Approximate Subspace-Sparse Recovery in the Presence of Corruptions via ℓ_1-Minimization

Ehsan Elhamifar, *Member, IEEE,* Mahdi Soltanolkotabi, *Member, IEEE,* and S. Shankar Sastry, *Fellow, IEEE*

***Abstract*—High-dimensional data often lie in low-dimensional subspaces corresponding to different classes they belong to. Finding sparse representations of data points in a dictionary built from the collection of data helps to uncover the low-dimensional subspaces and, as a result, address important problems such as compression, clustering, classification, subset selection and more. However, an important challenge related to real-world datasets is that the collection of data is often corrupted by measurement or process noise. In this paper, we address the problem of recovering sparse representations for noisy data points in a dictionary whose columns correspond to noisy data points lying close to a union of subspaces. We consider a constrained ℓ_1-minimization program and study conditions under which the solution of the optimization recovers a representation of a noisy point as a linear combination of a few noisy points from the same subspace. Our framework is based on a novel generalization of the null-space property to the setting where data lie in multiple subspaces, the number of data points in each subspace exceeds the dimension of the subspace, and all data points are corrupted by noise. We do not impose any randomness assumption on the arrangement of subspaces or distribution of data points in each subspace. We show that, under appropriate conditions that depend on relationships among data points within and between subspaces, the solution of our proposed optimization satisfies a desired approximate subspace-sparse recovery. More specifically, we show that a noisy data point, close to one of the subspaces, will be reconstructed using data points from the same subspace with a small error and that the coefficients corresponding to data points in other subspaces will be sufficiently small.**

***Index Terms*—Low-dimensional subspaces, sparse representation, noisy data points, ℓ_1-minimization, subspace incoherence, subspace inradius, approximate recovery.**

I. Introduction

HIGH-DIMENSIONAL datasets are ubiquitous in many areas of science and engineering, such as signal and image processing, computer vision, information retrieval, bio and health informatics, energy systems, robotics and more. Real-world data, however, often lie close to low-dimensional subspaces instead of being uniformly distributed in the high-dimensional ambient space [1], [2], [3], [4], [5], [6]. Exploiting and recovering the low-dimensional structures in data, in fact, is the key to efficiently address a variety of important problems such as clustering [6], [7], [8], [9], [10], [11], classification [12], [13], compression [4], [14], [15], [16], subset selection [17], [18], [19], visualization as well as other applications [20], [21], [22], [23].

E. Elhamifar is currently a postdoctoral scholar in the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, USA. E-mail: ehsan@eecs.berkeley.edu.

M. Soltanolkotabi is currently a postdoctoral scholar in the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, USA. E-mail: mahdisol@berkeley.edu.

S. Shankar Sastry is a professor of the Electrical Engineering and Computer Sciences Department, University of California at Berkeley, USA. E-mail: sastry@eecs.berkeley.edu.

Sparse representation techniques provide effective tools to exploit and uncover the low-dimensional structures in datasets [24], [25], [26]. More specifically, given a measurement $\boldsymbol{y} \in \mathbb{R}^n$ and a dictionary or a sensing matrix $\boldsymbol{A} \in \mathbb{R}^{n \times N}$, which has a nontrivial null-space, the goal of sparse recovery is to find a representation $\boldsymbol{c} \in \mathbb{R}^N$ of $\boldsymbol{y}$ as a linear combination of the columns of $\boldsymbol{A}$, such that $\boldsymbol{c}$ has only a few nonzero coefficients. A computationally efficient method to achieve this goal is to solve the ℓ_1-minimization program

$$\min \|\boldsymbol{c}\|_1 \quad \text{s.t.} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{c}. \tag{I.1}$$

In fact, $\|\boldsymbol{c}\|_1$, which is the sum of the absolute values of elements of $\boldsymbol{c}$, is the convex envelope of the cardinality of $\boldsymbol{c}$ and is known to recover sparse solutions, under appropriate conditions on the dictionary and the sparsity level [24], [25], [26], [27], [28].

Sparse representation-based methods can be divided into two categories, depending on the type of dictionaries being used. The first group of methods use fixed pre-defined dictionaries, such as the ones built from Wavelets, Fourier basis, Random Projections and so on [27], [29], [30]. The second group of methods, which form an important class of data analysis algorithms, use adaptive dictionaries built from the collection of data, where the columns of the dictionary $\boldsymbol{A}$ correspond to data points [6], [12], [31]. Under the assumption that the data points lie in a union of subspaces, with the number of data in each subspace being larger than the dimension of the subspace, a sparse representation of $\boldsymbol{y}$, ideally, corresponds to a subspace-sparse representation. In other words, $\boldsymbol{y}$ can be written as a linear combination of a few data points that lie in the same low-dimensional subspace. In fact, subspace-sparse recovery is the key requirement for the success of sparse representation-based clustering, classification, compression and subset selection algorithms, which has been the subject of recent studies in the literature [6], [11], [13], [17], [32], [33], [34]. One can show that when data points perfectly lie in subspaces, under appropriate conditions on the principal angles between subspaces and the distribution of data, the solution of ℓ_1-minimization perfectly recovers a subspace-sparse representation [6], [32], [33].

An important challenge related to real-world datasets is that data points are often corrupted by measurement or process noise. In other words, not only $\boldsymbol{y}$, but also all columns of the dictionary $\boldsymbol{A}$ are corrupted by noise. As a result, standard analysis tools related to the first group of sparse recovery methods,

in which the predefined dictionary $\boldsymbol{A}$ is uncorrupted while the measurement $\boldsymbol{y}$ is noisy, are not applicable [20], [35], [36], [37]. Therefore, there is a need to develop subspace-sparse recovery algorithms and study their theoretical guarantees in the setting where data lie in a union of subspaces and are corrupted by noise.

Recently, [38], [39] studied the problem of subspace-sparse recovery in the presence of noise using the unconstrained optimization program

$$\min \lambda\|\boldsymbol{c}\|_1 + \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{c}\|_{\ell_2}^2, \tag{I.2}$$

where the regularization parameter $\lambda > 0$ sets a trade-off between sparsity and reconstruction error objectives. It is proved that, when subspaces and/or data points are drawn randomly from appropriate distributions, under appropriate conditions on subspaces and data points and for certain values of λ, the solution of the above optimization recovers subspace-sparse representations for all data points.

Paper Contributions. In this paper, we study the problem of approximate subspace-sparse recovery in the presence of noise, where we do not impose any randomness assumption on the arrangement of subspaces or distribution of data points in each subspace. We assume that all data points are corrupted by Gaussian noise whose Euclidean norm is smaller than or equal to ε. Instead of the unconstrained minimization (I.2), we consider the constrained ℓ_1-minimization program

$$\min \|\boldsymbol{c}\|_1 \quad \text{s.t.} \quad \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{c}\|_{\ell_2} \leq \gamma\varepsilon, \tag{I.3}$$

where $\gamma > 0$ is a regularization parameter, which we determine in the paper.

We show that, under appropriate conditions on the data and subspaces, the solution of (I.3) satisfies the approximate subspace-sparse recovery property, i.e., 1) $\boldsymbol{y}$ will be reconstructed using data point from its underlying subspace with an error that is of the order of $O(\varepsilon)$; 2) coefficients corresponding to data points in other subspaces are sufficiently small, of the order of $O(\varepsilon)$. Our theoretical results relies on a novel generalization of the well-known null-space property, studied in conventional sparse recovery [40], [41], [42], [43], to the setting where 1) data lie in a union of subspaces, with the number of data points in each subspace typically being larger than the subspace dimension; 2) all data points are corrupted by noise.

Unlike conventional results on sparse recovery that assume only the measurement vector $\boldsymbol{y}$ is noise [44], [45], our work addresses the general framework where both the measurement and the dictionary columns are corrupted by noise, and data lie in a union of subspaces. In addition, our result contains, as a special case, existing result on sparse recovery in union of subspaces for the noise-free setting [6], [32], [33]. To the best of our knowledge, this is the first work analyzing the constrained ℓ_1-minimization program (I.3) for subspace-sparse recovery with noisy data. Finally, unlike state of the art, we impose no randomness randomness on data or subspaces. We allow for arbitrary arrangement of subspaces and arbitrary data points in each subspace. The only randomness assumption comes from the noise in the data, which we assume to be Gaussian.

Paper Organization. The organization of this paper is as follows. In Section II, we present the settings of our problem. We state the approximate subspace-sparse recovery problem and introduce appropriate definitions and notations. In Section III, we present our theoretical guarantees for our proposed constrained ℓ_1-minimization program. Finally, in Section IV, we conclude the paper and discuss open problems.

II. Problem Formulation and Main Results

In this section, we consider the problem of finding sparse representations for corrupted data points that lie close to a union of subspaces. Assume that we have L linear subspaces $\{\mathcal{S}_i\}_{i=1}^L$ in $\mathbb{R}^n$ of dimensions $\{d_i\}_{i=1}^L$. Let $\boldsymbol{X} \in \mathbb{R}^{n\times N}$ denote a matrix whose columns correspond to noise-free data points that lie in the union of the L subspaces. Without loss of generality, we assume that the columns of $\boldsymbol{X}$ have unit Euclidean norms. We denote by $\boldsymbol{X}_i \in \mathbb{R}^{n\times N_i}$ the N_i data points that lie in $\mathcal{S}_i$, hence $\sum_{i=1}^L N_i = N$. We can write

$$\boldsymbol{X} \triangleq \begin{bmatrix}\boldsymbol{X}_1 & \boldsymbol{X}_2 & \cdots & \boldsymbol{X}_L\end{bmatrix}\boldsymbol{\Gamma} \in \mathbb{R}^{n\times N}, \tag{II.1}$$

where $\boldsymbol{\Gamma} \in \mathbb{R}^{N\times N}$ is a permutation matrix, which is not necessarily known a priori. Given $\boldsymbol{x}$ that lies in one of the subspaces, the subspace-sparse recovery problem refers to the problem of finding a representation of $\boldsymbol{x}$ in the dictionary $\boldsymbol{X}$, as $\boldsymbol{x} = \boldsymbol{X}\boldsymbol{c}$, such that the nonzero coefficients of $\boldsymbol{c}$ correspond to a few data points that lie in the same subspace as that of $\boldsymbol{x}$. More specifically, considering the sparse optimization program

$$\boldsymbol{c}^* = \arg\min \|\boldsymbol{c}\|_{\ell_1} \quad \text{s.t.} \quad \boldsymbol{x} = \boldsymbol{X}\boldsymbol{c}, \tag{II.2}$$

one would like to have a few nonzero elements in $\boldsymbol{c}^*$ that correspond to data points lying in the same subspace of $\boldsymbol{x}$.

