a pool of numbers with d 1's and $m-d$ 0's. For $i \in [\![m]\!]$, denote by $X_i \in \{0,1\}$ the i-th outcome. It is easy to see that $L_m(d,k)$ and $\sum_{i=1}^k X_i$ have the same distribution. To show (31), it suffices to prove the following conditional expectation inequality:

$$\sqrt{L_m(d,k)} + \mathbb{E}[\sqrt{L_m(d,k+2)} \mid L_m(d,k)] \leq 2\mathbb{E}[\sqrt{L_m(d,k+1)} \mid L_m(d,k)]$$

Note that the above inequality follows if for all $0 \leq a \leq \min(d,k)$:

$$\sqrt{a} + \mathbb{E}\sqrt{a + X_{k+1} + X_{k+2}} \leq 2\mathbb{E}\sqrt{a + X_{k+1}}$$

It is easy to see that

$$\begin{aligned}
\mathbb{E}\sqrt{a + X_{k+1}} &= \frac{d-a}{m-k}\sqrt{a+1} + (1 - \frac{d-a}{m-k})\sqrt{a}. \\
\mathbb{E}\sqrt{a + X_{k+1} + X_{k+2}} &= \frac{d-a}{m-k} \times \frac{d-a-1}{m-k-1}\sqrt{a+2} + 2 \times \frac{d-a}{m-k} \times \frac{m-k-(d-a)}{m-k-1}\sqrt{a+1} \\
&\quad + \frac{m-k-(d-a)}{m-k} \times \frac{m-k-(d-a)-1}{m-k-1}\sqrt{a}.
\end{aligned}$$

By elementary algebra, it can be shown that

$$\begin{aligned}&2\mathbb{E}\sqrt{a+X_{k+1}}-\sqrt{a}-\mathbb{E}\sqrt{a+X_{k+1}+X_{k+2}}\\=&\frac{d-a}{m-k}\times\frac{d-a-1}{m-k-1}\times(2\sqrt{a+1}-\sqrt{a+2}-\sqrt{a})\geq 0,\end{aligned}$$

The inequality follows since $f(x)=\sqrt{x}$ is a concave function. Thus the proof is complete. ■

Lemma 14 *Let $\mathbf{x}(t)=(x_1(t),\ldots,x_m(t))^T\in\mathbb{R}^m$ be an m-dimensional function on $[0,\epsilon)$ such that: (1) $x_1(0)=1$ and for all $i\geq 2$, $x_i(0)=0$; (2) The derivative $\dot{x}_i(t)$ exists and is bounded for all $t\in(0,\epsilon)$. We have*

$$\lim_{t\downarrow 0^+}\frac{\|\mathbf{x}(t)\|_2-\|\mathbf{x}(0)\|_2}{t}=\lim_{t\downarrow 0^+}\dot{x}_1(t).$$

Proof

$$\begin{aligned}\lim_{t\downarrow 0^+}\frac{\|\mathbf{x}(t)\|_2-\|\mathbf{x}(0)\|_2}{t}&=\lim_{t\downarrow 0^+}\frac{(\sum_{i=1}^m x_i^2(t))^{1/2}-1}{t}\\&=\lim_{t\downarrow 0^+}\frac{\sum_{i=1}^m x_i^2(t)-1}{t}\left((\sum_{i=1}^m x_i^2(t))^{1/2}+1\right)^{-1}\\&=\frac{1}{2}\lim_{t\downarrow 0^+}\frac{\sum_{i=1}^m x_i^2(t)-1}{t}\\&=\frac{1}{2}\left(\lim_{t\downarrow 0^+}\frac{x_1^2(t)-1}{t}+\sum_{i=2}^m\lim_{t\downarrow 0^+}\frac{x_i^2(t)}{t}\right)\\&=\frac{1}{2}\left(\lim_{t\downarrow 0^+}\frac{x_1(t)-1}{t}(x_1(t)+1)+\sum_{i=2}^m\lim_{t\downarrow 0^+}\frac{x_i^2(t)}{t}\right)\\&=\lim_{t\downarrow 0^+}\frac{x_1(t)-1}{t}+\frac{1}{2}\sum_{i=2}^m\lim_{t\downarrow 0^+}\frac{x_i^2(t)}{t}.\end{aligned}$$

