Clustering a Union of Linear Subspaces Using Optimal Vectors

Abstract

This paper focuses on the Matrix Factorization based Clustering (MFC) method which is one of the few closed-form algorithms for the subspace clustering algorithm. Despite being simple, closed-form, and computation-efficient, MFC can outperform the other sophisticated subspace clustering methods in many challenging scenarios. We reveal the connection between MFC and the Innovation Pursuit (iPursuit) algorithm which was shown to be able to outperform the other spectral clustering based methods with a notable margin especially when the span of clusters are close. A novel theoretical study is presented which sheds light on the key performance factors of both algorithms (MFC/iPursuit) and it is shown that both algorithms can be robust to notable intersections between the span of clusters. Importantly, in contrast to the theoretical guarantees of other algorithms which emphasized on the distance between the subspaces as the key performance factor and without making the innovation assumption, it is shown that the performance of MFC/iPursuit mainly depends on the distance between the innovative components of the clusters.

1 INTRODUCTION

When data points lie in a single linear manifold, conventional techniques such as Principal Component Analysis (PCA) can be efficiently used to find the underlying low-dimensional structure [Zhang and Lerman, 2014, Lerman et al., 2015]. However, in many applications, the data points may be originating from multiple independent sources and a union of manifolds can better model the data [Vidal, 2011]. The subspace clustering problem is defined on how to learn these low dimensional manifolds when they are linear subspaces [Heckel and Bölcskei, 2013, Elhamifar and Vidal, 2013, Tsakiris and Vidal, 2017, Rahmani and Atia, 2017a, Peng et al., 2016, Lu et al., 2013, Feng et al., 2014, Patel et al., 2013, Wang et al., 2013, Li et al., 2021, Lu et al., 2013, Wang and Xu, 2016, You et al., 2016, Ji et al., 2017, Zhang et al., 2018, Klys et al., 2018, Peng et al., 2016, Menon et al., 2020, Jiang et al., 2018, Lipor et al., 2021] in a completely unsupervised way.

Summary of contributions: This paper focuses on analyzing two subspace clustering algorithms: Matrix Factorization based Clustering (MFC) and Innovation Pursuit (iPursuit). First we reveal the underlying connection between them and the presented analysis shows why they can notably outperform other spectral clustering based methods in the challenging scenarios. The main contributions of this work can be summarized as follows.

• It is shown that iPursuit is equivalent to MFC if we alter its ℓ_1-norm based cost function into a quadratic cost function and importantly, all the presented theoretical results are applicable to both algorithms.

• To the best of our knowledge, this paper presents the first comprehensive analysis of MFC/iPursuit algorithms and the presented analysis is not based on the restrictive innovation assumption used in [Rahmani and Atia, 2017a,b]. The MFC/iPursuit algorithms are analyzed and we establish deterministic and probabilistic sufficient conditions which guarantee that the computed adjacency matrix by MFC/iPursuit satisfies a defined quality requirement. Importantly, it is shown that in contrast to most of other clustering algorithms whose performance depend on the distance between the subspaces, the performance of MFC/iPursuit mainly depends on the distance between the innovative components of the clusters. Accordingly, even if the span of clusters intersect heavily, MFC/iPursuit can still provably satisfy the performance requirement.

Notation and Definitions: Given a matrix $\mathbf{A}$, $\|\mathbf{A}\|$ denotes its spectral norm, $\|\mathbf{A}\|_F$ denotes its Frobenius norm, and $\|\mathbf{A}\|_{p,1} = \sum_i \|\mathbf{a}_i\|_p$ where $\mathbf{a}_i$ denotes the ith column of $\mathbf{A}$ and $\mathbf{a}^i$ denotes the ith row of $\mathbf{A}$. For a vector $\mathbf{a}$, $\|\mathbf{a}\|_p$

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

denotes its ℓ_p-norm, $\mathbf{a}(i)$ denotes its ith element, and $\mathbf{a}[i:k]$ contains the elements of $\mathbf{a}$ whose indexes are from i to k. The elements of matrix $\mathbf{Y} = |\mathbf{X}|$ are equal to the absolute value of the elements of matrix $\mathbf{X}$. The subspace $\mathcal{U}^{\perp}$ is the complement of $\mathcal{U}$. $\mathbb{S}^{M_1-1}$ indicates the unit ℓ_2-norm sphere in $\mathbb{R}^{M_1}$. It is assumed that data matrix $\mathbf{D} \in \mathbb{R}^{M_1 \times M_2}$ can be represented as $\mathbf{D} = \mathbf{U}\Sigma\mathbf{V}^T$ where $\mathbf{U} \in \mathbb{R}^{M_1 \times r_d}$ is the matrix of left singular vectors, the diagonal matrix $\Sigma \in \mathbb{R}^{r_d \times r_d}$ contains the non-zero singular values, the columns of $\mathbf{V} \in \mathbb{R}^{M_2 \times r_d}$ are equal to the right singular vectors, r_d is the rank of $\mathbf{D}$, M_2 is the number of data points, and M_1 is the dimension of ambient space. The subspace $\mathcal{S} = \oplus_{i=1}^{m} \mathcal{S}_i$ is equal to the direct sum of subspaces $\{\mathcal{S}_i\}_{i=1}^{m}$ and $\dim(\mathcal{S})$ denotes the dimension of $\mathcal{S}$. Two adjacency matrices $\mathbf{A} \in \mathbb{R}^{M_2 \times M_2}$ and $\mathbf{B} \in \mathbb{R}^{M_2 \times M_2}$ are said to be equivalent when $\frac{\mathbf{a}_i}{\|\mathbf{a}_i\|_1} = \frac{\mathbf{b}_i}{\|\mathbf{b}_i\|_1}$ holds for all $1 \leq i \leq M_2$. RHS means right hand side and LHS means left hand side.

Distance between subspaces: Suppose $\mathbf{U}_1 \in \mathbb{R}^{M_1 \times r}$ and $\mathbf{U}_2 \in \mathbb{R}^{M_1 \times r}$ are orthonormal bases for r-dimensional subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively. Two different notions are used to express the affinity between two subspaces. One measure is $\|\mathbf{U}_1^T\mathbf{U}_2\|$. However, $\|\mathbf{U}_1^T\mathbf{U}_2\|$ is always equal to 1 when $\dim(\mathcal{S}_1 \cap \mathcal{S}_2) > 0$. The other measure of affinity between two subspaces is

$$\|\mathbf{U}_1^T\mathbf{U}_2\|_\sigma = \sqrt{\frac{\sum_{i=1}^{r} \cos^2\theta_i}{r}}$$

where $\{\theta_i\}_{i=1}^{r}$ are the principal angles between $\mathcal{S}_1$ and $\mathcal{S}_2$ [Soltanolkotabi et al., 2012]. Note that $\|\mathbf{U}_1^T\mathbf{U}_2\|_\sigma = 1$ only when $\mathcal{S}_1 = \mathcal{S}_2$.

1.1 DATA MODEL

Data Model 1 provides the details of the presumed model along with definition of the used symbols. To simplify the exposition and the analysis, it is assumed that the dimension of subspaces are equal, the number of data points in different clusters are equal, and a subspace $\mathcal{S}$ is used to define the intersection between the span of clusters.

Data Model 1. *The data matrix $\mathbf{D} \in \mathbb{R}^{M_1 \times M_2}$ can be written as $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_m]\mathbf{T}$ where $\mathbf{T} \in \mathbb{R}^{M_2 \times M_2}$ is an unknown permutation matrix. We define $\mathcal{S}_i$ as the column space of $\mathbf{D}_i$ and $\mathcal{S}_i \not\subset \mathcal{S}_j$ and $\mathcal{S}_j \not\subset \mathcal{S}_i$ for any $i \neq j$. The dimension of all subspaces is equal to r and there are n data points in each cluster, i.e., $\mathbf{D}_i \in \mathbb{R}^{M_1 \times n}$. The dimension of the intersection between subspaces is equal to s, i.e., dim $(\cap_{i=1}^{m}\mathcal{S}_i) = s$ and we define subspace $\mathcal{S} = \cap_{i=1}^{m}\mathcal{S}_i$. In addition, $\mathcal{S}_i \cap \mathcal{S}_j = \mathcal{S}$ for all $i \neq j$. The orthonormal matrix $\mathbf{U}_i \in \mathbb{R}^{M_1 \times r}$ is a basis for $\mathcal{S}_i$ and it can be written as $\mathbf{U}_i = [\mathbf{S}\ ,\ \dot{\mathbf{U}}_i]$ where orthonormal matrix $\mathbf{S} \in \mathbb{R}^{M_1 \times s}$ is a basis for $\mathcal{S} = \cap_{i=1}^{m}\mathcal{S}_i$ and $\dot{\mathbf{U}}_i \in \mathbb{R}^{M \times (r-s)}$ is a basis for $\mathcal{S}_i \cap \mathcal{S}^{\perp}$. The orthonormal matrix $\dot{\mathbf{U}}_i$ represents the component of $\mathcal{S}_i$ which does not lie in $\mathcal{S}$ and we call $\dot{\mathcal{S}}_i = span(\dot{\mathbf{U}}_i) = \mathcal{S}_i \cap \mathcal{S}^{\perp}$ the innovative component of $\mathcal{S}_i$. Each data point $\mathbf{d}_i$ which lies in $\mathcal{S}_{k_i}$ can be represented as*

$$\mathbf{d}_i = \mathbf{S}\alpha_i + \dot{\mathbf{U}}_{k_i}\beta_i\ , \tag{1}$$

where $\alpha_i \in \mathbb{R}^s$ and $\beta_i \in \mathbb{R}^{r-s}$.

In order to represent the association of each data point to its corresponding cluster, we define index k_i such that $\mathbf{d}_i \in \mathcal{S}_{k_i}$. Matrix $\mathbf{D}_{-k}$ includes all the columns of $\mathbf{D}$ except the ones which lie in $\mathcal{S}_k$. Matrices $\dot{\mathbf{D}}_j$ and $\bar{\mathbf{D}}_j$ are defined as $\dot{\mathbf{D}}_j = \dot{\mathbf{U}}_j^T\mathbf{D}_j$ and $\bar{\mathbf{D}}_j = \mathbf{S}^T\mathbf{D}_j$.

Algorithm 1 Data Clustering Using iPursuit

Input. The input is data matrix $\mathbf{D} \in \mathbb{R}^{M_1 \times M_2}$.

1. Project data points on $\mathbb{S}^{M-1}$. Set $\mathbf{d}_i$ equal to $\mathbf{d}_i/\|\mathbf{d}_i\|_2$ for all $1 \leq i \leq M_2$.

2. Direction search. Define $\mathbf{C}^* \in \mathbb{R}^{M_1 \times M_2}$ as optimal point of $\min_{\mathbf{C}} \|\mathbf{C}^T\mathbf{D}\|_1$ subject to $\text{diag}(\mathbf{C}^T\mathbf{D}) = \mathbf{1}$.

3. Define adjacency matrix $\mathbf{A} = \left|\mathbf{C}^T\mathbf{D}\right|$.

4. Normalize the ℓ_1-norm of each row of $\mathbf{A}$ (i.e., normalize degree to 1) and apply graph preprocessing steps (e.g., sparsifying adjacency matrix $\mathbf{A}$ via keeping few dominant non-zero elements of each row).