In real-world problems, however, data points often do not lie perfectly in subspaces, due to corruption by noise. Instead, they lie approximately close to a union of subspaces. In this paper, we address the problem of approximate subspace-sparse recovery in the presence of noise. More precisely, we assume that we have a collection of noisy data points $\boldsymbol{Y}_i \in \mathbb{R}^{n\times N_i}$ from each subspace $\mathcal{S}_i$, i.e.,

$$\boldsymbol{Y}_i = \boldsymbol{X}_i + \boldsymbol{Z}_i, \tag{II.3}$$

where $\boldsymbol{X}_i$ denotes the collection of noise-free data points, which lie at the intersection of $\mathcal{S}_i$ with the unit hypersphere, and $\boldsymbol{Z}_i$ denotes the random noise matrix, which has i.i.d elements drawn from the Gaussian distribution $\mathcal{N}(0, \frac{\epsilon^2}{n})$. As a result, each noise-free data point of unit Euclidean norm on each subspace is corrupted by a noise whose Euclidean norm is roughly less than or equal to ε, where

$$\varepsilon \triangleq \epsilon(1+\rho), \tag{II.4}$$

for a sufficiently small $\rho > 0$. We also assume that $\boldsymbol{x} \in \mathcal{S}_i$, which has unit Euclidean norm, is corrupted by a noise $\boldsymbol{z}$, which has i.i.d elements drawn from $\mathcal{N}(0, \frac{\epsilon^2}{n})$, giving rise to the noisy data point $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{z}$.

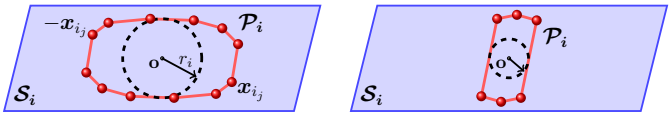

Fig. 1. Left: The subspace inradius associated with $\mathcal{S}_i$ is the radius of the largest Euclidean ball whose intersection with $\mathcal{S}_i$ is inscribed in the symmetrized convex hull of data points in $\mathcal{S}_i$. Right: When data are not well distributed in a subspace, i.e., they are close to a degenerate subspace, e.g., a line inside a plane, the subspace inradius decreases.

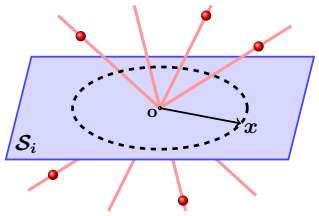

Fig. 2. The subspace incoherence associated with $\mathcal{S}_i$ is defined as the maximum inner product between an arbitrary vector of unit Euclidean norm in $\mathcal{S}_i$ and data points in other subspaces.

Remark II-A: Notice that for a Gaussian random vector $\boldsymbol{z} \in \mathbb{R}^n$ with i.i.d entries drawn from $\mathcal{N}(0, \frac{\varepsilon^2}{n})$, with high probability, we have $\|\boldsymbol{z}\|_{\ell_2} \leq \varepsilon$. For the sake of brevity, throughout the paper, we do not include explicitly the failure probability of $\|\boldsymbol{z}\|_{\ell_2} \leq \varepsilon$ in the probabilistic statements of our results.

For simplicity of notation, we denote $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{X}_i & \boldsymbol{X}_{-i}\end{bmatrix}$, where $\boldsymbol{X}_{-i}$ represents the collection of data points from all subspaces except $\mathcal{S}_i$. Similarly, we write $\boldsymbol{Y} = \begin{bmatrix}\boldsymbol{Y}_i & \boldsymbol{Y}_{-i}\end{bmatrix}$, where $\boldsymbol{Y}_{-i}$ denotes the collection of noisy data points from all subspaces except $\mathcal{S}_i$. We also use the convention $\boldsymbol{y} \in \mathcal{S}_i^\varepsilon$ to refer to a noisy data point that is the sum of a noise-free data point $\boldsymbol{x}$ in $\mathcal{S}_i$ with unit Euclidean norm and a noise $\boldsymbol{z}$ whose Euclidean norm is smaller than or equal to ε, i.e.,

$$\mathcal{S}_i^\varepsilon \triangleq \{\boldsymbol{y} \in \mathbb{R}^n : \ \boldsymbol{y} = \boldsymbol{x} + \boldsymbol{z}, \ \boldsymbol{x} \in \mathcal{S}_i, \ \|\boldsymbol{z}\|_{\ell_2} \leq \varepsilon\}. \quad \text{(II.5)}$$

Our goal is to find an approximate subspace-sparse representation, $\boldsymbol{c}^\top = \begin{bmatrix}\boldsymbol{c}_i^\top & \boldsymbol{c}_{-i}^\top\end{bmatrix}$, of a noisy data point $\boldsymbol{y}$ in the dictionary of corrupted data, $\boldsymbol{Y}$, as we define next.

Definition 2.1 (approximate subspace-sparse recovery): Consider a noisy data point $\boldsymbol{y}$ lying in $\mathcal{S}_i^\varepsilon$ and a noisy dictionary $\boldsymbol{Y}$, where the Euclidean norm of the noise on the its columns is less than or equal to ε. An approximate subspace-sparse recovery of $\boldsymbol{y}$ in $\boldsymbol{Y}$ corresponds to a representation $\boldsymbol{y} = \boldsymbol{Y}\boldsymbol{c}$, such that

1) $\boldsymbol{y}$ can be reconstructed, with high accuracy, using noisy data points from its own subspace, i.e.,

$$\|\boldsymbol{y} - \boldsymbol{Y}_i\boldsymbol{c}_i\|_{\ell_2} \leq O(\varepsilon); \quad \text{(II.6)}$$

2) the nonzero coefficients corresponding to noisy data points in other subspaces are sufficiently small, i.e.,

$$\|\boldsymbol{c}_{-i}\|_{\ell_1} \leq O(\varepsilon). \quad \text{(II.7)}$$

In order to achieve an approximate subspace-sparse representation for $\boldsymbol{y}$ in the dictionary $\boldsymbol{Y}$, in this paper, we consider the the constrained ℓ_1-minimization program

$$\min \|\boldsymbol{c}\|_{\ell_1} \quad \text{s. t.} \quad \|\boldsymbol{y} - \boldsymbol{Y}\boldsymbol{c}\|_{\ell_2} \leq \gamma\varepsilon, \quad \text{(II.8)}$$

where $\gamma > 0$ is a parameter that we determine in the paper. We investigate conditions on the data and subspaces under which the optimal solution of (II.8) achieves approximate subspace-sparse recovery.

The conditions that we derive depend on the inradius of convex bodies of the data in each subspace and the incoherence between subspaces, as we define next.

Definition 2.2 (subspace inradius): Let $\boldsymbol{X}_i \triangleq \begin{bmatrix}\boldsymbol{x}_{i1} & \boldsymbol{x}_{i2} & \cdots & \boldsymbol{x}_{iN_i}\end{bmatrix}$ be a matrix whose columns lie in $\mathcal{S}_i$. Denote by $P(\boldsymbol{X}_i)$ the symmetrized convex hull of $\boldsymbol{X}_i$. More precisely,

$$P(\boldsymbol{X}_i) \triangleq \operatorname{conv}(\pm\boldsymbol{x}_{i1}, \pm\boldsymbol{x}_{i2}, \ldots, \pm\boldsymbol{x}_{iN_i}). \quad \text{(II.9)}$$

The subspace inradius associated with $\mathcal{S}_i$, which we denote by r_i, is defined as the radius of the largest Euclidean ball whose intersection with $\mathcal{S}_i$ is inscribed in $P(\boldsymbol{X}_i)$, see Figure 1.

Definition 2.3 (subspace incoherence): The subspace incoherence associated with $\mathcal{S}_i$ is defined as

$$\mu_i \triangleq \max_{\boldsymbol{x} \in \mathcal{S}_i, \|\boldsymbol{x}\|_{\ell_2}=1} \left\|\boldsymbol{x}^\top \boldsymbol{X}_{-i}\right\|_{\ell_\infty}. \quad \text{(II.10)}$$

In other words, μ_i is the maximum inner product between an arbitrary vector of unit Euclidean norm in $\mathcal{S}_i$ and the columns of $\boldsymbol{X}_{-i}$, which correspond to data points in other subspaces, see Figure 2.

Notice that from the definition of the principal angles between subspaces, we always have $\mu_i \leq \max_{j \neq i} \cos\theta_{ij}$, where θ_{ij} denotes the smallest principal angle between $\mathcal{S}_i$ and $\mathcal{S}_j$.

In this paper, we show that, as long as the subspace incoherences between $\mathcal{S}_i$ and other subspaces are sufficiently small compared to the subspace inradius of $\mathcal{S}_i$, the optimization algorithm in (II.8), for an appropriate γ, finds an approximate subspace-sparse representation for any $\boldsymbol{y} \in \mathcal{S}_i^\varepsilon$. More specifically, we prove the following result.

Theorem 2.1: *Let* $\gamma \triangleq \max_i 2(1 + \frac{2\sqrt{2(\log N_i + \log n)}}{r_i})$. *Define* β *as*

$$\beta \triangleq (1 + \max_i \frac{3r_i}{r_i - (\mu_i + \varepsilon)})\frac{\gamma}{2} + \delta, \quad \text{(II.11)}$$

where $\delta > 0$ *is arbitrarily small. Then, for every* i *and every* $\boldsymbol{y} \in \mathcal{S}_i^\varepsilon$, *the solution of the optimization problem in (II.8), denoted by* $\boldsymbol{c}^{*\top} = \begin{bmatrix}\boldsymbol{c}_i^{*\top} & \boldsymbol{c}_{-i}^{*\top}\end{bmatrix}$, *with high probability, satisfies*

$$\|\boldsymbol{y} - \boldsymbol{Y}_i\boldsymbol{c}_i^*\|_{\ell_2} \leq \beta\varepsilon. \quad \text{(II.12)}$$

In addition, assuming $\varepsilon \leq \frac{\gamma}{2\beta+\gamma} r_i$, *with high probability, we have*

$$\left\|\boldsymbol{c}_{-i}^*\right\|_{\ell_1} \leq \frac{2\beta + \gamma}{2r_i}\varepsilon. \quad \text{(II.13)}$$

Notice that, from (II.11), a necessary condition for the approximate subspace-sparse recovery is to have $\mu_i + \varepsilon < r_i$ for all i. This in fact makes sense since from earlier results [6], [33], in the noise-free setting, perfect subspace-sparse recovery holds as long as $\mu_i < r_i$ holds for all i. Thus, given the fact that data points are corrupted by noise whose Euclidean norm is about ε, there is a need for adjustment of the condition by incorporating the noise level.

Remark II-B: Our results in Theorem 2.1 suggest that the smaller the ratio $(\mu_i + \varepsilon)/r_i$ is and the larger r_i is, the better recovery we obtain using the ℓ_1-minimization in (II.8). This is expected, since a larger r_i corresponds to a more even distribution of points in subspace i, i.e., farther from a degenerate subspace. On the other hand, a smaller μ_i corresponds to points in different subspaces being more dissimilar to each other.