By mean value theorem, for each $t\in(0,\epsilon)$, there exists $\delta_t\in(0,t)$ such that $x_1(t)-1=\dot{x}_1(\delta_t)t$. Thus the first term simply becomes $\lim_{t\downarrow 0^+}\dot{x}_1(t)$. By the same argument, for each $i\in\{2,\ldots,m\}$, $x_i(t)=\dot{x}_i(\delta_t)t$ for some $\delta_t\in(0,t)$. Since $\dot{x}_i(t)$ is bounded, we have

$$\lim_{t\downarrow 0^+}\frac{x_i^2(t)}{t}=\lim_{t\downarrow 0^+}\dot{x}_i(\delta_t)^2t=0.$$

Therefore the claim is verified. ■

Lemma 15 *Let $\mathbf{x}(t)=(x_1(t),\ldots,x_m(t))^T\in\mathbb{R}^m$ be an m-dimensional function on $[0,\epsilon)$ such that: (1) $x_i(0)=0$ for all $i=1,\ldots,m$; (2) The derivative $\dot{x}_i(t)$ exists for all $t\in(0,\epsilon)$. We have*

$$\lim_{t\downarrow 0^+}\frac{\|\mathbf{x}(t)\|_2}{t}=\|\lim_{t\downarrow 0^+}\dot{\mathbf{x}}(t)\|_2.$$

Proof

$$\lim_{t\downarrow 0^+}\frac{\|\mathbf{x}(t)\|_2}{t}=\lim_{t\downarrow 0^+}\left(\sum_{i=1}^m(\frac{x_i(t)}{t})^2\right)^{1/2}=\left(\sum_{i=1}^m(\lim_{t\downarrow 0^+}\frac{x_i(t)}{t})^2\right)^{1/2}=\|\lim_{t\downarrow 0^+}\dot{\mathbf{x}}(t)\|_2.$$

■

Lemma 16 *Let* $\mathbf{a}=(a_1,\ldots,a_m)^T\in\mathbb{R}^m$ *where* $a_1\neq 0$ *and* $\mathbf{x}(t)=(x_1(t),\ldots,x_m(t))^T\in\mathbb{R}^m$ *be an* m*-dimensional function on* $[0,\epsilon)$ *such that: (1)* $x_1(0)=1$ *and for all* $i\geq 2$, $x_i(0)=0$*; (2) The derivative* $\dot{x}_i(t)$ *exists and is bounded for all* $t\in(0,\epsilon)$*. We have*

$$\lim_{t\downarrow 0^+}\frac{|\mathbf{a}^T\mathbf{x}(t)|-|a_1|}{t}=|a_1|\lim_{t\downarrow 0^+}\dot{x}_1(t)+\boldsymbol{sgn}(a_1)\sum_{i=2}^m a_i\lim_{t\downarrow 0^+}\dot{x}_i(t).$$

Proof Without loss of generality, assume $a_1>0$. Since $x_1(0)=1$ and for all $i\geq 2$, $x_i(0)=0$, by continuity, for sufficiently small t, we have

$$\frac{|\mathbf{a}^T\mathbf{x}(t)|-|a_1|}{t}=\frac{|a_1x_1(t)+\sum_{i=2}^m a_ix_i(t)|-a_1}{t}=\frac{a_1x_1(t)-a_1+\sum_{i=2}^m a_ix_i(t)}{t}.$$

Therefore, by the same argument in the proof of Lemma 14,

$$\begin{aligned}\lim_{t\downarrow 0^+}\frac{|\mathbf{a}^T\mathbf{x}(t)|-|a_1|}{t}&=\lim_{t\downarrow 0^+}\frac{a_1x_1(t)-a_1}{t}+\lim_{t\downarrow 0^+}\sum_{i=2}^m\frac{a_ix_i(t)}{t}\\&=a_1\lim_{t\downarrow 0^+}\dot{x}_1(t)+\sum_{i=2}^m a_i\lim_{t\downarrow 0^+}\dot{x}_i(t).\end{aligned}$$

■

Lemma 17 *Let* $\mathbf{a}=(a_1,\ldots,a_m)^T\in\mathbb{R}^m$ *and* $\mathbf{x}(t)=(x_1(t),\ldots,x_m(t))^T\in\mathbb{R}^m$ *be an* m*-dimensional function on* $[0,\epsilon)$ *such that: (1)* $x_i(0)=0$ *for all* $i=1,\ldots,m$*; (2) The derivative* $\dot{x}_i(t)$ *exists for all* $t\in(0,\epsilon)$*. We have*

$$\lim_{t\downarrow 0^+}\frac{|\mathbf{a}^T\mathbf{x}(t)|}{t}=\left|\sum_{i=1}^m a_i\lim_{t\downarrow 0^+}\dot{x}_i(t)\right|.$$