5. Apply spectral clustering to $\mathbf{A} + \mathbf{A}^T$.

Output: The identified clusters.

2 RELATED WORK

Numerous approaches for subspace clustering were proposed in prior work including statistical-based approaches [Yang et al., 2006, Tipping and Bishop, 1999, Sugaya and Kanatani, 2004, Fischler and Bolles, 1981], spectral clustering based methods [Elhamifar and Vidal, 2013, Liu et al., 2013], the algebraic-geometric approach [Vidal et al., 2005], and iterative methods [Bradley and Mangasarian, 2000]. Much of the recent research work on subspace clustering is focused on spectral clustering [Von Luxburg, 2007] based methods [Dyer et al., 2013, Gao et al., 2015, Elhamifar and Vidal, 2013, Heckel and Bölcskei, 2013, Liu et al., 2013, Rahmani and Atia, 2017c, Soltanolkotabi et al., 2012, Wang et al., 2013, Chen and Lerman, 2009, Park et al., 2014].

The spectral clustering based algorithms are composed of two main steps and they only differ in the first step. First, an adjacency matrix is constructed via finding a neighborhood set for each data point and in the second step, the spectral graph clustering algorithm [Von Luxburg, 2007] is applied to the learned adjacency matrix. For instance, Sparse Subspace Clustering (SSC) [Elhamifar and Vidal, 2013] uses ℓ_1-minimization to construct a sparse adjacency matrix, Low-Rank Representation (LRR) [Liu et al., 2013] uses nuclear norm minimization to find the adjacency matrix, and the Thresholding based Subspace Clustering (TSC) method [Heckel and Bölcskei, 2013] simply uses the inner-product between the data points to construct the adjacency

matrix. In contrast to TSC which uses inner-product between the data points to construct the adjacency matrix, iPursuit [Rahmani and Atia, 2017a,c] utilized the directions of innovation to measure the similarity between the data points. The Matrix Factorization based Clustering (MFC) method [Kanatani, 2001, Costeira and Kanade, 1998, Boult and Brown, 1991] is a closed-from spectral clustering based method which utilizes the right singular vectors of the data to construct the adjacency matrix.

Algorithm 2 Matrix Factorization based Clustering (MFC)

Input. The input is data matrix $\mathbf{D} \in \mathbb{R}^{M_1 \times M_2}$.

1. Project data points on $\mathbb{S}^{M-1}$**.** Set $\mathbf{d}_i$ equal to $\mathbf{d}_i / \|\mathbf{d}_i\|_2$ for all $1 \leq i \leq M_2$.

2. SVD: Compute $\mathbf{D} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ where the columns of $\mathbf{V} \in \mathbb{R}^{M_2 \times r_d}$ are equal to the right singular vectors.

3. Define $\mathbf{A} = \left|\mathbf{V}\mathbf{V}^T\right|$.

4. Similar to Step 4 in Algorithm 1.

5. Similar to Step 5 in Algorithm 1.

Output: The identified clusters.

2.1 A BRIEF OVERVIEW OF IPURSUIT (ALGORITHM 1)

Suppose that data matrix $\mathbf{D}$ follows Data Model 1. If the span of clusters satisfy Assumption 1, then we say that Innovation Assumption holds.

Assumption 1. *For each subspace $\mathcal{S}_i$, we have $\mathcal{S}_i \notin \bigoplus_{k \neq i} \mathcal{S}_k$.*

Define orthonormal matrix $\mathbf{P}_i$ such that the column-space of $\mathbf{P}$ is equal to $\mathcal{P}_i = \bigoplus_{k \neq i} \mathcal{S}_k$. If the innovation assumption holds, then the rank of $(\mathbf{I} - \mathbf{P}_i\mathbf{P}_i^T)\mathbf{U}_i$ is greater than zero and we define $\vec{\mathcal{S}}_i$ as the column-space of $(\mathbf{I} - \mathbf{P}_i\mathbf{P}_i^T)\mathbf{U}_i$. The geometrical idea behind iPursuit is that if we can find a direction in $\vec{\mathcal{S}}_i$, it is orthogonal to all the clusters except $\mathcal{S}_i$ and this fact can be used to distinguish $\mathcal{S}_i$ from the rest of clusters. Specifically, in order to find a direction in $\vec{\mathcal{S}}_{k_i}$ corresponding to each $\mathbf{d}_i$, [Rahmani and Atia, 2017a,c] proposed to find this direction (dubbed the direction of innovation corresponding to $\mathbf{d}_i$) as the optimal point of

$$\min_{\mathbf{c}} \ \|\mathbf{c}^T\mathbf{D}\|_1 \quad \text{subject to} \qquad \mathbf{c}^T\mathbf{d}_i = 1 \,. \tag{2}$$

The motivation behind the design of (2) was that the direction of innovation corresponding to $\mathbf{d}_i$ can be computed via looking for a vector which is orthogonal to the maximum number of data points. Although the innovation assumption was used to design iPursuit, in [Rahmani and Atia, 2017c,a] it was numerically shown that it is not essential in the performance of iPursuit.

The authors of [Rahmani and Atia, 2017c,a] presented an analysis of (2) which is limited to a two cluster scenario and it was based on the Innovation Assumption to prove that the optimal point of (2) lies in $\vec{\mathcal{S}}_{k_i}$. In contrast, the presented theoretical study (a) does not require the innovation assumption, (b) guarantees a completely different requirement, (c) is the first thorough analysis of MFC, (d) reveals the connection between iPursuit and MFC, and importantly (e) it shows the importance of the incoherence between the innovative components.

3 ANALYZING A SPECTRAL CLUSTERING BASED METHOD

The difference between different spectral clustering based algorithms is in the way that they compute the adjacency matrix. Accordingly, we should define proper metrics using which we could determine how accurate/useful is the estimated adjacency matrix. The authors of [Soltanolkotabi et al., 2012] used the number of false connections (any non-zero connection between two nodes/data-points while they belong to different clusters) as a metric to assess the estimated adjacency matrix. However, the graph clustering algorithms such as spectral clustering can yield an exact clustering of the data even if there are a significant amount of false connections in the estimated adjacency matrix provided that the estimated weights on the true connections are sufficiently stronger than the weights of the false connections. Therefore, in this paper, we use the following criteria to assess the quality of a adjacency matrix and we analyze the subspace clustering algorithms to reveal if/how they satisfy Requirement 1.

Requirement 1. *Suppose $\mathbf{A} \in \mathbb{R}^{M_2 \times M_2}$ is the estimated adjacency matrix. We require all the columns of $\mathbf{A}$ to satisfy*

$$\frac{\kappa}{m-1} \|\mathbf{a}_{i\mathcal{I}_i^{\perp}}\|_p^p < \|\mathbf{a}_{i\mathcal{I}_i}\|_p^p \,,$$

where $\mathcal{I}_i = \{j \ \mid \ k_i = k_j\}$, $\mathcal{I}_i^{\perp} = \{j \ \mid \ k_i \neq k_j\}$, $k_i = \arg\max_j \|\mathbf{U}_j^T\mathbf{d}_i\|_2$, and $\mathbf{a}_{i\mathcal{I}_i}$ contains the elements of $\mathbf{a}_i$ whose indexes are in $\mathcal{I}_i$.

The parameter κ is chosen greater than 1 and it determines how well the adjacency matrix represents the clustering structure of the data. Evidently, the higher is κ, the more challenging it is for a subspace clustering algorithm to satisfy Requirement 1. In the following sections, we discuss the role of parameter p and we analyze MFC/iPursuit such that they satisfy Requirement 1 with $p = 1/p = 2$.

Remark 1. *Even if $\mathbf{A}$ satisfies Requirement 1 with a large κ, it does not necessarily mean that Spectral Clustering yields exact clustering. Similarly, proving that $\mathbf{A}$ does not contain any false connection (as in [Soltanolkotabi et al., 2012]) also does not guarantee exact clustering. However, these measures are useful to assess how clear the estimated $\mathbf{A}$ represents the clustering structure. In addition, although Requirement 1 does not guarantee exact clustering by the spectral clustering step, it is very similar to the sufficient*

condition stated in [Ling and Strohmer, 2020] to guarantee that the spectral clustering algorithm yields the exact clustering. Specifically, [Ling and Strohmer, 2020] proves that if

$$\max_i \|\mathbf{a}_{i\mathcal{I}_i^{\perp}}\|_1 < \frac{\min_k \gamma_2(\mathcal{L}(A_k))}{4} ,$$

then the spectral clustering algorithm studied in [Ling and Strohmer, 2020] yields an exact clustering where $\gamma_2(\mathcal{L}(A_k))$ is the second smallest eigenvalue of graph Laplacian w.r.t. the kth cluster and $\mathbf{A}_k \in \mathbb{R}^{n\times n}$.

4 THEORETICAL STUDIES

This section focuses on analyzing MFC/iPursuit and revealing the key factors in its performance. First, we discuss the underlying connection between iPursuit and MFC and this interesting connection is utilized to analyze both algorithms using similar techniques. In the presented results, we utilize the parameters defined in the following definition.

Definition 1. *Suppose $\mathbf{D}$ follows Data Model 1. We define $\Delta_{\min} = \min_j\{\inf_{\substack{\|\mathbf{u}\|=1\\ \mathbf{u}\in\mathcal{S}_j}} \|\mathbf{u}^T\mathbf{D}_j\|_p^p\}_{j=1}^m$, $\dot{\Delta}_{\max} = \max_j\{\sup_{\substack{\|\mathbf{u}\|=1\\ \mathbf{u}\in\mathbb{R}^{r-s}}} \|\mathbf{u}^T\dot{\mathbf{D}}_j\|_p^p\}_{j=1}^m$, $\bar{\Delta}_{\max} = \max_j\{\sup_{\substack{\|\mathbf{u}\|=1\\ \mathbf{u}\in\mathbb{R}^{s}}} \|\mathbf{u}^T\bar{\mathbf{D}}_j\|_p^p\}_{i=1}^m$,and $\phi = \max_{j\neq t} \|\dot{\mathbf{U}}_t^T\dot{\mathbf{U}}_j\|$. In addition, when $y > x$, we define $\sigma_l(\frac{x}{y},\delta) = \frac{x-2\sqrt{x\log\frac{2M_2}{\delta}}}{y+2\sqrt{(y-x)\log\frac{2M_2}{\delta}}+2\log\frac{2M_2}{\delta}-2\sqrt{x\log\frac{2M_2}{\delta}}}$ and $\sigma_u(\frac{x}{y},\delta) = \frac{x+2\sqrt{x\log\frac{2M_2}{\delta}}+2\log\frac{2M_2}{\delta}}{y+2\sqrt{x\log\frac{2M_2}{\delta}}+2\log\frac{2M_2}{\delta}-2\sqrt{(y-x)\log\frac{2M_2}{\delta}}}$.*

The parameters $\Delta_{\min}$, $\dot{\Delta}_{\max}$, and $\bar{\Delta}_{\max}$ are similar to permeance statistic [Lerman et al., 2015] which indicates how well the data points are distributed inside the subspaces. For instance, when the columns of $\mathbf{D}_i$ in $\mathcal{S}_i$ are concentrated around a direction, the value of $\inf_{\substack{\|\mathbf{u}\|=1\\ \mathbf{u}\in\mathcal{S}_i}} \|\mathbf{u}^T\mathbf{D}_i\|_p^p$ is small in comparison to when the data points are uniformly distributed in $\mathcal{S}_i$. Although the permeance statistic appears in the presented results, it does not necessarily mean that iPursuit and MFC require a uniform distribution of data pints inside the subspaces and the reason that it appears is that the sufficient conditions guarantee the performance under the worst case scenarios. The parameter ϕ indicates how close the innovative components $\{\dot{\mathcal{S}}_i\}_{i=1}^m$ are to each other.