Example II-C: Let $1/\kappa \triangleq \max_i(\mu_i + \varepsilon)/r_i$, where $\kappa > 1$, from the necessary condition that r_i must be greater than $\mu_i + \varepsilon$, as stated earlier. Also, let $c \triangleq 2\sqrt{2(\log N_{i^*} + \log n)}$ and $r \triangleq r_{i^*}$, where i^* is the index for which we obtain the maximum value in the definition of γ in Theorem 2.1. Hence, we have

$$\gamma = 2\,(1 + \frac{c}{r}). \tag{II.14}$$

In addition, using the definition of β in (II.11), we can write

$$\beta = (4 + \frac{3}{\kappa - 1})(1 + \frac{c}{r}) + \delta. \tag{II.15}$$

Clearly, the larger the value of subspace inradius r is, the less error tolerance γ we can allow for the reconstruction of a given $\boldsymbol{y}$ and, at the same time, the reconstruction of $\boldsymbol{y}$ using noisy points in its own subspace has a smaller error. In addition, as κ increases, the error on the reconstruction of $\boldsymbol{y}$ using noisy points in its own subspace decreases. In the limiting case of κ being large enough, we obtain

$$\|\boldsymbol{y} - \boldsymbol{Y}_i \boldsymbol{c}_i^*\|_{\ell_2} \leq 4\,(1 + \frac{c}{r})\,\varepsilon. \tag{II.16}$$

In the next section, we provide the required theoretical analysis tools to prove the above result. In fact, our theory relies on a novel generalization of the null-space property [40], [41], [42], [43] to the setting where 1) data lie in a union of subspaces, with the number of data points in each subspace typically larger than the subspace dimension; 2) all data points are corrupted by noise.

III. Approximate Subspace-Sparse Recovery Theory

In this section, we consider the ℓ_1-minimization program

$$\begin{aligned} \begin{bmatrix} \boldsymbol{c}_i^* \\ \boldsymbol{c}_{-i}^* \end{bmatrix} &= \operatorname{argmin} \left\| \begin{bmatrix} \boldsymbol{c}_i \\ \boldsymbol{c}_{-i} \end{bmatrix} \right\|_{\ell_1} \\ &\text{s.t.} \quad \left\| \boldsymbol{y} - \begin{bmatrix} \boldsymbol{Y}_i & \boldsymbol{Y}_{-i} \end{bmatrix} \begin{bmatrix} \boldsymbol{c}_i \\ \boldsymbol{c}_{-i} \end{bmatrix} \right\|_{\ell_2} \leq \gamma\varepsilon, \end{aligned} \tag{III.1}$$

and investigate conditions under which we achieve approximate subspace-sparse recovery for an arbitrary noisy data point $\boldsymbol{y} \in \mathcal{S}_i^\varepsilon$. More precisely, we investigate conditions under which the optimal solution of (III.1) approximately reconstructs $\boldsymbol{y}$ from noisy data points in its own subspace, i.e., $\|\boldsymbol{y} - \boldsymbol{Y}_i \boldsymbol{c}_i^*\|_{\ell_2}$ is bounded by $O(\varepsilon)$, and the coefficients corresponding to noisy data points in other subspaces are sufficiently small, i.e., $\|\boldsymbol{c}_{-i}^*\|_{\ell_1}$ is of the order of $O(\varepsilon)$.

A. Preliminary Lemmas

To prove the main results of the paper, we make use of the following Lemmas. The proof of the first Lemma can be found in [33] and we provide the proofs of the other two Lemmas in the Appendix.

Lemma 3.1: Given a noise-free data point $\boldsymbol{x} \in \mathcal{S}_i$, the ℓ_1-norm of the optimal solution of the minimization program

$$\boldsymbol{c}_i^* = \operatorname{argmin} \|\boldsymbol{c}\|_{\ell_1} \quad \text{s.t.} \quad \boldsymbol{x} = \boldsymbol{X}_i \boldsymbol{c}, \tag{III.2}$$

satisfies the following inequality

$$\|\boldsymbol{c}_i^*\|_{\ell_1} \leq \frac{\|\boldsymbol{x}\|_{\ell_2}}{r_i}. \tag{III.3}$$

In other words, the upper bound on the minimum ℓ_1-norm representation of a noise-free data point $\boldsymbol{x}$ in $\mathcal{S}_i$ in terms of noise-free data points in $\mathcal{S}_i$ is proportional to the Euclidean norm of $\boldsymbol{x}$ and is inversely proportional to the subspace inradius r_i.

Lemma 3.2: For $\boldsymbol{Z}_i \in \mathbb{R}^{n \times N_i}$ with i.i.d entries drawn from $\mathcal{N}(0, \frac{\epsilon^2}{n})$ and a given $\boldsymbol{c}_i \in \mathbb{R}^{N_i}$, with probability at least $1 - \frac{1}{(nN_i)^2}$, we have

$$\|\boldsymbol{Z}_i \boldsymbol{c}_i\|_{\ell_2} \leq 2\varepsilon \sqrt{2(\log N_i + \log n)}\, \|\boldsymbol{c}_i\|_{\ell_1}. \tag{III.4}$$

The result of the above Lemma implies that given $\boldsymbol{Y}_i = \boldsymbol{X}_i + \boldsymbol{Z}_i$ whose columns are noisy data points in $\mathcal{S}_i^\varepsilon$, the linear combination $\boldsymbol{Y}_i \boldsymbol{c}_i$ corresponds to perturbing the noise-free vector $\boldsymbol{X}_i \boldsymbol{c}_i$ lying in $\mathcal{S}_i$ with a noise whose Euclidean norm is bounded above by (III.4).

Lemma 3.3: Given a noisy data point in the i-th subspace, $\boldsymbol{y} \in \mathcal{S}_i^\varepsilon$, consider the ℓ_1-minimization program

$$\boldsymbol{c}^* = \operatorname{argmin} \|\boldsymbol{c}\|_{\ell_1} \quad \text{s.t.} \quad \|\boldsymbol{y} - \boldsymbol{Y}\boldsymbol{c}\|_{\ell_2} \leq \gamma\varepsilon, \tag{III.5}$$

with $\gamma \triangleq \max_i 2(1 + \frac{2\sqrt{2(\log N_i + \log n)}}{r_i})$. With probability at least $1 - \frac{1}{(nN_i)^2}$, we have

$$\|\boldsymbol{c}^*\|_{\ell_1} \leq \frac{1}{r_i}. \tag{III.6}$$

Thus, for an appropriately chosen error tolerance, the upper bound on the ℓ_1-norm of the optimal representation of a noisy data point in $\mathcal{S}_i^\varepsilon$, as a linear combination of all noisy data points in $\boldsymbol{Y}$, is inversely proportional to the subspace inradius r_i.

As a consequence of Lemmas 3.2 and 3.3, for the optimal solution of (III.1), we have

$$\begin{aligned} \|\boldsymbol{Z}_i \boldsymbol{c}_i^*\|_{\ell_2} &\leq 2\varepsilon \sqrt{2(\log N_i + \log n)}\, \|\boldsymbol{c}_i^*\|_{\ell_1} \\ &\leq 2\varepsilon\, \frac{\sqrt{2(\log N_i + \log n)}}{r_i}, \end{aligned} \tag{III.7}$$

where we used the fact that $\|\boldsymbol{c}_i^*\|_{\ell_1} \leq \|\boldsymbol{c}^*\|_{\ell_1} \leq \frac{1}{r_i}$.

B. Main Results

In this section, we prove our main result in Theorem 2.1. To do so, we consider an arbitrary vector $\tilde{\boldsymbol{y}}$ that lies close to $\mathcal{S}_i$ and whose Euclidean norm is larger than the approximate recovery noise level, i.e., $\|\tilde{\boldsymbol{y}}\|_{\ell_2} > \beta\varepsilon$, where $\beta > 0.5\gamma$. We consider the following ℓ_1-minimization programs,

$$\boldsymbol{a}_i(\tilde{\boldsymbol{y}}) = \operatorname{argmin} \|\boldsymbol{a}\|_{\ell_1} \quad \text{s.t.} \quad \|\tilde{\boldsymbol{y}} - \boldsymbol{X}_i\boldsymbol{a}\|_{\ell_2} \leq \frac{\gamma}{2}\varepsilon, \tag{III.8}$$

$$\boldsymbol{a}_{-i}(\tilde{\boldsymbol{y}}) = \operatorname{argmin} \|\boldsymbol{a}\|_{\ell_1} \quad \text{s.t.} \quad \|\tilde{\boldsymbol{y}} - \boldsymbol{Y}_{-i}\boldsymbol{a}\|_{\ell_2} \leq \gamma\varepsilon. \tag{III.9}$$

In other words, in (III.8), we consider approximate reconstruction of $\tilde{\boldsymbol{y}}$ using noise-free data points in $\mathcal{S}_i$, and in (III.9), we consider approximate reconstruction of $\tilde{\boldsymbol{y}}$ using noisy data points in subspaces other than $\mathcal{S}_i^{\varepsilon}$.

The structure of our theoretical analysis in the paper is as follows. First, in Theorem 3.1, we find conditions based on the inradius and incoherence of subspaces under which we have $\|\boldsymbol{a}_i(\tilde{\boldsymbol{y}})\|_{\ell_1} < \|\boldsymbol{a}_{-i}(\tilde{\boldsymbol{y}})\|_{\ell_1}$, for every $\tilde{\boldsymbol{y}}$. Our result corresponds to a novel generalization of the null-space property [40], [41], [42], [43] to the case where 1) data lie in a union of subspaces, with the number of data points in each subspace typically larger than the subspace dimension; 2) all data points are corrupted by noise. Then, in Theorems 3.2 and 3.3, we show that if the noisy multi-subspace null-space property holds, i.e., $\|\boldsymbol{a}_i(\tilde{\boldsymbol{y}})\|_{\ell_1} < \|\boldsymbol{a}_{-i}(\tilde{\boldsymbol{y}})\|_{\ell_1}$, for every $\tilde{\boldsymbol{y}}$, then the optimization problem (III.1) achieves approximate subspace-sparse recovery according to Definition 2.1.

For brevity of the notation, we denote $\boldsymbol{a}_i(\tilde{\boldsymbol{y}})$ and $\boldsymbol{a}_{-i}(\tilde{\boldsymbol{y}})$ by $\boldsymbol{a}_i$ and $\boldsymbol{a}_{-i}$, respectively, whenever the argument $\tilde{\boldsymbol{y}}$ is clear from the context. To characterize the set of admissible $\tilde{\boldsymbol{y}}$ in our theoretical analysis, we make use of the following definition.

Definition 3.1: We denote by $\mathbb{W}_i(\beta, \gamma, \varepsilon)$ the set of all $\tilde{\boldsymbol{y}}$ with $\|\tilde{\boldsymbol{y}}\|_{\ell_2} > \beta\varepsilon$, which can be written as the sum of a noise-free vector in $\mathcal{S}_i$ and a noise whose Euclidean norm is smaller than or equal to $0.5\gamma\varepsilon$, i.e.,

$$\begin{aligned}\mathbb{W}_i(\beta, \gamma, \varepsilon) \triangleq \{\tilde{\boldsymbol{y}} \in \mathbb{R}^n : \ & \|\tilde{\boldsymbol{y}}\|_{\ell_2} \geq \beta\varepsilon, \ \tilde{\boldsymbol{y}} = \boldsymbol{y} + \boldsymbol{z}, \\ & \boldsymbol{y} \in \mathcal{S}_i, \ \|\boldsymbol{z}\|_{\ell_2} \leq 0.5\gamma\varepsilon\}.\end{aligned} \tag{III.10}$$

Next, we show that for a suitable value of γ, which depends on the subspace inradius, and for suitable values of β, the noisy multi-subspace null-space property holds.