Proof The proof is similar to that of Lemma 15. ■

Appendix B. Additional Simulations

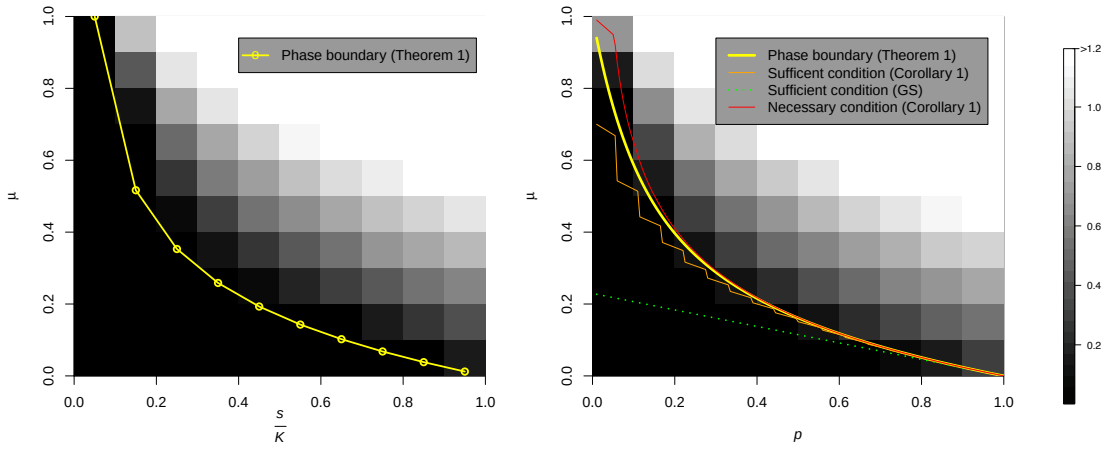

Figure B.1: Local recovery errors for the s-sparse Gaussian model (Left) and the Bernoulli(p)-Gaussian model (Right), with the number of dictionary atoms $K = 20$. See Figure 1 for simulation details.

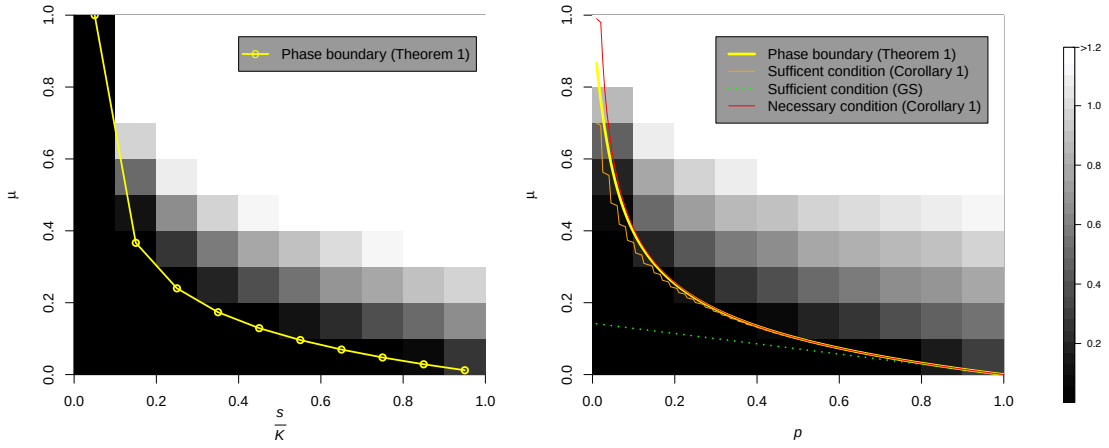

Figure B.2: Local recovery errors for the s-sparse Gaussian model (Left) and the Bernoulli(p)-Gaussian model (Right), with the number of dictionary atoms $K = 50$. See Figure 1 for simulation details.

References

P.A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, and Praneeth Netrapalli. Learning sparsely used overcomplete dictionaries via alternating minimization. *arXiv preprint arXiv:1310.7991v2*, 2014a.