Remark 2. *It is important to note that ϕ only measures the affinity between the innovative components $\{\dot{\mathcal{S}}_i\}_{i=1}^m$. In other word, even if two subspaces $\mathcal{S}_i$ and $\mathcal{S}_j$ heavily intersect such that $\|\mathbf{U}_i^T\mathbf{U}_j\|_\sigma$ is nearly equal to 1, $\|\dot{\mathbf{U}}_i^T\dot{\mathbf{U}}_j\|$ could be small if the innovative components are incoherent with each other. In the following results, it is shown that in contrast to most of subspace segmentation methods whose performance depend on $\max_{j\neq t}\|\mathbf{U}_t^T\mathbf{U}_j\|_\sigma$, the performance of iPursuit and MFC mainly depends on the distance between the innovative components.*

4.1 THE CONNECTION BETWEEN IPURSUIT AND MFC

The cost function of iPursuit (2) encourages the optimal direction $\mathbf{c}_i^*$ to be orthogonal to the maximum number of data points. If the innovation assumption (Assumption 1) holds and $\mathbf{c}_i^* \in \vec{\mathcal{S}_{k_i}}$ for all the data points, then $\mathbf{A} = |\mathbf{D}^T\mathbf{C}^*|$ does not include any false connection. However, in practice the innovation assumption is not essential and $\mathbf{A} = |\mathbf{D}^T\mathbf{C}^*|$ can yield an accurate clustering of the data even if $|\mathbf{D}^T\mathbf{C}^*|$ is not a sparse matrix [Rahmani and Atia, 2017c, Ling and Strohmer, 2020]. A direct conclusion is that it may not be essential to employ ℓ_1-norm in the cost function of (2). Accordingly, in this section, we investigate an iPursuit algorithm whose i^{th} optimal direction is obtained as the optimal point of

$$\min_{\mathbf{c}} \|\mathbf{c}^T\mathbf{D}\|_2 \quad \text{subject to} \quad \mathbf{c}^T\mathbf{d}_i = 1 . \tag{3}$$

The following lemma shows that the iPursuit algorithm which employs ℓ_2-norm to compute the optimal directions is equivalent to MFC.

Lemma 1. *Define $\mathbf{C}^*$ as the optimal point of*

$$\min_{\mathbf{C}} \|\mathbf{D}^T\mathbf{C}\|_{2,1} \quad \textit{subject to} \quad \textit{diag}(\mathbf{C}^T\mathbf{D}) = \boldsymbol{1} ,$$

and define $\mathbf{A} = \left|\mathbf{D}^T\mathbf{C}^*\right|$. Then $\mathbf{A}(i,j) = \frac{|\mathbf{v}^{i\,T}\mathbf{v}^j|}{\|\mathbf{v}^i\|_2^2}$.

Lemma 1 shows that iPursuit is equivalent to MFC when ℓ_2-norm is employed to compute the optimal vectors (note that the denominator of $\mathbf{A}(i,j) = \frac{|\mathbf{v}^{i\,T}\mathbf{v}^j|}{\|\mathbf{v}^i\|_2^2}$ is the same for all the entries of a row and Step 4 of Algorithm 2 normalizes the degree of the nodes and keeps the dominant entries of each row). We leverage this connection between MFC and iPursuit to provide an analysis which is applicable to both algorithms. In the following theoretical results, p appears as a parameter in the sufficient conditions. If $p = 1$, the sufficient condition corresponds to iPursuit and if $p = 2$, then the sufficient condition corresponds to MFC.

4.2 AN ANALYSIS FOR MFC AND IPURSUIT

The following theorem provides a sufficient condition to guarantee that Requirement 1 is satisfied. The presented results are applicable to both iPursuit and MFC since it is assumed that $\mathbf{A} = |\mathbf{D}^T\mathbf{C}^*|$ where $\mathbf{C}^*$ is obtained via solving

$$\min_{\mathbf{C}} \|\mathbf{D}^T\mathbf{C}\|_{p,1} \quad \text{subject to} \quad \text{diag}(\mathbf{C}^T\mathbf{D}) = \mathbf{1} . \tag{4}$$

Theorem 2. *Suppose that $\mathbf{D}$ follows Data Model 1 and*

$\mathbf{A} = |\mathbf{D}^T \mathbf{C}^*|$ *where* $\mathbf{C}^*$ *is the optimal point of (4). If*

$$\min_i \frac{\|\beta_i\|_2^p}{\|\mathbf{d}_i\|_2^p} \Delta_{\min} \geq \dot{\Delta}_{\max} \kappa \left(\frac{1}{\kappa + (m-1)} + \frac{m-1}{\kappa + (m-1)} \phi^p \right), \tag{5}$$

then $\mathbf{A}$ *satisfies Requirement 1.*

In contrast to former theoretical results which require $\max_{j \neq t} \|\mathbf{U}_t^T \mathbf{U}_j\|_\sigma$ to be sufficiently small, the presented guarantee is concerned with $\max_{j \neq t} \|\dot{\mathbf{U}}_t^T \dot{\mathbf{U}}_j\|$ and note that $\max_{j \neq t} \|\dot{\mathbf{U}}_t^T \dot{\mathbf{U}}_j\|$ can stay small even if the subspaces have a high dimension of intersection (i.e., $\|\mathbf{U}_t^T \mathbf{U}_j\|_\sigma$ is nearly equal to 1). When m, the number of clusters, is large, the sufficient condition can be roughly simplified into $\phi^p \leq \frac{\Delta_{\min}}{\kappa \, \dot{\Delta}_{\max}} \min_i \frac{\|\beta_i\|_2^p}{\|\mathbf{d}_i\|_2^p}$, which means that the higher is the dimension of intersection, the more distanced the innovative components should be. The sufficient condition requires all the data points to have a sufficiently strong projection on the innovative component.

4.3 PROBABILISTIC GUARANTEES

In this section, we simplify the result presented in Theorem 2 in two steps. First, we presume a random model for the distribution of the data points and in the second step, we consider a random model for the generation of the subspaces. We start with the first step as follows.

Assumption 2. *Each matrix* $\mathbf{D}_i \in \mathbb{R}^{M_1 \times n}$ *is generated as* $\mathbf{D}_i = \mathbf{U}_i \mathbf{G}_i$ *where the elements of* $\mathbf{G}_i \in \mathbb{R}^{r \times n}$ *are sampled independently from* $\mathcal{N}(0, \frac{1}{\sqrt{r}})$.

Assumption 2 ensures that the distribution of $\frac{\mathbf{d}_i}{\|\mathbf{d}_i\|_2}$ is uniformly at random on $\mathbb{S}^{M_1 - 1} \cap \mathcal{S}_{k_i}$. Note that $\mathbb{E}[\|\mathbf{d}_i\|_2^2] = 1$ and in the following theorems, we do not normalize the ℓ_2-norm of the data points to make the analysis easier. In this section, we derive the guarantees for $p = 2$ and similar guarantees for $p = 1$ can be established.

Theorem 3. *Suppose* $\mathbf{D}$ *follows Data Model 1, matrices* $\{\mathbf{D}_i\}_{i=1}^m$ *are generated as in Assumption 2, and adjacency matrix* $\mathbf{A}$ *is computed as in Theorem 2 with* $p = 2$. *If*

$$(\frac{n}{r} - \eta_{\delta_r}) \sigma_l \left(\frac{r-s}{r}, \delta \right) \geq \kappa (\frac{1}{\kappa + m - 1} + \phi^2 \frac{m-1}{\kappa + m - 1}) \left(\frac{n}{r} + \frac{r-s}{r} \eta_{\delta_{r-s}} \right) \tag{6}$$

where $\eta_{\delta_x} = \max(\frac{4 z_{\delta_x}}{3} \log \frac{2\,x\,m}{\delta}, \sqrt{4 \frac{n(x+3)}{x^2} \log \frac{2xm}{\delta}})$ *and* $z_{\delta_x} = 1 + 2\sqrt{\frac{1}{x} \log \frac{2nm}{\delta}} + \frac{2}{x} \log \frac{2nm}{\delta}$, *then Requirement 1 with* $p = 2$ *is satisfied with probability at least* $1 - 5\delta$.

Theorem 3 reveals several interesting points about the requirements of the algorithms. First it confirms our intuition about the relation between the dimension of subspaces and the required number of data points. The sufficient condition states that n/r should be sufficiently large to ensure that Requirement 1 is satisfied. When n/r is sufficiently large, then $(\frac{n}{r} - \eta_{\delta_r})$ is nearly equal to n/r. Therefore, when m is large, the sufficient condition roughly states that ϕ^2 should be sufficiently smaller than $\frac{1}{\kappa} \frac{r-s}{r}$. In other word, Theorem 3 clearly indicates that the higher is the dimension of intersection, the more separable their innovative components should be. Next, we further simplify the sufficient condition via assuming a random model for the distribution of subspaces.

Theorem 4. *Suppose* $\mathbf{D}$ *and* $\mathbf{A}$ *are generated as in Theorem 3 and* $\{\dot{\mathcal{S}}_i\}_{i=1}^m$ *and* $\mathcal{S}$ *are chosen independently and uniformly at random. If*

$$\left(\frac{n}{r} - \eta_{\delta_r} \right) \sigma_l \left(\frac{r-s}{r}, \delta \right) \geq \kappa \left(\frac{n}{r} + \frac{r-s}{r} \eta_{\delta_{r-s}} \right) \left(\frac{1}{\kappa + m - 1} + \frac{c_\delta (r-s)^2}{M_1} \frac{m-1}{\kappa + m - 1} \right)$$

then Requirement 1 is satisfied with probability at least $1 - 6\delta$, *where* $c_\delta = 3 \max\left(1, \sqrt{\frac{8 M_1 \pi}{(M_1 - 1)(r-s)}}, \sqrt{\frac{16 M_1 \log \frac{mr}{\delta}}{(M_1 - 1)(r-s)}}\right)$.

If we simplify the sufficient condition, Theorem 4 roughly states that M_1 should be sufficiently larger than $\kappa r (r - s) \sqrt{\log m}$. The main reason is that the subspaces and their innovative components are generated uniformly at random and the higher is the dimension of the ambient space, the less coherent they are in expectation.