Theorem 3.1 (Noisy Multi-Subspace Null-Space Property): Let $\gamma \triangleq \max_i 2\,(1 + \frac{2\sqrt{2(\log N_i + \log n)}}{r_i})$. *Define* β *as*

$$\beta \triangleq (1 + \max_i \frac{3r_i}{r_i - (\mu_i + \varepsilon)})\,\frac{\gamma}{2} + \delta, \tag{III.11}$$

where $\delta > 0$ *is an arbitrarily small nonnegative number. Then, for every* $\tilde{\boldsymbol{y}}$ *which belongs to* $\mathbb{W}_i(\beta, \gamma, \varepsilon)$*, the solutions of the optimization programs* (III.8) *and* (III.9) *satisfy*

$$\|\boldsymbol{a}_i(\tilde{\boldsymbol{y}})\|_{\ell_1} < \|\boldsymbol{a}_{-i}(\tilde{\boldsymbol{y}})\|_{\ell_1}. \tag{III.12}$$

Proof: Consider $\tilde{\boldsymbol{y}}$ in $\mathbb{W}_i(\beta, \gamma, \varepsilon)$. We can write

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{x}} + \tilde{\boldsymbol{z}}, \tag{III.13}$$

where from (III.10), we have $\tilde{\boldsymbol{x}} \in \mathcal{S}_i$ and $\|\tilde{\boldsymbol{z}}\|_{\ell_2} \leq 0.5\gamma\varepsilon$. Since $\|\tilde{\boldsymbol{y}}\|_{\ell_2} > \beta\varepsilon$, we have that $\|\tilde{\boldsymbol{x}}\|_{\ell_2} > (\beta - 0.5\gamma)\varepsilon$. We prove the result of the theorem in the following steps.

Step 1: We find an upper bound on the ℓ_1-norm of the solution of (III.8) for $\tilde{\boldsymbol{y}}$, i.e., we show that

$$\|\boldsymbol{a}_i\|_{\ell_1} \leq \frac{\|\tilde{\boldsymbol{x}}\|_{\ell_2}}{r_i}. \tag{III.14}$$

Step 2: We find a lower bound on the ℓ_1-norm of the solution of (III.9) for $\tilde{\boldsymbol{y}}$, i.e., we show that, with high probability,

$$\frac{\|\tilde{\boldsymbol{x}}\|_{\ell_2} - 3\gamma\varepsilon/2}{\mu_i + \varepsilon} \leq \|\boldsymbol{a}_{-i}\|_{\ell_1}. \tag{III.15}$$

Step 3: Combining the results of steps 1 and 2 and using the definition of β in (III.11), we show that

$$\|\boldsymbol{a}_i\|_{\ell_1} \leq \frac{\|\tilde{\boldsymbol{x}}\|_{\ell_2}}{r_i} < \frac{\|\tilde{\boldsymbol{x}}\|_{\ell_2} - 3\gamma\varepsilon/2}{\mu_i + \varepsilon} \leq \|\boldsymbol{a}_{-i}\|_{\ell_1}, \tag{III.16}$$

obtaining the desired result.

Proof of step 1: Our goal is to find an upper bound on the ℓ_1-norm of the solution of (III.8) for $\tilde{\boldsymbol{y}}$, defined in (III.13). Since $\tilde{\boldsymbol{x}}$ lies in $\mathcal{S}_i$, it can be written as a linear combination of noise-free data points in $\boldsymbol{X}_i$. Let

$$\boldsymbol{b}_i = \operatorname{argmin} \|\boldsymbol{b}\|_{\ell_1} \quad \text{s.t.} \quad \tilde{\boldsymbol{x}} = \boldsymbol{X}_i\boldsymbol{b}. \tag{III.17}$$

From Lemma 3.1 we have $\|\boldsymbol{b}_i\|_{\ell_1} \leq \frac{\|\tilde{\boldsymbol{x}}\|_{\ell_2}}{r_i}$. In addition, using (III.13), we can write $\tilde{\boldsymbol{y}}$ as

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{x}} + \tilde{\boldsymbol{z}} = \boldsymbol{X}_i\boldsymbol{b}_i + \tilde{\boldsymbol{z}}, \tag{III.18}$$

where $\|\tilde{\boldsymbol{z}}\|_{\ell_2} \leq \gamma\varepsilon/2$. As a result, $\boldsymbol{b}_i$ is a feasible solution for the ℓ_1-minimization program in (III.8). Hence, using the fact that $\boldsymbol{a}_i$ is the optimal solution of (III.8), we obtain

$$\|\boldsymbol{a}_i\|_{\ell_1} \leq \|\boldsymbol{b}_i\|_{\ell_1} \leq \frac{\|\tilde{\boldsymbol{x}}\|_{\ell_2}}{r_i}. \tag{III.19}$$

Proof of step 2: Our goal is to find a lower bound on the ℓ_1-norm of the solution of (III.9) for $\tilde{\boldsymbol{y}}$, defined in (III.13). By the feasibility of $\boldsymbol{a}_{-i}$ for the optimization program (III.9), we can write

$$\tilde{\boldsymbol{y}} = \boldsymbol{Y}_{-i}\boldsymbol{a}_{-i} + \boldsymbol{v}, \tag{III.20}$$

where $\|\boldsymbol{v}\|_{\ell_2} \leq \gamma\varepsilon$. Substituting the above equation into (III.13), we can write

$$\tilde{\boldsymbol{x}} = \boldsymbol{Y}_{-i}\boldsymbol{a}_{-i} + (\tilde{\boldsymbol{z}} - \boldsymbol{v}), \tag{III.21}$$

where, $\|\tilde{\boldsymbol{z}} - \boldsymbol{v}\|_{\ell_2} \leq 3\gamma\varepsilon/2$. Multiplying both sides of the above equation on the left by $\tilde{\boldsymbol{x}}^\top / \|\tilde{\boldsymbol{x}}\|_{\ell_2}$ and using the Hölder's inequality, we obtain

$$\begin{aligned}\|\tilde{\boldsymbol{x}}\|_{\ell_2} &\leq \left\| \frac{\tilde{\boldsymbol{x}}^\top}{\|\tilde{\boldsymbol{x}}\|_{\ell_2}} \boldsymbol{Y}_{-i} \right\|_{\ell_\infty} \|\boldsymbol{a}_{-i}\|_{\ell_1} + \frac{3}{2}\gamma\varepsilon \\ &\leq \left(\left\| \frac{\tilde{\boldsymbol{x}}^\top}{\|\tilde{\boldsymbol{x}}\|_{\ell_2}} \boldsymbol{X}_{-i} \right\|_{\ell_\infty} + \left\| \frac{\tilde{\boldsymbol{x}}^\top}{\|\tilde{\boldsymbol{x}}\|_{\ell_2}} \boldsymbol{Z}_{-i} \right\|_{\ell_\infty} \right) \|\boldsymbol{a}_{-i}\|_{\ell_1} \\ &\quad + \frac{3}{2}\gamma\varepsilon \quad \leq \quad (\mu_i + \varepsilon)\,\|\boldsymbol{a}_{-i}\|_{\ell_1} + \frac{3}{2}\gamma\varepsilon,\end{aligned} \tag{III.22}$$

where we used the fact that the Euclidean norm of each column of $\boldsymbol{Z}_{-i} \in \mathbb{R}^{n\times(N-N_i)}$ is at most ε, with high probability. Hence, we obtain the following lower bound on the optimal solution of (III.9),

$$\frac{\|\tilde{\boldsymbol{x}}\|_{\ell_2} - 3\gamma\varepsilon/2}{\mu_i + \varepsilon} \leq \|\boldsymbol{a}_{-i}\|_{\ell_1}. \tag{III.23}$$

Proof of step 3: Using the definition of β in (III.11), it is easy to verify that, we have

$$\frac{\mu_i + \varepsilon}{r_i} < 1 - \frac{3\gamma}{2\beta - \gamma}. \tag{III.24}$$

In addition, using the fact that $\|\tilde{\boldsymbol{x}}\|_{\ell_2} \geq (\beta - 0.5\gamma)\varepsilon$, we have

$$\frac{\mu_i + \varepsilon}{r_i} < 1 - \frac{3\eta\varepsilon}{(\beta-\eta)\varepsilon} \leq \frac{\|\tilde{\boldsymbol{x}}\|_{\ell_2} - 3\gamma\varepsilon/2}{\|\tilde{\boldsymbol{x}}\|_{\ell_2}}, \tag{III.25}$$

from which we obtain

$$\frac{\|\tilde{\boldsymbol{x}}\|_{\ell_2}}{r_i} < \frac{\|\tilde{\boldsymbol{x}}\|_{\ell_2} - 3\eta\varepsilon}{\mu_i + \varepsilon}. \tag{III.26}$$

Finally, combining (III.26) with the results of steps 1 and 2, we obtain the desired result of the theorem, i.e.,

$$\|\boldsymbol{a}_i\|_{\ell_1} \leq \frac{\|\tilde{\boldsymbol{x}}\|_{\ell_2}}{r_i} < \frac{\|\tilde{\boldsymbol{x}}\|_{\ell_2} - 3\gamma\varepsilon/2}{\mu_i + \varepsilon} \leq \|\boldsymbol{a}_{-i}\|_{\ell_1}. \tag{III.27}$$

■

The result of Theorem 3.1 shows that for a suitable value of the regularization parameter γ and for a suitable β, the noisy multi-subspace null-space property $\|\boldsymbol{a}_i(\tilde{\boldsymbol{y}})\|_{\ell_1} < \|\boldsymbol{a}_{-i}(\tilde{\boldsymbol{y}})\|_{\ell_1}$ holds, for every $\tilde{\boldsymbol{y}} \in \mathbb{W}_i(\beta,\gamma,\varepsilon)$. In the next two theorems, we show that if the noisy multi-subspace null-space property holds, then the ℓ_1-minimization program in (III.1), with high probability, achieves approximate subspace-sparse recovery.