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon. Learning sparsely used overcomplete dictionaries. In *Conference on Learning Theory*, pages 123–137, 2014b.

Alekh Agarwal, Animashree Anandkumar, and Praneeth Netrapalli. Exact recovery of sparsely used overcomplete dictionaries. *arXiv preprint arXiv:1309.1952*, 2014c.

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, 2006a.

Michal Aharon, Michael Elad, and Alfred M. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006b.

Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization–provably. In *Proceedings of the 44th Annual ACM Symposium on Theory of Computing*, pages 145–162, 2012a.

Sanjeev Arora, Rong Ge, Ankur Moitra, and Sushant Sachdeva. Provable ICA with unknown gaussian noise, with implications for gaussian mixtures and autoencoders. In *Advances in Neural Information Processing Systems*, pages 2375–2383, 2012b.

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

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

Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data*. Springer, 2011.

Emanuel Candes and Terrence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3): 287–314, 1994.

David L. Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences of the United States of America*, 100(5):2197–2202, 2003.

Michael Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer Science & Business Media, 2010.

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.

J-J Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Transactions on Information Theory*, 50(6):1341–1344, 2004.

Quan Geng, Huan Wang, and John Wright. On the local correctness of ℓ^1-minimization for dictionary learning. *arXiv preprint arXiv:1101.5672*, 2011.

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-factorisation via ℓ_1-minimisation. *IEEE Transactions on Information Theory*, 56(7):3523–3539, 2010.

Rémi Gribonval, Rodolphe Jenatton, Francis Bach, Martin Kleinsteuber, and Matthias Seibert. Sample complexity of dictionary learning and other matrix factorizations. *arXiv preprint arXiv:1312.3790*, 2013.

Rémi Gribonval, Rodolphe Jenatton, and Francis Bach. Sparse and spurious: dictionary learning with noise and outliers. *IEEE Transactions on Information Theory*, 61(11): 6298–6319, 2015.

Roger Grosse, Rajat Raina, Helen Kwong, and Andrew Ng. Shift-invariance sparse coding for audio classification. *arXiv preprint arXiv:1206.5241*, 2012.

Christopher 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.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010.

Julien Mairal, Francis Bach, and Jean Ponce. Sparse modeling for image and vision processing. *Foundations and Trends in Computer Graphics and Vision*, 8(2-3):85–283, 2014.

Stephane Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 3rd edition, 2008.

Andreas Maurer and Massimiliano Pontil. k-dimensional coding schemes in Hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, 2010.

Nishant A. Mehta and Alexander G. Gray. Sparsity-based generalization bounds for predictive sparse coding. In *Proceedings of the 30th International Conference on Machine Learning*, pages 36–44, 2013.

Bruno Olshausen and David Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1997.

Gabriel Peyré. Sparse modeling of textures. *Journal of Mathematical Imaging and Vision*, 34(1):17–31, 2009.

Mark D. Plumbley. Dictionary learning for ℓ_1-exact sparse coding. In *Independent Component Analysis and Signal Separation*, pages 406–413. Springer, 2007.

Ben Recht, Christopher Re, Joel Tropp, and Victor Bittorf. Factoring nonnegative matrices with linear programs. In *Advances in Neural Information Processing Systems*, pages 1214–1222, 2012.

Ron Rubinstein, Alfred M. Bruckstein, and Michael Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

Karin Schnass. On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD. *Applied and Computational Harmonic Analysis*, 37(3):464–491, 2014.

Karin Schnass. Local identification for overcomplete dictionaries. *arXiv preprint arXiv:1401.6354v1*, 2015.

Daniel Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. *arXiv preprint arXiv:1206.5882*, 2012.

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, 2017.

Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

Daniel Vainsencher, Shie Mannor, and Alfred M. Bruckstein. The sample complexity of dictionary learning. *Journal of Machine Learning Research*, 12:3259–3281, 2011.

Yu Wang, Siqi Wu, and Bin Yu. Global identifiability of complete dictionary learning through ℓ_1-minimization. *Manuscript*.

Siqi Wu, Antony Joseph, Ann S. Hammonds, Susan E. Celniker, Bin Yu, and Erwin Frise. Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. *Proceedings of the National Academy of Sciences of the United States of America*, 113(16):4290–4295, 2016.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68:49–57, 2006.

Michael Zibulevsky, Barak Pearlmutter, et al. Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation*, 13(4):863–882, 2001.