Remark 3. *The main purpose of the presented analysis is to demonstrate the key performance factors of the MFC/iPursuit algorithms and to show why they are notably robust to the strong intersection between the span of clusters. If we want to go further and use the theoretical results to compare MFC/iPursuit against the other subspace clustering algorithms, we need to analyze the other methods using the utilized criteria (Requirement 1). Although it goes beyond the scope of this paper, Section 4.5 presents a full analysis of the TSC algorithm based on Requirement 1 to show why MFC can strongly outperform TSC while their computation complexities are not much different.*

4.4 WITH THE INNOVATION ASSUMPTION

The innovation assumption (Assumption 1) is not essential in the performance of MFC/iPursuit and we did not use it in any of the presented studies. However, the innovation assumption can be utilized to establish stronger guarantees. In this section, two theorems are presented whose only difference with Theorem 2 and Theorem 4 is that they assume that Assumption 3 (stated bellow) holds.

Assumption 3. *It is assumed that* $\mathbf{D}$ *follows Data Model 1 and dim*$(\dot{\mathcal{S}}_i \cap \mathcal{P}_i) = 0$ *where* $\mathcal{P}_i = \oplus_{k \neq i} \mathcal{S}_i$.

Assumption 3 ensures that each innovative component $\dot{\mathcal{S}}_i$ is independent from the direct sum of all the other subspaces. The following theorem presumes that Assumption 3 holds.

Theorem 5. *Suppose* $\mathbf{D}$ *follows Assumption 3, define* $\vec{\mathcal{S}}_i$ *as the column space of* $(\mathbf{I}-\mathbf{P}_i\mathbf{P}_i^T)\mathbf{U}_i$*, define* $\vec{\mathbf{U}}_i$ *as a basis for* $\vec{\mathcal{S}}_i$*, and assume* $\mathbf{A}=|\mathbf{D}^T\mathbf{C}^*|$ *where* $\mathbf{C}^*$ *is the optimal point of (4). If*

$$\min_i \frac{\|\beta_i\|_2^p}{\|\mathbf{d}_i\|_2^p} \min_i \|\vec{\mathbf{U}}_{k_i}^T \dot{\mathbf{U}}_{k_i}\|_m^p \geq \frac{\kappa}{\kappa+m-1}\frac{\dot{\Delta}_{\max}}{\Delta_{\min}}, \quad (7)$$

then Requirement 1 is satisfied where $\|\vec{\mathbf{U}}_{k_i}^T \dot{\mathbf{U}}_{k_i}\|_m$ *is the minimum singular value of* $\vec{\mathbf{U}}_{k_i}^T \dot{\mathbf{U}}_{k_i}$.

The subspace $\vec{\mathcal{S}}_i$ was defined as the projection of $\mathcal{S}_i$ onto $(\oplus_{k\neq i}\mathcal{S}_k)^\perp$ which is equivalent to the projection of $\dot{\mathcal{S}}_i$ onto $(\oplus_{k\neq i}\mathcal{S}_k)^\perp$. The closer is $\dot{\mathcal{S}}_i$ to $\vec{\mathcal{S}}_i$, the more incoherent is $\dot{\mathcal{S}}_i$ with the innovative component of the other clusters since $\vec{\mathcal{S}}_i$ is orthogonal to $\oplus_{j\neq i}\dot{\mathcal{S}}_j$. This is the reason we have $\|\vec{\mathbf{U}}_{k_i}^T \dot{\mathbf{U}}_{k_i}\|_m$ on the LHS of (7) because

$$\|\vec{\mathbf{U}}_{k_i}^T \dot{\mathbf{U}}_{k_i}\|_m = \min_{\|u\|_2=1} \|\vec{\mathbf{U}}_{k_i}^T \dot{\mathbf{U}}_{k_i}\mathbf{u}\|_2$$

is a measure of coherence between $\dot{\mathcal{S}}_i$ and $\vec{\mathcal{S}}_i$. Therefore, similar to Theorem 2, Theorem 5 states that the weaker is the projection of data points onto the innovative components, the more distanced the innovative components should be. The major difference between the condition of Theorem 2 and that of Theorem 5 is that in (7) m plays a stronger role and (7) states that increasing m (provided that it does not increase the coherency between the innovative components) can enhance the chance of MFC/iPursuit to satisfy Requirement 1. The following theorem provides a more explicit sufficient condition via assuming the random data model used in Theorem 4.

Theorem 6. *Suppose* $\mathbf{D}$ *and* $\mathbf{A}$ *are generated as in Theorem 4 and assume that* $M_1 > s+(r-s)m$*. If*

$$\sigma_l\left(\frac{r-s}{r},\delta\right)\sigma_l\left(\frac{\vartheta}{M_1},\delta\right) \geq \frac{\kappa}{\kappa+m-1}\frac{\frac{n}{r}+\frac{r-s}{r}\eta_{\delta_{r-s}}}{\frac{n}{r}-\eta_{\delta_r}} \quad (8)$$

where $\vartheta = M_1 - \big(s+(r-s)(m-1)\big)$*, then Requirement 1 with* $p=2$ *is satisfied with probability at least* $1-6\delta-\epsilon$ *where* ϵ *is the probability that the rank of* $\mathbf{D}$ *is less* $s+(r-s)m$.

Note that Theorem 6 does not need to explicitly presume that Assumption 3 holds because when $M_1 > s+(r-s)m$, Assumption 3 is satisfied with an overwhelming probability [Vershynin, 2010]. The sufficient condition roughly states that when n/r is large enough, then $\frac{r-s}{s}\frac{\vartheta}{M_1}$ should be sufficiently larger than $\frac{\kappa}{\kappa+m}$ to guarantee that the requirement is satisfied with high probability. The value of $\frac{\vartheta}{M_1}$ increases when M_1 increases and it converges to 1 when r_d/M_1 decreases.

Theorem 2, Theorem 5, and Theorem 6 indicate that if $\mathbf{D}$ follows Data Model 1, then the larger is the number of clusters, the more likely it is for MFC/iPursuit to satisfy Requirement 1 provided that increasing m does not increase the coherency between $\{\dot{\mathcal{S}}_i\}_{i=1}^m$. This fact might sound counter intuitive, but it is an accurate prediction. For instance, suppose that $\mathbf{D}$ is generated as in Theorem 6, the first n columns of $\mathbf{D}$ lie in $\mathcal{S}_1$, $n=200$, $r=10$, $s=8$, and $M_1=400$. Define

$$\mathbf{a}_{\mathcal{S}_1} = \frac{1}{n}\sum_{i=1}^n \mathbf{a}_i$$

where $\mathbf{a}_i$ is the ith column of $\mathbf{A}$. Therefore, $\mathbf{a}_{\mathcal{S}_1}$ is the average of the first n columns of $\mathbf{A}$ which are corresponding to data points in $\mathcal{S}_1$. Figure 1 shows $\mathbf{a}_{\mathcal{S}_1}$ with different values of m for the adjacency matrices computed by MFC and the TSC algorithm [Heckel and Bölcskei, 2013] which computes $\mathbf{A}=|\mathbf{D}^T\mathbf{D}|$. Ideally, we should observe that the expected value of the elements of $\mathbf{a}_{\mathcal{S}_1}[1:n]$ are sufficiently larger than the expected value of the elements of $\mathbf{a}_{\mathcal{S}_1}[n:M_2]$. One can observe that when $\mathbf{A}=|\mathbf{D}^T\mathbf{D}|$, the elements of $\mathbf{a}_{\mathcal{S}_1}[1:n]$ are not much distinguishable from the elements of $\mathbf{a}_{\mathcal{S}_1}[n:M_2]$ with both $m=2$ and $m=10$. In contrast, when $\mathbf{A}=|\mathbf{D}^T\mathbf{C}^*|$ and when $m=10$, $\|\mathbf{a}_{\mathcal{S}_1}[1:n]\|_2$ is clearly larger than $\frac{1}{m-1}\|\mathbf{a}_{\mathcal{S}_1}[n:M_2]\|_2$. The last plot of Figure 1 shows the effect of m on the quality of the computed adjacency matrix in a more clear way. Define parameter $\hat{\kappa}$ as follows

$$\kappa' = \frac{(m-1)\ \|\mathbf{a}_{\mathcal{S}_1}[1:n]\|_2^2}{\|\mathbf{a}_{\mathcal{S}_1}[n:M_2]\|_2^2}. \quad (9)$$

Parameter κ' shows how clear the adjacency matrix separates the data points in $\mathcal{S}_1$ from the other clusters. The last plot (first from right), shows κ' versus m for both MFC and TSC. One can observe that κ' notably increases as m increases when $\mathbf{A}=|\mathbf{D}^T\mathbf{C}^*|$ which means that the quality of the estimated adjacency matrix improves as m increases. In sharp contrast, increasing m does not show a positive/negative impact on the computed adjacency matrix by Algorithm 3.

It is important to note that the conclusion that the performance of MFC/iPursuit improves if m increases is not a general rule. When M_1 is not sufficiently large, as m increases, the distance between the subspaces (and the distance between their innovative components) decreases and it degrades the performance of the algorithms. Moreover, the reason that in Theorem 4 and Theorem 6 the coherency between the subspaces decreases as M_1 increases is due to the presumed model for the generation of the subspaces and it is not a general rule that ϕ decreases as M_1 increases.

Algorithm 3 Inner-Product based Subspace Clustering [Heckel and Bölcskei, 2013] (TSC Algorithm)

Input. The input is data matrix $\mathbf{D} \in \mathbb{R}^{M_1 \times M_2}$.

1. Data Preprocessing. Normalize the ℓ_2-norm of the columns of $\mathbf{D}$, i.e., set $\mathbf{d}_i$ equal to $\mathbf{d}_i / \|\mathbf{d}_i\|_2$ for all $1 \leq i \leq M_2$.
2. Define $\mathbf{A} = \left|\mathbf{D}^T \mathbf{D}\right|$.
3. Similar to Step 4 in Algorithm 1.
4. Similar to Step 5 in Algorithm 1.
Output: The identified clusters.

4.5 COMPARISON WITH THE TSC ALGORITHM

In this section, we theoretically compare the TSC algorithm against against MFC/iPursuit. Both MFC/iPursuit and Algorithm 3 use inner-product as the kernel function to measure the similarity between data points. However, in sharp contrast to Algorithm 3, MFC/iPursuit computes the inner-product between the directions of innovation and the data points as opposed to computing the inner-product between the data points. In [Rahmani and Atia, 2017b,c] and in this paper, it is shown that this difference makes MFC/iPursuit able to notably outperform TSC in most of scenarios. In order to clarify the reason behind this performance difference, we provide similar analysis for Algorithm 3 and we compare the requirements of MFC/iPursuit against those of Algorithm 3. Although the presented theorems only include sufficient conditions (not necessary conditions), their comparison is insightful.