Theorem 3.2 (Approximate Reconstruction): Let $\gamma \triangleq \max_i 2(1 + \frac{2\sqrt{2(\log N_i + \log n)}}{r_i})$. *Assume that there exists* $\beta > 0.5\gamma$ *such that for every* $\tilde{\boldsymbol{y}} \in \mathbb{W}_i(\beta,\gamma,\varepsilon)$, *the solutions of the optimization programs* (III.8) *and* (III.9) *satisfy* $\|\boldsymbol{a}_i(\tilde{\boldsymbol{y}})\|_{\ell_1} < \|\boldsymbol{a}_{-i}(\tilde{\boldsymbol{y}})\|_{\ell_1}$. *Then the solution of our proposed optimization program in* (III.1), *with probability at least* $1 - \frac{1}{(nN_i)^2}$, *satisfies*

$$\|\boldsymbol{y} - \boldsymbol{Y}_i\boldsymbol{c}_i^*\|_{\ell_2} \leq \beta\varepsilon. \tag{III.28}$$

Proof: Let $\boldsymbol{c}^* = \begin{bmatrix}\boldsymbol{c}_i^* \\ \boldsymbol{c}_{-i}^*\end{bmatrix}$ be the solution of the ℓ_1-minimization program (III.1). For the sake of contradiction, assume that the condition in (III.28) does not hold, i.e., $\|\boldsymbol{y} - \boldsymbol{Y}_i\boldsymbol{c}_i^*\|_{\ell_2} > \beta\varepsilon$. Since $\boldsymbol{c}^*$ is a feasible solution of the optimization program (III.1), we can write

$$\boldsymbol{y} = \boldsymbol{Y}_i\boldsymbol{c}_i^* + \boldsymbol{Y}_{-i}\boldsymbol{c}_{-i}^* + \boldsymbol{e}, \tag{III.29}$$

where $\|\boldsymbol{e}\|_{\ell_2} \leq \gamma\varepsilon$. Define

$$\tilde{\boldsymbol{y}} \triangleq \boldsymbol{y} - \boldsymbol{Y}_i\boldsymbol{c}_i^*. \tag{III.30}$$

Note that by our assumption, we have $\|\tilde{\boldsymbol{y}}\|_{\ell_2} > \beta\varepsilon$. We arrive at contradiction by taking the following three steps.

Step 1: Let $\boldsymbol{a}_{-i}$ be the solution of the optimization program (III.9) for $\tilde{\boldsymbol{y}}$ defined in (III.30). We show that $\begin{bmatrix}\boldsymbol{c}_i^{*\top} & \boldsymbol{a}_{-i}^\top\end{bmatrix}^\top$ is a feasible solution of (III.1), and satisfies

$$\left\|\begin{bmatrix}\boldsymbol{c}_i^* \\ \boldsymbol{a}_{-i}\end{bmatrix}\right\|_{\ell_1} \leq \left\|\begin{bmatrix}\boldsymbol{c}_i^* \\ \boldsymbol{c}_{-i}^*\end{bmatrix}\right\|_{\ell_1}. \tag{III.31}$$

Step 2: Let $\boldsymbol{a}_i$ be the solution of the optimization program (III.8) for $\tilde{\boldsymbol{y}}$ defined in (III.30). We show that $\begin{bmatrix}\boldsymbol{c}_i^{*\top} + \boldsymbol{a}_i^\top & \boldsymbol{0}\end{bmatrix}^\top$ is a feasible solution of (III.1).

Step 3: Combining the results of the first two steps with the main assumption of the theorem, i.e., $\|\boldsymbol{a}_i\|_{\ell_1} < \|\boldsymbol{a}_{-i}\|_{\ell_1}$, we obtain

$$\left\|\begin{bmatrix}\boldsymbol{c}_i^* + \boldsymbol{a}_i \\ \boldsymbol{0}\end{bmatrix}\right\|_{\ell_1} < \left\|\begin{bmatrix}\boldsymbol{c}_i^* \\ \boldsymbol{a}_{-i}\end{bmatrix}\right\|_{\ell_1} \leq \left\|\begin{bmatrix}\boldsymbol{c}_i^* \\ \boldsymbol{c}_{-i}^*\end{bmatrix}\right\|_{\ell_1}. \tag{III.32}$$

contradicting the optimality of $\begin{bmatrix}\boldsymbol{c}_i^{*\top} & \boldsymbol{c}_{-i}^{*\top}\end{bmatrix}^\top$ for the optimization program (III.1).

Proof of step 1: From (III.29), we have $\tilde{\boldsymbol{y}} = \boldsymbol{Y}_{-i}\boldsymbol{c}_{-i}^* + \boldsymbol{e}$. In other words, $\tilde{\boldsymbol{y}}$ can be approximately written as a linear combination of noisy data points in $\boldsymbol{Y}_{-i}$. Since $\|\boldsymbol{e}\|_{\ell_2} \leq \gamma\varepsilon$, we have that $\boldsymbol{c}_{-i}^*$ is a feasible solution of the optimization program (III.9). Let $\boldsymbol{a}_{-i}$ be the optimal solution of (III.9) for $\tilde{\boldsymbol{y}}$, hence

$$\|\boldsymbol{a}_{-i}\|_{\ell_1} \leq \|\boldsymbol{c}_{-i}^*\|_{\ell_1}. \tag{III.33}$$

We can write $\tilde{\boldsymbol{y}}$ as

$$\tilde{\boldsymbol{y}} = \boldsymbol{Y}_{-i}\boldsymbol{a}_{-i} + \boldsymbol{v}, \tag{III.34}$$

where $\|\boldsymbol{v}\|_{\ell_2} \leq \gamma\varepsilon$. Using (III.34) and the definition of $\tilde{\boldsymbol{y}}$ in (III.30), i.e., $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{Y}_i\boldsymbol{c}_i^*$, we can write $\boldsymbol{y}$ as

$$\boldsymbol{y} = \boldsymbol{Y}_i\boldsymbol{c}_i^* + \boldsymbol{Y}_{-i}\boldsymbol{a}_{-i} + \boldsymbol{v}. \tag{III.35}$$

Since $\|\boldsymbol{v}\|_{\ell_2} \leq \gamma\varepsilon$, we have that $\begin{bmatrix}\boldsymbol{c}_i^* \\ \boldsymbol{a}_{-i}\end{bmatrix}$ is a feasible solution of the ℓ_1-minimization program (III.1). Thus, using (III.33), we obtain the desired result of step 1, i.e.,

$$\left\|\begin{bmatrix}\boldsymbol{c}_i^* \\ \boldsymbol{a}_{-i}\end{bmatrix}\right\|_{\ell_1} \leq \left\|\begin{bmatrix}\boldsymbol{c}_i^* \\ \boldsymbol{c}_{-i}^*\end{bmatrix}\right\|_{\ell_1}. \tag{III.36}$$

Another result that we use in the proof of step 2 is the fact that, with probability at least $1 - \frac{1}{(nN_i)^2}$, we have

$$\|\boldsymbol{a}_{-i}\|_{\ell_1} \leq \left\|\begin{bmatrix}\boldsymbol{c}_i^* \\ \boldsymbol{a}_{-i}\end{bmatrix}\right\|_{\ell_1} \leq \left\|\begin{bmatrix}\boldsymbol{c}_i^* \\ \boldsymbol{c}_{-i}^*\end{bmatrix}\right\|_{\ell_1} \leq \frac{1}{r_i}. \tag{III.37}$$

which follows from Lemma 3.3.

Proof of step 2: Since $\boldsymbol{y} \in \mathcal{S}_i^\varepsilon$, we can write $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{z}$, where $\boldsymbol{x}$ is a noise-free data point of unit Euclidean norm in $\mathcal{S}_i$ and $\boldsymbol{z}$ corresponds to noise whose Euclidean norm is bounded above by ε. Therefore, we can rewrite $\tilde{\boldsymbol{y}}$ as

$$\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{Y}_i\boldsymbol{c}_i^* = (\boldsymbol{x} - \boldsymbol{X}_i\boldsymbol{c}_i^*) + (\boldsymbol{z} - \boldsymbol{Z}_i\boldsymbol{c}_i^*). \tag{III.38}$$

Note that $\boldsymbol{x} - \boldsymbol{X}_i\boldsymbol{c}_i^*$ is a vector in $\mathcal{S}_i$, since it is a linear combination of noise-free data points in $\mathcal{S}_i$. Also, from Lemmas 3.2 and 3.3, we have that $\|\boldsymbol{z} - \boldsymbol{Z}_i\boldsymbol{c}_i^*\|_{\ell_2} \leq 0.5\gamma\varepsilon$ holds with probability at least $1 - \frac{1}{(nN_i)^2}$. Thus, $\tilde{\boldsymbol{y}}$ can be written as the sum of a vector in $\mathcal{S}_i$ plus a noise term whose Euclidean norm, with high probability, is bounded above by $0.5\gamma\varepsilon$, hence,

$\tilde{\boldsymbol{y}} \in \mathbb{W}_i(\beta,\gamma,\varepsilon)$. Thus, for $\tilde{\boldsymbol{y}}$, the optimization program (III.8) has a feasible solution, which we denote by $\boldsymbol{a}_i$. Note that using the fact that $\tilde{\boldsymbol{y}} \in \mathbb{W}_i(\beta,\gamma,\varepsilon)$ and the assumption of the theorem, we have

$$\|\boldsymbol{a}_i\|_{\ell_1} < \|\boldsymbol{a}_{-i}\|_{\ell_1}. \tag{III.39}$$

By the optimality of $\tilde{\boldsymbol{y}}$ for the ℓ_1-minimization program (III.8), we can write

$$\tilde{\boldsymbol{y}} = \boldsymbol{X}_i\boldsymbol{a}_i + \boldsymbol{v}, \tag{III.40}$$

where $\|\boldsymbol{v}\|_{\ell_2} \leq 0.5\gamma\varepsilon$. Using (III.40) and the definition of $\tilde{\boldsymbol{y}}$ in (III.30), i.e., $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{Y}_i\boldsymbol{c}_i^*$, we can write $\boldsymbol{y}$ as

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{Y}_i\boldsymbol{c}_i^* + \boldsymbol{X}_i\boldsymbol{a}_i + \boldsymbol{v}\\ &= \boldsymbol{Y}_i(\boldsymbol{c}_i^* + \boldsymbol{a}_i) + (\boldsymbol{v} - \boldsymbol{Z}_i\boldsymbol{a}_i),\end{aligned} \tag{III.41}$$

where the second equality follows from the definition of $\boldsymbol{X}_i = \boldsymbol{Y}_i - \boldsymbol{Z}_i$. Thus, if $\|\boldsymbol{v} - \boldsymbol{Z}_i\boldsymbol{a}_i\|_{\ell_2} < \gamma\varepsilon$, we have that $\begin{bmatrix}\boldsymbol{c}_i^* + \boldsymbol{a}_i \\ \boldsymbol{0}\end{bmatrix}$ is a feasible solution of the optimization program (III.1), hence obtaining the desired result of step 2. Notice that combining (III.39) and (III.37), with probability at least $1 - \frac{1}{(nN_i)^2}$, we have $\|\boldsymbol{a}_i\|_{\ell_1} < \frac{1}{r_i}$. Hence, using Lemma 3.2, the inequality $\|\boldsymbol{v} - \boldsymbol{Z}_i\boldsymbol{a}_i\|_{\ell_2} < \gamma\varepsilon$ holds with high probability.

Proof of step 3: Based on the assumption of the theorem, since $\tilde{\boldsymbol{y}} \in \mathbb{W}_i(\beta,\gamma,\varepsilon)$, we have that $\|\boldsymbol{a}_i\|_{\ell_1} < \|\boldsymbol{a}_{-i}\|_{\ell_1}$. Hence, using the results of steps 1 and 2, we obtain

$$\begin{aligned}\left\|\begin{bmatrix}\boldsymbol{c}_i^* + \boldsymbol{a}_i \\ \boldsymbol{0}\end{bmatrix}\right\|_{\ell_1} &\leq \|\boldsymbol{c}_i^*\|_{\ell_1} + \|\boldsymbol{a}_i\|_{\ell_1}\\ &< \left\|\begin{bmatrix}\boldsymbol{c}_i^* \\ \boldsymbol{a}_{-i}\end{bmatrix}\right\|_{\ell_1} \leq \left\|\begin{bmatrix}\boldsymbol{c}_i^* \\ \boldsymbol{c}_{-i}^*\end{bmatrix}\right\|_{\ell_1}.\end{aligned} \tag{III.42}$$

This contradicts the optimality of $\begin{bmatrix}\boldsymbol{c}_i^* \\ \boldsymbol{c}_{-i}^*\end{bmatrix}$ for the optimization program (III.1). Hence, we must have

$$\|\boldsymbol{y} - \boldsymbol{Y}_i\boldsymbol{c}_i^*\|_{\ell_2} \leq \beta\varepsilon. \tag{III.43}$$

■

Up to this point, we have shown that, under appropriate conditions, for any noisy data point $\boldsymbol{y} \in \mathcal{S}_i^\varepsilon$, the solution of the ℓ_1-minimization program (III.1) is such that $\boldsymbol{y}$ will be reconstructed with high accuracy using noisy data points from its own subspace. Next, we show that, in the optimal solution, the coefficients corresponding to data points in other subspaces will be sufficiently small, provided that the noise level ε is not very large. More specifically, we prove the following result.

Theorem 3.3 (Approximate Support Recovery): Let $\boldsymbol{c}^{*\top} = \begin{bmatrix}\boldsymbol{c}_i^{*\top} & \boldsymbol{c}_{-i}^{*\top}\end{bmatrix}$ *be the solution of the optimization program in* (III.1) *for a noisy data point* $\boldsymbol{y}$ *in* $\mathcal{S}_i^\varepsilon$. *Assume that* $\varepsilon \leq \frac{\gamma}{2\beta+\gamma} r_i$ *and that the approximate reconstruction condition* $\|\boldsymbol{y} - \boldsymbol{Y}_i\boldsymbol{c}_i^*\|_{\ell_2} \leq \beta\varepsilon$ *holds. Then, we have*

$$\|\boldsymbol{c}_{-i}^*\|_{\ell_1} \leq \frac{\beta + \gamma/2}{r_i}\varepsilon. \tag{III.44}$$

Proof: Since $\boldsymbol{y}$ is a noisy data point in $\mathcal{S}_i^\varepsilon$, we can write $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{z}$, where $\boldsymbol{x}$ is a noise-free data point in $\mathcal{S}_i$ and $\boldsymbol{z}$ corresponds to noise whose Euclidean norm is smaller than or equal to ε. Define $\tilde{\boldsymbol{y}} \triangleq \boldsymbol{y} - \boldsymbol{Y}_i\boldsymbol{c}_i^*$, hence, from the assumption of the theorem, we have $\|\tilde{\boldsymbol{y}}\|_{\ell_2} \leq \beta\varepsilon$. We can write $\tilde{\boldsymbol{y}}$ as

$$\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{Y}_i\boldsymbol{c}_i^* = \underbrace{(\boldsymbol{x} - \boldsymbol{X}_i\boldsymbol{c}_i^*)}_{\triangleq \tilde{\boldsymbol{x}}} + \underbrace{(\boldsymbol{z} - \boldsymbol{Z}_i\boldsymbol{c}_i^*)}_{\triangleq \tilde{\boldsymbol{z}}}. \tag{III.45}$$

Notice that $\tilde{\boldsymbol{x}}$ is a vector in $\mathcal{S}_i$, since it is a linear combination of noise-free data points in $\mathcal{S}_i$, and $\tilde{\boldsymbol{z}}$ corresponds to noise whose Euclidean norm is bounded as $\|\tilde{\boldsymbol{z}}\|_{\ell_2} \leq 0.5\gamma\varepsilon$, using Lemmas 3.2 and 3.3. We prove the result in (III.44) by taking the following three steps.

Step 1: First, we show that the minimum ℓ_1-norm of representing $\tilde{\boldsymbol{x}}$, the noise-free part of $\tilde{\boldsymbol{y}}$, using $\boldsymbol{X}_i$, the noise-free data points in $\mathcal{S}_i$, is bounded by

$$\begin{aligned}\min\ & \|\boldsymbol{b}\|_{\ell_1} &&\leq \tfrac{2\beta+\gamma}{2r_i}\varepsilon\\ \text{s.t.}\ & \tilde{\boldsymbol{x}} = \boldsymbol{X}_i\boldsymbol{b}.\end{aligned} \tag{III.46}$$

Step 2: Next, we show that the minimum ℓ_1-norm of the approximate representation of $\tilde{\boldsymbol{y}}$ in terms of noisy data points in $\mathcal{S}_i^\varepsilon$, i.e., $\boldsymbol{Y}_i$, is bounded by

$$\begin{aligned}\min \|\boldsymbol{b}\|_{\ell_1} &&\leq \quad \min \|\boldsymbol{b}\|_{\ell_1}\\ \text{s.t.}\quad \|\tilde{\boldsymbol{y}} - \boldsymbol{Y}_i\boldsymbol{b}\|_{\ell_2} \leq \gamma\varepsilon &&\quad \text{s.t.}\quad \tilde{\boldsymbol{x}} = \boldsymbol{X}_i\boldsymbol{b}.\end{aligned} \tag{III.47}$$

Step 3: Finally, we prove that, for the solution of the optimization program (III.1), the ℓ_1-norm of the coefficients corresponding to noisy data points in subspaces other than $\mathcal{S}_i^\varepsilon$, i.e., $\|\boldsymbol{c}_{-i}^*\|_{\ell_1}$, is bounded by

$$\begin{aligned}\|\boldsymbol{c}_{-i}^*\|_{\ell_1} \quad \leq \quad & \min \|\boldsymbol{b}\|_{\ell_1}\\ & \text{s.t.}\quad \|\tilde{\boldsymbol{y}} - \boldsymbol{Y}_i\boldsymbol{b}\|_{\ell_2} \leq \gamma\varepsilon.\end{aligned} \tag{III.48}$$

Combining the results of steps 1 to 3, we obtain (III.44).

Proof of step 1: Let $\boldsymbol{b}_i$ be the solution of the ℓ_1-minimization program

$$\boldsymbol{b}_i \ = \ \operatorname{argmin} \|\boldsymbol{b}\|_{\ell_1} \quad \text{s.t.} \quad \tilde{\boldsymbol{x}} = \boldsymbol{X}_i\boldsymbol{b}. \tag{III.49}$$

Using (III.45), we can write $\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{y}} - \tilde{\boldsymbol{z}}$, where $\|\tilde{\boldsymbol{y}}\|_{\ell_2} \leq \beta\varepsilon$ and $\|\tilde{\boldsymbol{z}}\|_{\ell_2} \leq 0.5\gamma\varepsilon$. As a result, the Euclidean norm of $\tilde{\boldsymbol{x}}$ is bounded by $\|\tilde{\boldsymbol{x}}\|_{\ell_2} \leq (\beta + \gamma/2)\varepsilon$. Thus, using Lemma 3.1, we obtain

$$\|\boldsymbol{b}_i\|_{\ell_1} \leq \frac{\|\tilde{\boldsymbol{x}}\|_{\ell_2}}{r_i} \leq \frac{\beta + \eta}{r_i}\varepsilon. \tag{III.50}$$

Proof of step 2: Let $\boldsymbol{b}_i$ be the solution of the optimization program (III.49), hence,

$$\tilde{\boldsymbol{x}} = \boldsymbol{X}_i\boldsymbol{b}_i. \tag{III.51}$$

Since, using (III.45), we have $\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{y}} - \tilde{\boldsymbol{z}}$, and also $\boldsymbol{X}_i = \boldsymbol{Y}_i - \boldsymbol{Z}_i$, we can rewrite (III.51) as

$$\tilde{\boldsymbol{y}} = \boldsymbol{Y}_i\boldsymbol{b}_i + (\tilde{\boldsymbol{z}} - \boldsymbol{Z}_i\boldsymbol{b}_i). \tag{III.52}$$

Thus, if we show that $\|\tilde{\boldsymbol{z}} - \boldsymbol{Z}_i\boldsymbol{b}_i\|_{\ell_2} \leq \gamma\varepsilon$, then we obtain (III.47), since $\boldsymbol{b}_i$ is the optimal solution of the right hand-side of (III.47), while it is also a feasible solution of the left hand-side of (III.47). Notice that the columns of $\boldsymbol{X}_i$ are data points

in a d_i-dimensional subspace of $\mathbb{R}^n$. As a result, from the linear programming theory, the optimal solution of the right hand-side of (III.47), $\boldsymbol{b}_i$, has a support whose size is at most d_i. Thus, using the fact that the Euclidean norm of the columns of $\boldsymbol{Z}_i$ is at most ε, we have $\|\boldsymbol{Z}_i\boldsymbol{b}_i\|_{\ell_2} \leq \varepsilon \|\boldsymbol{b}_i\|_{\ell_1}$. Now, using the result of step 1 in (III.50), i.e., $\|\boldsymbol{b}_i\|_{\ell_1} \leq \frac{\beta+\eta}{r_i}\varepsilon$, and the assumption of the theorem on the noise level, i.e., $\varepsilon \leq \frac{\gamma}{2\beta+\gamma} r_i$, we obtain

$$\begin{aligned}\|\tilde{\boldsymbol{z}} - \boldsymbol{Z}_i\boldsymbol{b}_i\|_{\ell_2} &\leq \|\tilde{\boldsymbol{z}}\|_{\ell_2} + \|\boldsymbol{Z}_i\boldsymbol{b}_i\|_{\ell_2} \leq \eta\varepsilon + \varepsilon\|\boldsymbol{b}_i\|_{\ell_1} \\ &\leq \eta\varepsilon + \frac{\beta+\eta}{r_i}\varepsilon^2 \leq 2\eta\varepsilon.\end{aligned} \tag{III.53}$$

Proof of step 3: Let $\boldsymbol{b}_i'$ be the optimal solution of the right hand-side of (III.48), i.e.,

$$\boldsymbol{b}_i' \;=\; \operatorname{argmin} \|\boldsymbol{b}\|_{\ell_1} \quad \text{s.t.} \quad \|\tilde{\boldsymbol{y}} - \boldsymbol{Y}_i\boldsymbol{b}\|_{\ell_2} \leq \gamma\varepsilon. \tag{III.54}$$

For the sake of contradiction, assume that the inequality in (III.48) does not hold, so we have $\|\boldsymbol{b}_i'\|_{\ell_1} < \|\boldsymbol{c}_{-i}^*\|_{\ell_1}$. Using the definition of $\tilde{\boldsymbol{y}}$ in (III.45), i.e., $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{Y}_i\boldsymbol{c}_i^*$, we have

$$\|\tilde{\boldsymbol{y}} - \boldsymbol{Y}_i\boldsymbol{b}_i'\|_{\ell_2} = \|\boldsymbol{y} - \boldsymbol{Y}_i(\boldsymbol{c}_i^* + \boldsymbol{b}_i')\|_{\ell_2} \leq \gamma\varepsilon. \tag{III.55}$$

As a result, $\begin{bmatrix}\boldsymbol{c}_i^* + \boldsymbol{b}_i' \\ \boldsymbol{0}\end{bmatrix}$ is a feasible solution of the optimization problem (III.1). Moreover, we have

$$\left\|\begin{bmatrix}\boldsymbol{c}_i^* + \boldsymbol{b}_i' \\ \boldsymbol{0}\end{bmatrix}\right\|_{\ell_1} \leq \|\boldsymbol{c}_i^*\|_{\ell_1} + \|\boldsymbol{b}_i'\|_{\ell_1} < \left\|\begin{bmatrix}\boldsymbol{c}_i^* \\ \boldsymbol{c}_{-i}^*\end{bmatrix}\right\|_{\ell_1}, \tag{III.56}$$

which contradicts the optimality of $\begin{bmatrix}\boldsymbol{c}_i^* \\ \boldsymbol{c}_{-i}^*\end{bmatrix}$ for (III.1). Thus, we must have $\|\boldsymbol{c}_{-i}^*\|_{\ell_1} \leq \|\boldsymbol{b}_i'\|_{\ell_1}$. ■

Putting the results of Theorems 3.1, 3.2 and 3.3 together, we arrive at our main theoretical results, guaranteeing approximate subspace-sparse recovery in the presenese of noise using the ℓ_1-minimization program in (III.1).