Theorem 7. *Suppose* $\mathbf{D}$ *follows Data Model 1. If*

$$\begin{aligned} 1 \geq \kappa \max_i \left\{ \frac{\|\alpha_i\|_2^p}{\|\mathbf{d}_i\|_2^p} \right\} \frac{\bar{\Delta}_{\max}}{\Delta_{\min}} + \\ \kappa \max_i \left\{ \frac{\|\beta_i\|_2^p}{\|\mathbf{d}_i\|_2^p} \right\} \phi^p \frac{\dot{\Delta}_{\max}}{\Delta_{\min}} \ , \end{aligned} \tag{10}$$

then $\mathbf{A} = |\mathbf{D}^T \mathbf{D}|$ *satisfies Requirement (1).*

There are two terms on the RHS of the sufficient condition where only the second term is weighted by ϕ. Even in the best case scenario where the innovative components are orthogonal to each other, i.e., $\phi = 0$, it may not be possible to satisfy the sufficient condition. For instance, suppose s/r is nearly equal to one and assume that the elements of β_i and α_i are sampled independently from $\mathcal{N}(0,1)$. In this scenario, $\mathbb{E}\left[\frac{\|\alpha_i\|_2^2}{\|\mathbf{d}_i\|_2^2}\right] = \frac{s}{m} \approx 1$ and it may not be possible to satisfy the sufficient condition even for $\kappa = 2$. The main reason is that when s/m is high, the inner-product value between data points in different clusters are high, no matter how well separated the innovative components are. In sharp contrast to Algorithm 3, MFC/iPursuit utilize the inner-product between the optimal directions and the data points to construct the adjacency matrix and when s/m is high, the optimal directions are strongly incoherent with $\mathcal{S}$ and this feature makes the role of the innovative components notably more significant. In order to make a more explicit comparison, we derive the sufficient condition for Algorithm 3 while it is assumed that the data is generated as in Theorem 4. The following theorem provides the result.

Theorem 8. *Suppose* $\mathbf{D}$ *is generated as in Theorem 4 and* $\mathbf{A} = |\mathbf{D}^T \mathbf{D}|$. *If*

$$\begin{aligned} &\left(\frac{n}{r} - \eta_{\delta_r}\right) \geq \kappa\, \sigma_u\left(\frac{s}{r}, \delta\right)\left(\frac{n}{r} + \frac{s}{r}\, \eta_{\delta_s}\right) + \\ &\kappa\, \sigma_u\left(\frac{r-s}{r}, \delta\right)\left(\frac{n}{r} + \frac{r-s}{r}\, \eta_{\delta_{r-s}}\right)\left(\frac{c_\delta (r-s)^2}{M_1}\right) \ , \end{aligned}$$

then Requirement 1 with $p = 2$ *is satisfied with probability at least* $1 - 9\delta$*, where* c_δ *was defined in Theorem 4 and* η_{δ_x} *was defined in Theorem 3.*

The first term on the RHS of the sufficient condition of Theorem 8 is the dominant term when s is large. When there are a sufficiently large number of data points in the clusters (n/r is large enough), the sufficient condition roughly states that $\kappa\, \frac{s}{r}$ should be sufficiently smaller than 1. However, it is not feasible to satisfy this condition in many scenarios. For instance, if we choose $\kappa = 2$, then the sufficient condition can be satisfied only when $s/r > 0.5$.

In summary, comparing the sufficient conditions suggests that in sharp contrast to Algorithm 3 which fails when the span of clusters are close, MFC/iPursuit can effectively leverage the innovative components of the clusters and if these innovative components are sufficiently separable (ϕ is sufficiently small), MFC/iPursuit might successfully distinguish the clusters.

5 NUMERICAL EXPERIMENTS

This paper does not present a new clustering algorithm and the main focus was to provide a deep understating and analysis of the MFC/iPursuit algorithms. We refer the reader to [Kanatani, 2001, Costeira and Kanade, 1998, Boult and Brown, 1991, Vidal, 2011, Rahmani and Atia, 2017a,b] for numerical studies of the MFC/iPursuit algorithms. The focus of the presented experiments are to demonstrate some of the features of the algorithms which was predicted by the presented theoretical studies. For iPursuit, MFC, and TSC, the graph preprocessing step (Step 4 in Algorithm 1) was done as follows. For each column of $\mathbf{A}$, 8 largest elements were kept and the rest of elements were set to zero. The clustering error is defined as $\frac{N_e}{M_2}$ where N_e is the total number of misclassified data points. In the appendix, we have included a simple numerical experiment showing that exact clustering can be achieved if Requirement 1 holds even for small values of κ.

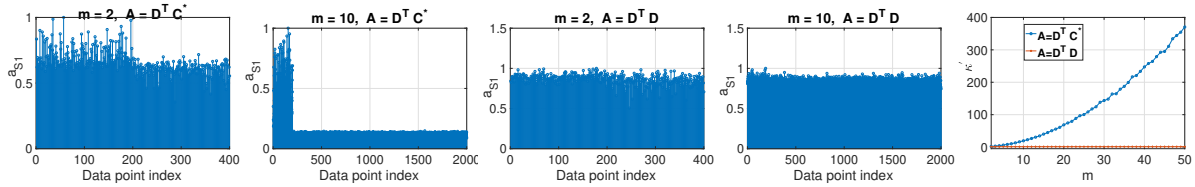

Figure 1: The first 4 plots (from LHS) show the elements of $\mathbf{a}_{\mathcal{S}_1} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{a}_i$ with different number of clusters for MFC and Algorithm 3. The first $n = 200$ data points lie in first cluster, $r = 10$, $s = 8$, and $M_1 = 400$. The last plot demonstrates parameter $\kappa^{'}$ defined in (9) versus m. One can observe that in this experiment increasing m improves the quality of the adjacency matrix computed by MFC.

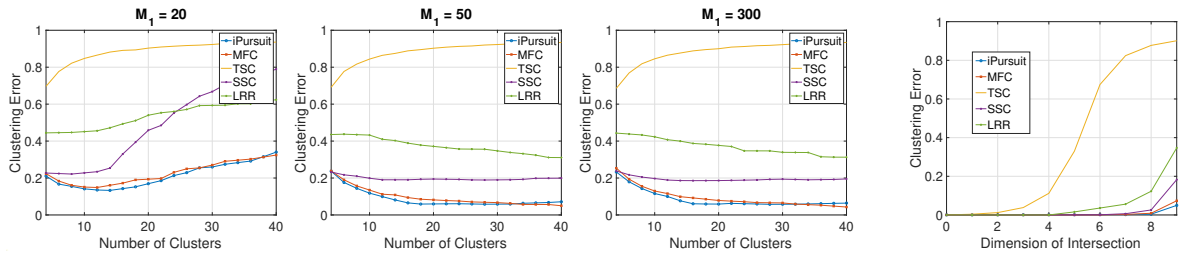

Figure 2: First three plots from left: Clustering error versus the number of clusters for different values of M_1 where $r = 10$, $s = 9$, and $n = 100$. First plot from right: This plot demonstrates clustering error versus s. In this experiment, $M_1 = 40$, $r = 10$, and $n = 100$.

5.1 THE DIMENSION OF INTERSECTION BETWEEN THE SUBSPACES

In the presented deterministic results (Theorem 2 and Theorem 5), we observed that $\frac{\|\beta_i\|_2}{\|\mathbf{d}_i\|_2}$ is an important factor in the performance of MFC/iPursuit and in the probabilistic results, this factor appeared as $\frac{r-s}{r}$. The purpose of this experiment is twofold. Firstly, we show that the accuracy of MFC/iPursuit degrades as s increases (since $\frac{r-s}{r}$ decreases). Secondly, it is shown that MFC/iPursuit are notably robust against intersection between the span of clusters comparing to most of other methods. The first plot (from right) in Figure 2 shows clustering error versus s where in this experiment $M_1 = 40$, $r = 10$, and $n = 100$ (the number of evaluation runs was 50). One can observe that the accuracy of MFC/iPursuit degrades as s increases. However, both of them notably outperform the other methods when s is high. The main reason is that as the presented theoretical studies indicated, the performance of MFC/iPursuit mainly depends on the coherency between the innovative components $\{\dot{\mathcal{S}}_i\}_{i=1}^{m}$ while most of other algorithms such as TSC require the span of clusters $\{\mathcal{S}_i\}_{i=1}^{m}$ to be sufficiently incoherent.

5.2 NUMBER OF CLUSTERS

In the theoretical results (Theorem 2 and Theorem 5), it was shown that the quality of the adjacency matrix computed by MFC/iPursuit might improve when m increases. Specifically, the theoretical results suggested that when data follows Data Model 1 and as long as increasing m does not increase the coherency between $\{\dot{\mathcal{S}}_i\}_{i=1}^{m}$, MFC/iPursuit can yield a better adjacency matrix (an adjacency matrix with higher $\min_i \frac{(m-1)\|\mathbf{a}_{i\,\mathcal{I}_i}\|_p^p}{\|\mathbf{a}_{i\,\mathcal{I}_i^{\perp}}\|_p^p}$) if m increases.

The first three plots (from left) in Figure 2 shows clustering error versus m for different values of M_1 where in this experiment $r = 10$, $s = 9$, and $n = 100$ (the number of evaluation runs was 50). One can observe that when $M_1 = 300$ and when $M_1 = 50$, the accuracy of MFC/iPursuit improves when m increases while when $M_1 = 20$, the accuracy degrades. The reason for this observation is that as the theoretical results indicated, both the number of clusters and the coherency between the innovative components contribute to the performance of the algorithms. When M_1 is not sufficiently large, increasing m increases the coherency between the innovative components and it degrades the performances of the algorithms.

CONCLUSION

It was shown that iPursuit is equivalent to a closed form matrix factorization based clustering algorithm if the direction search optimization problem is altered into a quadratic optimization problem. A novel analysis applicable to both algorithms were proposed which showed that in contrast to some of the other subspace clustering algorithms whose performance depend on the distance between the span of clusters, the performance of MFC/iPursuit mainly depends on the distance between the innovative components of the clusters.

References

Terrance E Boult and L Gottesfeld Brown. Factorization-based segmentation of motions. In *Proceedings of the IEEE workshop on visual motion*, pages 179–180. IEEE Computer Society, 1991.

Paul S Bradley and Olvi L Mangasarian. k-plane clustering. *Journal of Global Optimization*, 16(1):23–32, 2000.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

Guangliang Chen and Gilad Lerman. Spectral curvature clustering (SCC). *International Journal of Computer Vision*, 81(3):317–330, 2009.

João Paulo Costeira and Takeo Kanade. A multibody factorization method for independently moving objects. *International Journal of Computer Vision*, 29(3):159–179, 1998.

Eva L Dyer, Aswin C Sankaranarayanan, and Richard G Baraniuk. Greedy feature selection for subspace clustering. *The Journal of Machine Learning Research*, 14(1): 2487–2517, 2013.

Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (11):2765–2781, 2013.

Jiashi Feng, Zhouchen Lin, Huan Xu, and Shuicheng Yan. Robust subspace segmentation with block-diagonal prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3818–3825, 2014.

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

Hongchang Gao, Feiping Nie, Xuelong Li, and Heng Huang. Multi-view subspace clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4238–4246, 2015.

Reinhard Heckel and Helmut Bölcskei. Robust subspace clustering via thresholding. *arXiv preprint arXiv:1307.4891*, 2013.

Pan Ji, Tong Zhang, Hongdong Li, Mathieu Salzmann, and Ian Reid. Deep subspace clustering networks. *Advances in neural information processing systems*, 30:24–33, 2017.

Hao Jiang, Daniel P Robinson, Rene Vidal, and Chong You. A nonconvex formulation for low rank subspace clustering: algorithms and convergence analysis. *Computational Optimization and Applications*, 70(2):395–418, 2018.

Marius Junge, Qiang Zeng, et al. Noncommutative bennett and rosenthal inequalities. *The Annals of Probability*, 41 (6):4287–4316, 2013.