Theorem 3.4: Assume that the columns of $\boldsymbol{Y} \in \mathbb{R}^{n\times N}$ correspond to noisy data points lying in $\{\mathcal{S}_i^\varepsilon\}_{i=1}^L$, with N_i data points in each $\mathcal{S}_i^\varepsilon$. Consider the ℓ_1-minimization program in (III.1) with the γ defined as

$$\gamma \triangleq \max_i 2\Big(1 + \frac{2\sqrt{2(\log N_i + \log n)}}{r_i}\Big). \tag{III.57}$$

Define β as

$$\beta \triangleq \Big(1 + \max_i \frac{3r_i}{r_i - (\mu_i + \varepsilon)}\Big)\frac{\gamma}{2} + \delta, \tag{III.58}$$

where $\delta > 0$ is arbitrarily small. Then, for every $i \in \{1,\ldots,L\}$ and every $\boldsymbol{y} \in \mathcal{S}_i^\varepsilon$, the solution of the optimization problem in (III.1), with probability at least $1 - \frac{1}{(nN_i)^2}$, satisfies

$$\|\boldsymbol{y} - \boldsymbol{Y}_i\boldsymbol{c}_i^*\|_{\ell_2} \leq \beta\varepsilon. \tag{III.59}$$

In addition, assume that $\varepsilon \leq \frac{\gamma}{2\beta+\gamma} r_i$. Then, we have that

$$\|\boldsymbol{c}_{-i}^*\|_{\ell_1} \leq \frac{2\beta+\gamma}{2r_i}\varepsilon \tag{III.60}$$

holds with probability at least $1 - \frac{1}{(nN_i)^2}$.

Proof: Given the choice of γ in (III.57) and β in (III.58), from Theorem 3.1,we have that the multi-subspace noisy null-space property holds, i.e., for every $\tilde{\boldsymbol{y}}$ in $\mathbb{W}_i(\beta,\gamma,\varepsilon)$, we have $\|\boldsymbol{a}_i(\tilde{\boldsymbol{y}})\|_{\ell_1} < \|\boldsymbol{a}_{-i}(\tilde{\boldsymbol{y}})\|_{\ell_1}$, where $\boldsymbol{a}_i(\tilde{\boldsymbol{y}})$ and $\boldsymbol{a}_{-i}(\tilde{\boldsymbol{y}})$ denote the solutions of (III.8) and (III.9), respectively. As a result, the condition of the Theorem 3.2 is satisfied and we have that (III.59) holds, with high probability. Finally, given the approximate reconstruction condition and the assumption of the theorem on the maximum value of ε, from Theorem 3.3, we have that (III.60) holds, with high probability. ■

Notice that in all of our theoretical results so far we allow for arbitrary subspace arrangements and data distributions in subspaces, without any randomness assumption. This is in fact an advantage of our theoretical analysis with respect to [38], which assumes the more restricted setting where data in each subspace are distributed at random. In fact, assuming random distribution for data points, we can further show that in the solution of the ℓ_1-minimization program (III.1), the coefficients from the correct support, i.e., $\boldsymbol{c}_i^*$, not only reconstruct $\boldsymbol{y}$ with a high accuracy, but also have sufficiently large values. More specifically, we prove the following result.

Theorem 3.5 (Correct Support Detection): Assume that the noise-free data in each subspace $\mathcal{S}_i$, i.e., the columns of $\boldsymbol{X}_i$, are drawn uniformly at random from the intersection of the unit hypersphere with $\mathcal{S}_i$. Let $\boldsymbol{c}^{*\top} = \begin{bmatrix}\boldsymbol{c}_i^{*\top} & \boldsymbol{c}_{-i}^{*\top}\end{bmatrix}$ be the solution of the optimization program in (III.1) for a noisy data point $\boldsymbol{y}$ in $\mathcal{S}_i^\varepsilon$. Assume that the approximate reconstruction condition $\|\boldsymbol{y} - \boldsymbol{Y}_i\boldsymbol{c}_i^*\|_{\ell_2} \leq \beta\varepsilon$ holds. Then, with probability at least $1 - \frac{2}{N_i^2}$, we have

$$\|\boldsymbol{c}_i^*\|_{\ell_1} \geq \frac{1 - (\beta+1)\,\varepsilon}{2\sqrt{\frac{2\log N_i}{d_i}} + 2\sqrt{\frac{2\log N_i}{n}}\,\varepsilon}. \tag{III.61}$$

Proof: Our goal is to find a lower bound on the ℓ_1-norm of $\boldsymbol{c}_i^*$. Since $\boldsymbol{y}$ belongs to $\mathcal{S}_i^\varepsilon$, it can be written as $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{z}$, where $\boldsymbol{x}$ is a vector of unit Euclidean norm in $\mathcal{S}_i$ and $\boldsymbol{z}$ corresponds to noise, where $\|\boldsymbol{z}\|_{\ell_2} \leq \varepsilon$. Using the assumption of the theorem, i.e., $\|\boldsymbol{y} - \boldsymbol{Y}_i\boldsymbol{c}_i^*\|_{\ell_2} \leq \beta\varepsilon$, we can write

$$\boldsymbol{y} = \boldsymbol{Y}_i\boldsymbol{c}_i^* + \boldsymbol{e}, \tag{III.62}$$

where $\|\boldsymbol{e}\|_{\ell_2} \leq \beta\varepsilon$. Since $\boldsymbol{Y}_i = \boldsymbol{X}_i + \boldsymbol{Z}_i$, we can rewrite the above equation as

$$\boldsymbol{x} + \boldsymbol{z} - \boldsymbol{e} = (\boldsymbol{X}_i + \boldsymbol{Z}_i)\boldsymbol{c}_i^*. \tag{III.63}$$

Multiplying both sides of (III.63) from left by $\boldsymbol{x}^\top$, and taking the absolute values, we have

$$\left|\boldsymbol{x}^\top(\boldsymbol{x} + \boldsymbol{z} - \boldsymbol{e})\right| = \left|\boldsymbol{x}^\top(\boldsymbol{X}_i + \boldsymbol{Z}_i)\boldsymbol{c}_i^*\right|. \tag{III.64}$$

Note that the left hand-side of (III.64) is bounded by

$$1 - (\beta+1)\,\varepsilon \;\leq\; \left|\boldsymbol{x}^\top(\boldsymbol{x} + \boldsymbol{z} - \boldsymbol{e})\right|. \tag{III.65}$$

On the other hand, using the Hölder's inequality, the right hand-side of (III.64), with probability at least $1 - \frac{2}{N_i^2}$, is

bounded by

$$\begin{aligned}\left|\boldsymbol{x}^\top(\boldsymbol{X}_i+\boldsymbol{Z}_i)\boldsymbol{c}_i^*\right| &\le \left\|\boldsymbol{x}^\top(\boldsymbol{X}_i+\boldsymbol{Z}_i)\right\|_{\ell_\infty}\|\boldsymbol{c}_i^*\|_{\ell_1}\\ &\le \left(\left\|\boldsymbol{x}^\top\boldsymbol{X}_i\right\|_{\ell_\infty}+\left\|\boldsymbol{x}^\top\boldsymbol{Z}_i\right\|_{\ell_\infty}\right)\|\boldsymbol{c}_i^*\|_{\ell_1}\\ &\le 2\left(\sqrt{\frac{2\log N_i}{d_i}}+\sqrt{\frac{2\log N_i}{n}}\,\varepsilon\right)\|\boldsymbol{c}_i^*\|_{\ell_1}.\end{aligned} \tag{III.66}$$

The last inequality in the above follows from Lemma A.2 in the Appendix. Finally, using the lower-bound in (III.64) and the upper-bound in (III.65), for (III.66), we obtain

$$1-(\beta+1)\,\varepsilon \le 2\left(\sqrt{\frac{2\log N_i}{d_i}}+\sqrt{\frac{2\log N_i}{n}}\,\varepsilon\right)\|\boldsymbol{c}_i^*\|_{\ell_1}, \tag{III.67}$$

hence, we arrive at our desired result in (III.61). ■

IV. Conclusions

In this paper, we considered the problem of finding sparse representations for noisy data points in a dictionary that consists of noisy data lying close to a union of subspaces. More specifically, we assumed that the columns of the dictionary correspond to data points drawn from a union of subspaces and corrupted by Gaussian noise whose Euclidean norm is about ε. We studied a constrained ℓ_1-minimization program and showed that under appropriate conditions on the subspace-inradius and subspace-coherence parameters, the solution of the proposed optimization recovers a solution satisfying approximate subspace-sparse recovery. In other words, we showed that the a noisy data point will be reconstructed using data point from its underlying subspace with an error that is of the order of ε, while coefficients corresponding to data points in other subspaces are sufficiently small, of the order of ε. To achieve this result, we developed an analysis framework based on a novel generalization of the null-space property to the setting where data lie in multiple subspaces, the number of data points in each subspace exceeds the dimension of the subspace, and all data points are corrupted by noise. An important future avenue of research is investigating efficient optimizations and theoretical analysis for other types of data corruptions such as large sparse errors.

Appendix A

In the paper, we used the fact that for a Gaussian random vector $\boldsymbol{z}\in\mathbb{R}^n$ with i.i.d entries drawn from $\mathcal{N}(0,\frac{\epsilon^2}{n})$, with high probability, we have $\|\boldsymbol{z}\|_{\ell_2}\le\varepsilon$. To see this, notice that if $z_i\sim\mathcal{N}(0,\frac{\epsilon^2}{n})$, then $q_i\triangleq\frac{n}{\epsilon^2}z_i^2$ follows a χ^2 distribution with one degree of freedom. We use the following Lemma from [46], which provides a bound on linear combination of χ^2 random variables.

Lemma A.1: Let $q_1,\dots,q_n$ be independent χ^2 random variables, each with one degree of freedom. For any vector $\boldsymbol{a}=\begin{bmatrix}a_1 & \cdots & a_n\end{bmatrix}^\top\in\mathbb{R}^n_+$ with nonnegative entries, and for any $t>0$, we have

$$\Pr\left[\sum_{i=1}^n a_iq_i > \|\boldsymbol{a}\|_{\ell_1}+2\sqrt{t}\,\|\boldsymbol{a}\|_{\ell_2}+2t\,\|\boldsymbol{a}\|_{\ell_\infty}\right]\le e^{-t}. \tag{A.1}$$

If we set $a_i=\varepsilon^2/n$ for all $i=1,\dots,n$, then using the above lemma, we obtain that

$$\Pr\left[\|\boldsymbol{z}\|_{\ell_2}^2>\varepsilon^2(1+\rho)^2\right]\le e^{-\frac{(1+\rho)^2-\sqrt{2(1+\rho)^2-1}}{2}n} \tag{A.2}$$

holds for any $\rho>0$.

We also have the following Lemma, which provides a bound on the inner product between a fixed vector and a matrix of Gaussian random variables.

Lemma A.2: Assume $\boldsymbol{A}\in\mathbb{R}^{m\times N}$ has i.i.d entries drawn from $\mathcal{N}(0,\sigma^2)$. Let $\boldsymbol{x}\in\mathbb{R}^m$ be a vector of unit Euclidean norm. We have

$$\Pr\left[\left\|\boldsymbol{A}^\top\boldsymbol{z}\right\|_{\ell_\infty}\le 2\sqrt{2\log N}\,\sigma\right]\ge 1-\frac{1}{N^2}. \tag{A.3}$$

Appendix B
Proof of Lemma 3.2

Denote the j-th row of the noise matrix $\boldsymbol{Z}_i$ by $\boldsymbol{Z}_i^{(j)}\in\mathbb{R}^{N_i}$. We can write

$$\|\boldsymbol{Z}_i\boldsymbol{c}_i\|_{\ell_2}^2\le\sum_{j=1}^n\langle\boldsymbol{Z}_i^{(j)},\boldsymbol{c}_i\rangle^2\le\sum_{j=1}^n\left\|\boldsymbol{Z}_i^{(j)}\right\|_{\ell_\infty}^2\|\boldsymbol{c}_i\|_{\ell_1}^2 \tag{B.1}$$

Since each entry of $\boldsymbol{Z}_i$ has a standard deviation of $\frac{\varepsilon}{\sqrt{n}}$, with probability at least $1-\frac{1}{(nN_i)^2}$, we have

$$\left\|\boldsymbol{Z}_i^{(j)}\right\|_{\ell_\infty}\le 2\varepsilon\sqrt{\frac{2(\log N_i+\log n)}{n}},\quad\forall j\in\{1,\dots,n\}. \tag{B.2}$$

Substituting (B.2) into (B.1), we have that, with probability at least $1-\frac{1}{(nN_i)^2}$, the following inequality holds,

$$\|\boldsymbol{Z}_i\boldsymbol{c}_i\|_{\ell_2}\le 2\varepsilon\sqrt{2(\log N_i+\log n)}\,\|\boldsymbol{c}_i\|_{\ell_1}. \tag{B.3}$$

Appendix C
Proof of Lemma 3.3

Note that $\boldsymbol{y}$ can be written as $\boldsymbol{y}=\boldsymbol{x}+\boldsymbol{z}$, where $\boldsymbol{x}\in\mathcal{S}_i$ has unit Euclidean norm and $\|\boldsymbol{z}\|_{\ell_2}\le\varepsilon$. Notice that $\boldsymbol{x}$ can be written as a linear combination of noise-free data points in $\mathcal{S}_i$. Let

$$\boldsymbol{c}_i^*=\operatorname{argmin}\|\boldsymbol{c}\|_{\ell_1}\quad\text{s.t.}\quad\boldsymbol{x}=\boldsymbol{X}_i\boldsymbol{c}. \tag{C.1}$$

Then from Lemma 3.1, we have $\|\boldsymbol{c}_i^*\|_{\ell_1}\le\frac{1}{r_i}$. On the other hand, we can rewrite $\boldsymbol{x}=\boldsymbol{X}_i\boldsymbol{c}_i^*$ as

$$\boldsymbol{y}-\boldsymbol{z}=(\boldsymbol{Y}_i-\boldsymbol{Z}_i)\boldsymbol{c}_i^*, \tag{C.2}$$

from which we obtain,

$$\boldsymbol{y}=\boldsymbol{Y}_i\boldsymbol{c}_i^*+(\boldsymbol{z}-\boldsymbol{Z}_i\boldsymbol{c}_i^*). \tag{C.3}$$

From Lemma 3.2, with probability at least $1-\frac{1}{(nN_i)^2}$, we have

$$\|\boldsymbol{z}-\boldsymbol{Z}_i\boldsymbol{c}_i^*\|_{\ell_2}\ \le\ \varepsilon\left(1+\frac{2\sqrt{2(\log N_i+\log n)}}{r_i}\right). \tag{C.4}$$

As a result, with high probability, $\begin{bmatrix}\boldsymbol{c}_i^*\\ \boldsymbol{0}\end{bmatrix}$ is a feasible solution of the optimization program (III.5). Thus, we must have

$$\|\boldsymbol{c}^*\|_{\ell_1}\le\|\boldsymbol{c}_i^*\|_{\ell_1}\le\frac{1}{r_i}. \tag{C.5}$$

REFERENCES

[1] R. Basri and D. Jacobs, "Lambertian reflection and linear subspaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 3, pp. 218–233, 2003.

[2] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography," *International Journal of Computer Vision*, vol. 9, no. 2, pp. 137–154, 1992.

[3] T. Hastie and P. Simard, "Metrics and models for handwritten character recognition," *Statistical Science*, vol. 13, no. 1, pp. 54–65, 1998.

[4] W. Hong, J. Wright, K. Huang, and Y. Ma, "Multi-scale hybrid linear models for lossy image representation," *IEEE Trans. on Image Processing*, vol. 15, no. 12, pp. 3655–3671, 2006.

[5] A. Yang, J. Wright, Y. Ma, and S. Sastry, "Unsupervised segmentation of natural images via lossy data compression," *Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 212–225, 2008.

[6] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[7] ——, "Sparse subspace clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[8] G. Lerman and T. Zhang, "Probabilistic recovery of multiple subspaces in point clouds by geometric ℓ_p minimization," *http://arxiv.org/abs/1002.1994*, 2010.

[9] P. Favaro, R. Vidal, and A. Ravichandran, "A closed form solution to robust subspace estimation and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[10] R. Liu, Z. Lin, and Z. S. F. DelaTorre, "Fixed-rank representation for unsupervised visual learning," *CVPR*, 2012.

[11] R. Heckel and H. Blcskei, "Noisy subspace clustering via thresholding," *IEEE International Symposium on Information Theory (ISIT)*, pp. 1382–1386, 2013.

[12] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[13] E. Elhamifar and R. Vidal, "Block-sparse recovery via convex optimization," *IEEE Transactions on Signal Processing*, 2012.

[14] M. Mishali, Y. C. Eldar, and A. Elron, "Xampling: Signal acquisition and processing in union of subspaces," *IEEE Transactions on Signal Processing*, vol. 59, no. 10, pp. 4719–4734, Oct. 2011.

[15] E. Elhamifar and R. Vidal, "Sparse manifold clustering and embedding," in *Neural Information Processing Systems*, 2011.

[16] T. Blumensath, "Sampling and reconstructing signals from a union of linear subspaces," *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4660–4671, 2011.

[17] E. Elhamifar, G. Sapiro, and R. Vidal, "See all by looking at a few: Sparse modeling for finding representative objects," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[18] E. Esser, M. Moller, S. Osher, G. Sapiro, and J. Xin, "A convex model for non-negative matrix factorization and dimensionality reduction on physical space," *IEEE Transactions on Image Processing*, vol. 21, no. 7, pp. 3239–3252, 2012.

[19] E. Elhamifar, G. Sapiro, and R. Vidal, "Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery," *Neural Information Processing Systems*, 2012.

[20] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Inform. Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.

[21] A. Zhang, N. Fawaz, S. Ioannidis, and A. Montanari, "Guess who rated this movie: Identifying users through subspace clustering," *Uncertainty in Artificial Intelligence (UAI)*, 2012.

[22] X. Wang, S. Atev, J. Wright, and G. Lerman, "Fast subspace search via grassmannian based hashing," *International Conference of Computer Vision (ICCV)*, 2013.

[23] Q. Qiu and G. Sapiro, "Learning transformations for classification forests," *International Conference on Learning Representations (ICLR)*, 2014.

[24] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797–829, 2006.

[25] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[26] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society B*, vol. 58, no. 1, pp. 267–288, 1996.

[27] A. Bruckstein, D. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, no. 1, pp. 34–81, Feb. 2009.

[28] N. Meinshausen and B. Yu, "Lasso-type recovery of sparse representations for high-dimensional data," *Annals of Statistics*, vol. 37, no. 1, pp. 246–270, 2009.

[29] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[30] E. J. Candès and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[31] E. Elhamifar and R. Vidal, "Robust classification using structured sparse representation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[32] ——, "Clustering disjoint subspaces via sparse representation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2010.

[33] M. Soltanolkotabi and E. J. Candes, "A geometric analysis of subspace clustering with outliers," *Annals of Statistics*, 2012.

[34] Y. Wang, H. Xu, and C. Leng, "Provable subspace clustering: When lrr meets ssc," *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[35] D. L. Donoho, "For most large underdetermined systems of linear equations, the minimal ell^1-norm near-solution approximates the sparsest near-solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 7, pp. 907–934, 2006.

[36] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," in *Compte Rendus de l'Academie des Sciences, Paris, Serie I*, vol. 346, 2008, pp. 589–592.

[37] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Compressed sensing of block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Trans. Signal Processing*, vol. 58, no. 6, pp. 3042–3054, June 2010.

[38] M. Soltanolkotabi, E. Elhamifar, and E. J. Candes, "Robust subspace clustering," *Annals of Statistics*, 2014.

[39] Y. Wang and H. Xu, "Noisy sparse subspace clustering," *International Conference on Machine Learning (ICML)*, 2013.

[40] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *PNAS*, vol. 100, no. 5, pp. 2197–2202, 2003.

[41] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Information Theory*, vol. 49, no. 12, pp. 3320–3325, Dec. 2003.

[42] M. Stojnic, F. Parvaresh, and B. Hassibi, "On the reconstruction of block-sparse signals with and optimal number of measurements," *IEEE Trans. Signal Processing*, vol. 57, no. 8, pp. 3075–3085, Aug. 2009.

[43] E. van den Berg and M. Friedlander, "Theoretical and empirical results for recovery from multiple measurements," *IEEE Trans. Information Theory*, vol. 56, no. 5, pp. 2516–2527, 2010.

[44] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. on Information Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.

[45] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique*, vol. 346, no. 9-10, pp. 589–592, 2008.

[46] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *The Annals of Statistics*, vol. 28, no. 5, pp. 1302–1338, 2000.