Ken-ichi Kanatani. Motion segmentation by subspace separation and model selection. In *Proceedings Eighth IEEE International Conference on computer Vision. ICCV 2001*, volume 2, pages 586–591. IEEE, 2001.

Jack Klys, Jake Snell, and Richard Zemel. Learning latent subspaces in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 6444–6454, 2018.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2005.

Gilad Lerman, Michael B McCoy, Joel A Tropp, and Teng Zhang. Robust computation of linear models by convex relaxation. *Foundations of Computational Mathematics*, 15(2):363–410, 2015.

Weiwei Li, Mostafa Rahmani, and Ping Li. Provable data clustering via innovation search. *arXiv preprint arXiv:2108.06888*, 2021.

Shuyang Ling and Thomas Strohmer. Certifying global optimality of graph cuts via semidefinite relaxation: A performance guarantee for spectral clustering. *Foundations of Computational Mathematics*, 20(3):367–421, 2020.

John Lipor, David Hong, Yan Shuo Tan, and Laura Balzano. Subspace clustering using ensembles of k-subspaces. *Information and Inference: A Journal of the IMA*, 10(1): 73–107, 2021.

Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013.

Canyi Lu, Jiashi Feng, Zhouchen Lin, and Shuicheng Yan. Correlation adaptive subspace segmentation by trace lasso. In *Proceedings of the IEEE international conference on computer vision*, pages 1345–1352, 2013.

Vishnu Menon, Gokularam Muthukrishnan, and Sheetal Kalyani. Subspace clustering without knowing the number of clusters: A parameter free approach. *IEEE Transactions on Signal Processing*, 68:5047–5062, 2020.

Dohyung Park, Constantine Caramanis, and Sujay Sanghavi. Greedy subspace clustering. In *Advances in Neural Inf. Processing Systems*, pages 2753–2761, 2014.

Vishal M Patel, Hien Van Nguyen, and Rene Vidal. Latent space sparse subspace clustering. In *Proceedings of the IEEE international conference on computer vision*, pages 225–232, 2013.

Xi Peng, Shijie Xiao, Jiashi Feng, Wei-Yun Yau, and Zhang Yi. Deep subspace clustering with sparsity prior. In *IJCAI*, pages 1925–1931, 2016.

Mostafa Rahmani and George Atia. Innovation pursuit: A new approach to the subspace clustering problem. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2874–2882, 2017a.

Mostafa Rahmani and George K Atia. Innovation pursuit: A new approach to subspace clustering. *IEEE Transactions on Signal Processing*, 65(23):6276–6291, 2017b.

Mostafa Rahmani and George K Atia. Subspace clustering via optimal direction search. *IEEE Signal Processing Letters*, 24(12):1793–1797, 2017c.

Mahdi Soltanolkotabi, Emmanuel J Candes, et al. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–2238, 2012.

Yasuyuki Sugaya and Kenichi Kanatani. Geometric structure of degeneracy for multi-body motion segmentation. In *Statistical Methods in Video Processing*, pages 13–25. Springer, 2004.

Michael E Tipping and Chris M Bishop. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482, 1999.

Manolis C. Tsakiris and Rene Vidal. Hyperplane clustering via dual principal component pursuit. In *International Conference on Machine Learning, ICML*, volume 70, pages 3472–3481, 2017.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Rene Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 2(28):52–68, 2011.

Rene Vidal, Yi Ma, and Shankar Sastry. Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12): 1945–1959, 2005.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Yu-Xiang Wang and Huan Xu. Noisy sparse subspace clustering. *The Journal of Machine Learning Research*, 17 (1):320–360, 2016.

Yu-Xiang Wang, Huan Xu, and Chenlei Leng. Provable subspace clustering: When lrr meets ssc. *Advances in Neural Information Processing Systems*, 26:64–72, 2013.

Allen Y Yang, Shankar R Rao, and Yi Ma. Robust statistical estimation and segmentation of multiple subspaces. In *Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 99–99, 2006.

Chong You, Daniel Robinson, and Rene Vidal. Scalable sparse subspace clustering by orthogonal matching pursuit. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3918–3927, 2016.

Liheng Zhang, Marzieh Edraki, and Guo-Jun Qi. Cappronet: Deep feature learning via orthogonal projections onto capsule subspaces. In *Advances in Neural Information Processing Systems*, pages 5814–5823, 2018.

Teng Zhang and Gilad Lerman. A novel m-estimator for robust PCA. *J. of Machine Learning Research*, 15(1): 749–808, 2014.

.

A APPENDIX

A.1 REQUIREMENT 1

We discussed the fact that Requirement 1 indicates how clear the estimated adjacency matrix represents the clustering structure of the data and it is similar to the sufficient condition established in [Ling and Strohmer, 2020] which guarantees that the spectral clustering algorithm can yield exact clustering. In this experiment, we assume that $m = 4$ and $n = 100$ which means $M_2 = 400$. In order to construct $\mathbf{A}$, we sample each element of $\mathbf{A}$ from half-normal distribution and we normalize the elements such that $\frac{\kappa}{m-1}\|\mathbf{a}_{i\mathcal{I}_i^{\perp}}\|_1 = \|\mathbf{a}_{i\mathcal{I}_i}\|_1$, for all $1 \leq i \leq M_2$. Figure 3 shows clustering error of the spectral clustering algorithm versus κ. One can observe that even a small value of κ can guarantee exact clustering. Although the minimum value of κ for which we can guarantee exact clustering depends on the distribution of the elements of $\mathbf{A}$, but it shows that (as the results in [Ling and Strohmer, 2020] suggests), exact clustering can be achieved if the false connections are sufficiently weaker than the true connections.

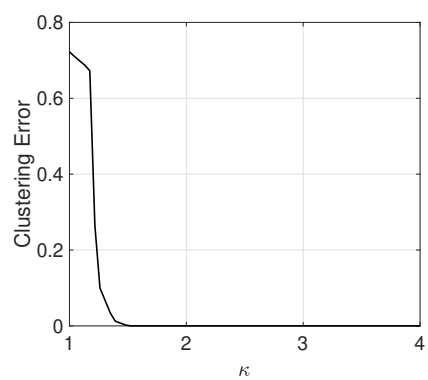

Figure 3: Clustering error versus parameter κ.

B PROOF OF THE PRESENTED THEORETICAL RESULTS

In this section, the proofs for the presented results are presented.

Proof of Lemma 1
The optimal point of (3) is equivalent to the optimal point of

$$\min_{\mathbf{c}} \ \mathbf{c}^T\mathbf{D}\mathbf{D}^T\mathbf{c} \quad \text{subject to} \quad \mathbf{c}^T\mathbf{d}_i = 1 \,, \tag{11}$$

whose Lagrangian function is as follows

$$\mathbf{c}^T\mathbf{D}\mathbf{D}^T\mathbf{c} + \gamma(\mathbf{c}^T\mathbf{d}_i - 1) \,, \tag{12}$$

where γ is the Lagrangian multiplier. Therefore, an optimal solution of (3) which lies in the column space of $\mathbf{D}$ is equal to

$$\mathbf{c}_i^* = \frac{\mathbf{U}\Sigma^{-2}\mathbf{U}^T\,\mathbf{d}_i}{\mathbf{d}_i^T\,\mathbf{U}\Sigma^{-2}\mathbf{U}^T\,\mathbf{d}_i} \,. \tag{13}$$

In addition, $\mathbf{d}_i = \mathbf{U}\Sigma\mathbf{v}^i$ where $\mathbf{v}^i$ is the ith row of $\mathbf{V} \in \mathbb{R}^{M_2\times r_d}$. Accordingly,

$$\mathbf{c}_i^* = \frac{\mathbf{U}\Sigma^{-1}\mathbf{v}^i}{\|\mathbf{v}^i\|_2^2}$$

and $\mathbf{d}_j^T\mathbf{c}_i^* = \frac{\mathbf{v}^{j\,T}\mathbf{v}^i}{\|\mathbf{v}^i\|_2^2}$.

Proof of Theorem 2
In the MFC/iPursuit method, the ith column of the estimated adjacency is equal to $\mathbf{D}^T\mathbf{c}_i^*$. In order to guarantee that Requirement 1 holds, it is sufficient to show that

$$\|\mathbf{D}_{k_i}^T\mathbf{c}_i^*\|_p^p > \frac{\kappa}{\kappa + (m-1)}\|\mathbf{D}^T\mathbf{c}_i^*\|_p^p \tag{14}$$

holds for all the columns. In order to guarantee that this inequality holds, we establish a lower-bound for the LHS and an upper-bound for the RHS and we derive the final sufficient condition such that the lower-bound is larger than the upper-bound.

First we derive a lower bound for $\|\mathbf{D}_{k_i}^T\mathbf{c}_i^*\|_p^p$. According to the linear constraint,

$$\mathbf{c}_i^{*T}\mathbf{d}_i = 1 \,,$$

and since $\mathbf{d}_i \in \mathcal{S}_{k_i}$, we conclude that

$$\|\mathbf{U}_{k_i}^T\mathbf{c}_i^*\|_2 \geq \frac{1}{\|\mathbf{d}_i\|_2} \,. \tag{15}$$

All the columns of $\mathbf{D}_{k_i}$ lie in $\mathcal{S}_{k_i}$. Thus,

$$\|\mathbf{D}_{k_i}^T\mathbf{c}_i^*\|_p = \|\mathbf{D}_{k_i}^T\mathbf{U}_{k_i}\mathbf{U}_{k_i}^T\mathbf{c}_i^*\|_p \,, \tag{16}$$

and according to (15),

$$\|\mathbf{D}_{k_i}^T\mathbf{c}_i^*\|_p^p \geq \frac{1}{\|\mathbf{d}_i\|_2^p} \inf_{\substack{\|\mathbf{u}\|=1\\ \mathbf{u}\in\mathcal{S}_{k_i}}} \|\mathbf{u}^T\mathbf{D}_{k_i}\|_p^p \,. \tag{17}$$

Using (17) we can derive the following lower-bound

$$\begin{aligned}\|\mathbf{D}_{k_i}^T\mathbf{c}_i^*\|_p^p &\geq \frac{1}{\|\mathbf{d}_i\|_2^p}\min_j \left\{ \inf_{\substack{\|\mathbf{u}\|=1\\ \mathbf{u}\in\mathcal{S}_j}} \|\mathbf{u}^T\mathbf{D}_j\|_p^p \right\}_{j=1}^m \\ &= \frac{1}{\|\mathbf{d}_i\|_2^p}\,\Delta_{\min} \,.\end{aligned} \tag{18}$$

Next, we establish an upper-bound for the RHS of (14). Corresponding to each data point $\mathbf{d}_i$, we define vector $\mathbf{d}_i^\perp$ as

$$\mathbf{d}_i^\perp = \frac{\dot{\mathbf{U}}_{k_i}\beta_i}{\|\beta_i\|_2^2} \,, \tag{19}$$

where $\beta_i \in \mathbb{R}^{r-s}$ was defined in (1). Note that

$$\mathbf{d}_i^T\mathbf{d}_i^\perp = \frac{(\mathbf{S}\alpha_i + \dot{\mathbf{U}}_{k_i}\beta_i)^T\dot{\mathbf{U}}_{k_i}\beta_i}{\|\beta_i\|_2^2} = 1 \,.$$

Since $\mathbf{d}_i^\perp$ is in the feasible set of

$$\min_{\mathbf{c}} \ \|\mathbf{c}^T\mathbf{D}\|_p^p \quad \text{subject to} \qquad \mathbf{c}^T\mathbf{d}_i = 1 \,, \tag{20}$$

we can conclude that

$$\|\mathbf{D}^T\mathbf{c}_i^*\|_p \leq \|\mathbf{D}^T\mathbf{d}_i^\perp\|_p \,. \tag{21}$$

In addition,

$$\|\mathbf{D}^T\mathbf{d}_i^\perp\|_p^p = \|\mathbf{D}_{k_i}^T\mathbf{d}_i^\perp\|_p^p + \|\mathbf{D}_{-k_i}^T\mathbf{d}_i^\perp\|_p^p \,. \tag{22}$$

The value of $\|\mathbf{D}_{k_i}^T\mathbf{d}_i^\perp\|_p^p$ can be upper-bounded as

$$\begin{aligned}\|\mathbf{D}_{k_i}^T\mathbf{d}_i^\perp\|_p^p &= \|\mathbf{d}_i^\perp\|_2^p\|\mathbf{D}_{k_i}^T(\mathbf{d}_i^\perp/\|\mathbf{d}_i^\perp\|_2)\|_p^p \\ &\leq \|\mathbf{d}_i^\perp\|_2^p \sup_{\|\mathbf{u}\|=1} \|\mathbf{u}^T\dot{\mathbf{D}}_{k_i}\|_p^p \\ &\leq \|\mathbf{d}_i^\perp\|_2^p\,\dot{\Delta}_{\max} \,.\end{aligned} \tag{23}$$

The columns of $\mathbf{D}_{-k_i}$ lies in a union of $m-1$ linear manifolds and we can rewrite $\|\mathbf{D}_{-k_i}^T\mathbf{d}_i^\perp\|_p^p$ as

$$\|\mathbf{D}_{-k_i}^T\mathbf{d}_i^\perp\|_p^p = \sum_{j\neq k_i} \|\mathbf{D}_j^T\mathbf{d}_i^\perp\|_p^p \,, \tag{24}$$

using which we can conclude that

$$\|\mathbf{D}_{-k_i}^T\mathbf{d}_i^\perp\|_p^p \leq (m-1)\max_{\substack{j\\ j\neq k_i}}(\|\mathbf{D}_j^T\mathbf{d}_i^\perp\|_p^p) \,. \tag{25}$$

Define

$$x = \arg\max_{\substack{j\\ j\neq k_i}} \|\mathbf{D}_j^T\mathbf{d}_i^\perp\|_p^p \,.$$

Note that

$$\begin{aligned}\mathbf{D}_x^T\mathbf{d}_i^\perp &= \left(\mathbf{S}\mathbf{S}^T\mathbf{D}_x + \dot{\mathbf{U}}_x\dot{\mathbf{U}}_x^T\mathbf{D}_x\right)^T\mathbf{d}_i^\perp \\ &= (\dot{\mathbf{U}}_x\dot{\mathbf{U}}_x^T\mathbf{D}_x)^T\mathbf{d}_i^\perp \,.\end{aligned} \tag{26}$$

Accordingly, we can rewrite $\|\mathbf{D}_x^T \mathbf{d}_i^\perp\|_p^p$ as

$$\begin{aligned}
&\|\mathbf{D}_x^T \mathbf{d}_i^\perp\|_p^p = \|(\dot{\mathbf{U}}_x \dot{\mathbf{U}}_x^T \mathbf{D}_x)^T \mathbf{d}_i^\perp\|_p^p \\
&= \|(\dot{\mathbf{U}}_x \dot{\mathbf{U}}_x^T \mathbf{D}_x)^T \dot{\mathbf{U}}_{k_i} \dot{\mathbf{U}}_{k_i}^T \mathbf{d}_i^\perp\|_p^p \\
&\leq \|\mathbf{d}_i^\perp\|_2^p \|\dot{\mathbf{U}}_{k_i}^T \dot{\mathbf{U}}_x\|^p \sup_{\substack{\|\mathbf{u}\|=1 \\ \mathbf{u}\in\mathbb{R}^{r-s}}} \|\mathbf{u}^T \dot{\mathbf{U}}_x^T \mathbf{D}_x\|_p^p \\
&\leq \|\mathbf{d}_i^\perp\|_2^p \max_{j\neq t} \|\dot{\mathbf{U}}_t^T \dot{\mathbf{U}}_j\|^p \max_j \left(\sup_{\|\mathbf{u}\|=1} \|\mathbf{u}^T \dot{\mathbf{D}}_j\|_p^p \right) .
\end{aligned} \tag{27}$$

Therefore, according to (27) and (25),

$$\|\mathbf{D}_{-k_i}^T \mathbf{d}_i^\perp\|_p^p \leq (m-1) \|\mathbf{d}_i^\perp\|_2^p \, \phi^p \, \dot{\Delta}_{\max} , \tag{28}$$

where

$$\|\mathbf{d}_i^\perp\|_2^p = \frac{1}{\|\beta_i\|_2^p} . \tag{29}$$

According to (18), (23), and (28), if

$$\min_i \frac{\|\beta_i\|_2^p}{\|\mathbf{d}_i\|_2^p} \Delta_{\min} \geq \dot{\Delta}_{\max} \left(\frac{\kappa}{\kappa + (m-1)} + \kappa \frac{m-1}{\kappa + (m-1)} \phi^p \right),$$

then Requirement 1 is satisfied.

Proof of Theorem 7

In order to prove that Requirement1 1 is satisfied, it is enough to prove that

$$\|\mathbf{d}_i^T \mathbf{D}_{k_i}\|_p^p \geq \frac{\kappa}{m-1} \|\mathbf{d}_i^T \mathbf{D}_{-k_i}\|_p^p . \tag{30}$$

In order to ensure that (30) holds, we derive the sufficient condition which guarantees that a lower-bound on the LHS of (30) is larger than or equal to an upper-bound on the RHS of (30).

First we derive a lower-bound on the LHS. Note that $\mathbf{d}_i$ and the columns of $\mathbf{D}_{k_i}$ lie in the same cluster. Accordingly,

$$\begin{aligned}
\|\mathbf{d}_i^T \mathbf{D}_{k_i}\|_p^p &\geq \|\mathbf{d}_i\|_2^p \inf_{\substack{\|\mathbf{u}\|=1 \\ \mathbf{u}\in\mathcal{S}_{k_i}}} \|\mathbf{u}^T \mathbf{D}_{k_i}\|_p^p \\
&\geq \|\mathbf{d}_i\|_2^p \min_j \left\{ \inf_{\substack{\|\mathbf{u}\|=1 \\ \mathbf{u}\in\mathcal{S}_{k_j}}} \|\mathbf{u}^T \mathbf{D}_{k_j}\|_p^p \right\} .
\end{aligned} \tag{31}$$

Next, we derive an upper-bound on the RHS of (30). The matrix $\mathbf{D}_{-k_i}$ contains the columns of matrices $\{\mathbf{D}_j\}_{j\neq k_i}$. Therefore,

$$\begin{aligned}
\|\mathbf{d}_i^T \mathbf{D}_{-k_i}\|_p^p &= \sum_{j\neq k_i} \|\mathbf{d}_i^T \mathbf{D}_j\|_p^p \\
&\leq (m-1) \max_{\substack{j \\ j\neq k_i}} \left\{ \|\mathbf{d}_i^T \mathbf{D}_j\|_p^p \right\} .
\end{aligned} \tag{32}$$

Define

$$x = \arg\max_{\substack{j \\ j\neq k_i}} \|\mathbf{d}_i^T \mathbf{D}_j\|_p . \tag{33}$$

According to the presumed data model, $\mathbf{d}_i^T \mathbf{D}_x$ can be expanded as follows

$$\begin{aligned}
\mathbf{d}_i^T \mathbf{D}_x &= (\mathbf{S}\alpha_i + \dot{\mathbf{U}}_{k_i}\beta_i)^T (\mathbf{S}\mathbf{S}^T \mathbf{D}_x + \dot{\mathbf{U}}_x \dot{\mathbf{U}}_x^T \mathbf{D}_x) \\
&= \alpha_i^T \mathbf{S}^T \mathbf{D}_x + \alpha_i^T (\mathbf{S}^T \dot{\mathbf{U}}_x) \dot{\mathbf{D}}_x + \beta_i^T (\dot{\mathbf{U}}_{k_i}^T \mathbf{S}) \mathbf{S}^T \mathbf{D}_x \\
&\quad + \beta_i^T (\dot{\mathbf{U}}_{k_i}^T \dot{\mathbf{U}}_x) \dot{\mathbf{D}}_x \\
&= \alpha_i^T \mathbf{S}^T \mathbf{D}_x + \beta_i^T (\dot{\mathbf{U}}_{k_i}^T \dot{\mathbf{U}}_x) \dot{\mathbf{D}}_x
\end{aligned} \tag{34}$$

According to (34),

$$\|\mathbf{d}_i^T \mathbf{D}_x\|_p^p \leq \|\alpha_i^T \mathbf{S}^T \mathbf{D}_x\|_p^p + \|\beta_i^T (\dot{\mathbf{U}}_{k_i}^T \dot{\mathbf{U}}_x) \dot{\mathbf{D}}_x\|_p^p . \tag{35}$$

Next, we establish an upper-bound on each component on the RHS of (35). First, we bound $\alpha_i^T \mathbf{S}^T \mathbf{D}_x$ as follows

$$\begin{aligned}
&\|\alpha_i^T \mathbf{S}^T \mathbf{D}_x\|_p^p = \|\alpha_i^T \bar{\mathbf{D}}_x\|_p^p = \|\alpha_i\|_2^p \left\| \frac{\alpha_i^T \bar{\mathbf{D}}_x}{\|\alpha_i\|_2} \right\|_2^p \\
&\leq \|\alpha_i\|_2^p \sup_{\substack{\|\mathbf{u}\|=1 \\ \mathbf{u}\in\mathbb{R}^s}} \|\mathbf{u}^T \bar{\mathbf{D}}_x\|_p^p \\
&\leq \|\alpha_i\|_2^p \max_j \left\{ \sup_{\substack{\|\mathbf{u}\|=1 \\ \mathbf{u}\in\mathbb{R}^s}} \|\mathbf{u}^T \bar{\mathbf{D}}_j\|_p^p \right\} .
\end{aligned} \tag{36}$$

Finally, the last term of the RHS of (35) is upper-bounded as follows,

$$\begin{aligned}
&\|\beta_i^T (\dot{\mathbf{U}}_{k_i}^T \dot{\mathbf{U}}_x) \dot{\mathbf{D}}_x\|_p^p \leq \\
&\quad \|\beta_i\|_2^p \max_{j\neq k} \left\{ \|\dot{\mathbf{U}}_k^T \dot{\mathbf{U}}_j\|^p \right\} \max_k \left\{ \sup_{\substack{\|\mathbf{u}\|=1 \\ \mathbf{u}\in\mathbb{R}^{r-s}}} \|\mathbf{u}^T \dot{\mathbf{D}}_k\|_p^p \right\} .
\end{aligned} \tag{37}$$

Therefore, according to (31), (36), and (37), if (10) holds, then Requirement 1 is satisfied.

Proof of Theorem 3

In order to prove that Requirement 1 is satisfied, we only need to guarantee that (5) holds. Therefore, a lower bound for $\Delta_{\min}$ and $\min_i \frac{\|\beta_i\|_2^p}{\|\mathbf{d}_i\|_2^p}$ and an upper-bound for $\dot{\Delta}_{\max}$ are established. In the presented proof, we utilize the following lemma whose proof is available in the next section.

Lemma 9. *Suppose* $\{\mathbf{g}_i \in \mathbb{R}^N\}_{i=1}^n$ *are random i.i.d.* $\mathcal{N}(0, \frac{1}{N}\mathbf{I})$ *vectors, i.e.,* $\mathbb{E}[\mathbf{g}_i \mathbf{g}_i^T] = \frac{1}{N}\mathbf{I}$. *If* $N > 2$, *then*

$$\begin{aligned}
\sup_{\|\mathbf{u}\|=1} \sum_{i=1}^n (\mathbf{u}^T \mathbf{g}_i)^2 &\leq \frac{n}{N} + \eta \\
\inf_{\|\mathbf{u}\|=1} \sum_{i=1}^n (\mathbf{u}^T \mathbf{g}_i)^2 &\geq \frac{n}{N} - \eta
\end{aligned} \tag{38}$$

with probability at least $1 - 2\delta$ *where* $\eta = \max\left(\frac{4z}{3} \log \frac{2N}{\delta}, \sqrt{4 \frac{n(3+N)}{N^2} \log \frac{2N}{\delta}} \right)$, *and* $z = 1 + 2\sqrt{\frac{1}{N} \log \frac{2n}{\delta}} + \frac{2}{N} \log \frac{2n}{\delta}$.

Note that for any $t > 0$

$$\mathbb{P}\left[\Delta_{\min} < t\right] = \mathbb{P}\left[\Delta_1 < t \text{ or } \Delta_2 < t \text{ ... or } \Delta_m < t\right] \leq \sum_{i=1}^{m} \mathbb{P}\left[\Delta_i < t\right] , \tag{39}$$

where $\Delta_j = \inf_{\substack{\|\mathbf{u}\|=1 \\ \mathbf{u}\in\mathcal{S}_j}} \|\mathbf{u}^T\mathbf{D}_j\|_p^p$. According to (39), Lemma 9, and the definition of η_{δ_x},

$$\Delta_{\min} > \left(\frac{n}{r} - \eta_{\delta_r}\right) \tag{40}$$

with probability at least $1 - 2\delta$ because

$$\begin{aligned} &\mathbb{P}\left[\Delta_{\min} < \frac{n}{r} - \eta_{\delta_r}\right] \\ &\leq \sum_j \mathbb{P}\left[\inf_{\substack{\|\mathbf{u}\|=1 \\ \mathbf{u}\in\mathcal{S}_j}} \|\mathbf{u}^T\mathbf{D}_j\|_p^p < \frac{n}{r} - \eta_{\delta_r}\right] \\ &\leq \sum_{i=1}^{m} 2\delta/m = 2\delta . \end{aligned} \tag{41}$$

Similarly, we can conclude that

$$\dot{\Delta}_{\max} < \frac{r-s}{r}\left(\frac{n}{r-s} + \eta_{\delta_{r-s}}\right) \tag{42}$$

with probability at least $1 - 2\delta$.

Next we establish a lower-bound for $\frac{\|\beta_i\|_2^2}{\|\mathbf{d}_i\|_2^2}$. According to the presumed data model, the distribution of $r\|\beta_i\|_2^2$ is chi-square with $r - s$ degree of freedom and the distribution of $r\|\mathbf{d}_i\|_2^2$ is chi-square with r degree of freedom. First, we review the following lemma from [Laurent and Massart, 2000] which provides tail bounds for the chi-square distribution.

Lemma 10. *Let X_r be a chi-squared random variable with r-degrees of freedom. Then, for each $t > 0$,*

$$\begin{aligned} &\mathbb{P}\left[X_r \geq r + 2\sqrt{rt} + 2t\right] \leq e^{-t} , \\ &\mathbb{P}\left[X_r \leq r - 2\sqrt{rt}\right] \leq e^{-t} . \end{aligned} \tag{43}$$

The distribution of $\frac{\|\beta_i\|_2^2}{\|\mathbf{d}_i\|_2^2}$ is equivalent to the distribution of $\frac{1}{1+X_s/X_{r-s}}$ where X_s and X_{r-s} are chi-square random variables with s and $r-s$ degree of freedom, respectively. Therefore, in order to establish a lower-bound for $\|\beta_i\|_2^2/\|\mathbf{d}_i\|_2^2$, we derive an upper-bond for X_s/X_{r-s}. For any $t_1 > 0$ and $t_2 > 0$ we have

$$\begin{aligned} &\mathbb{P}\left[\frac{X_s}{X_{r-s}} \leq \frac{t_1}{t_2}\right] \geq \mathbb{P}\left[X_s \leq t_1 \text{ and } X_{r-s} \geq t_2\right] \\ &= \mathbb{P}\left[X_s \leq t_1\right] \mathbb{P}\left[X_{r-s} \geq t_2\right] , \end{aligned} \tag{44}$$

because X_s and X_{r-s} are independent. Therefore, according to (44) and according to Lemma 10, if $r - s - 2\sqrt{(r-s)\log\frac{2M_2}{\delta}} > 0$, then

$$\mathbb{P}\left[\frac{X_s}{X_{r-s}} \leq \frac{s+2\sqrt{s\log\frac{2M_2}{\delta}}+2\log\frac{2M_2}{\delta}}{r-s-2\sqrt{(r-s)\log\frac{2M_2}{\delta}}}\right] \geq 1 - \frac{\delta}{M_2} . \tag{45}$$

Using (45), we can conclude that

$$\mathbb{P}\left[\min_i \frac{\|\beta_i\|_2^2}{\|\mathbf{d}_i\|_2^2} \geq \frac{r-s-2\sqrt{(r-s)t_\delta}}{r+2\sqrt{s\log\frac{2M_2}{\delta}}+2t_\delta-2\sqrt{(r-s)t_\delta}}\right] \geq 1-\delta , \tag{46}$$

where $t_\delta = \log\frac{2M_2}{\delta}$. Therefore, according to (40), (42), (46), and (5), if the sufficient condition of Theorem 3 holds, then Requirement 1 is satisfied with probability at least $1 - 5\delta$.

Proof of Theorem 4

In order to guarantee that Requirement 1 is satisfied with high probability, we only need to guarantee that the sufficient condition of Theorem 3 holds. Accordingly, an upper-bound for ϕ is derived and it is used to write the new sufficient condition. First we review the following lemma whose proof is provided in the next section.

Lemma 11. *Suppose that orthonormal matrices $\mathbf{U}_1$ and $\mathbf{U}_2$ span two independent randomly generated r-dimensional subspaces. Then,*

$$\mathbb{P}\left[\|\mathbf{U}_1^T\mathbf{U}_2\| > \sqrt{\frac{c_\delta r^2}{M_1}}\right] \leq \delta ,$$

where $\sqrt{c_\delta} = 3\max\left(1, \sqrt{\frac{8M_1\pi}{(M_1-1)r}}, \sqrt{\frac{8M_1\log r/\delta}{(M_1-1)r}}\right)$.

Let us redefine $c_\delta = 3\max\left(1, \sqrt{\frac{8M_1\pi}{(M_1-1)(r-s)}}, \sqrt{\frac{16M_1\log\frac{mr}{\delta}}{(M_1-1)(r-s)}}\right)$. Therefore, according to Lemma 11,

$$\mathbb{P}\left[\phi > \sqrt{\frac{c_\delta(r-s)^2}{M_1}}\right] \leq \delta , \tag{47}$$

because

$$\begin{aligned} &\mathbb{P}\left[\phi > \sqrt{\frac{c_\delta(r-s)^2}{M_1}}\right] \\ &\leq \sum_i \sum_{j\neq i} \mathbb{P}\left[\|\mathbf{U}_i\mathbf{U}_j\| > \sqrt{\frac{c_\delta(r-s)^2}{M_1}}\right] \\ &\leq \sum_i \sum_j \delta/m^2 . \end{aligned} \tag{48}$$

According to (48) and according to Theorem 3, if the sufficient condition of Theorem 4 holds, then Requirement 1 is satisfied with probability at least $1-6\delta$.

Proof of Theorem 8

In order to prove that Requirement 1 is satisfied, we only need to guarantee that (10) holds. Therefore, we establish upper-bounds for $\max_i \left\{\frac{\|\alpha_i\|_2^p}{\|\mathbf{d}_i\|_2^p}\right\}$, $\max_i \left\{\frac{\|\beta_i\|_2^p}{\|\mathbf{d}_i\|_2^p}\right\}$, $\dot{\Delta}_{\max}$, and $\bar{\Delta}_{\max}$ and we establish a lower-bound for $\Delta_{\min}$.

As it was shown in the proof of Theorem 3, the distribution of $\|\beta_i\|_2^2/\|\mathbf{d}_i\|_2^2$ is equivalent to the distribution of $\frac{X_{r-s}}{X_{r-s}+X_s}$ where X_{r-s} and X_s are chi-squared random variables with $r-s$ and r degrees of freedom, respectively. Similarly, the distribution of $\|\alpha_i\|_2^2/\|\mathbf{d}_i\|_2^2$ is equivalent to the distribution of $\frac{X_s}{X_{r-s}+X_s}$. Therefore, we can use Lemma 10 to establish an upper-bound for $\max_i \left\{\|\alpha_i\|_2^2/\|\mathbf{d}_i\|_2^2\right\}$ and an upper-bound for $\max_i \left\{\|\beta_i\|_2^2/\|\mathbf{d}_i\|_2^2\right\}$. Note that $\frac{X_{r-s}}{X_{r-s}+X_s} = \frac{1}{1+X_s/X_{r-s}}$ and $\frac{X_s}{X_{r-s}+X_s} = \frac{1}{1+X_{r-s}/X_s}$ which means that we only need to bound $\frac{X_s}{X_{r-s}}$. According to Lemma 10,

$$\mathbb{P}\left[\frac{X_s}{X_{r-s}} \leq \frac{s-2\sqrt{s\log\frac{2M_2}{\delta}}}{r-s+2\sqrt{(r-s)\log\frac{2M_2}{\delta}}+2\log\frac{2M_2}{\delta}}\right] \leq \frac{\delta}{M_2}\,,$$

$$\mathbb{P}\left[\frac{X_{r-s}}{X_s} \leq \frac{(r-s)-2\sqrt{(r-s)\log\frac{2M_2}{\delta}}}{s+2\sqrt{s\log\frac{2M_2}{\delta}}+2\log\frac{2M_2}{\delta}}\right] \leq \frac{\delta}{M_2}\,.$$

Therefore,

$$\mathbb{P}\left[\max_i \left\{\|\beta_i\|_2^2/\|\mathbf{d}_i\|_2^2\right\} \geq \frac{r-s+2\sqrt{(r-s)t_\delta}+2t_\delta}{r+2\sqrt{(r-s)t_\delta}+2t_\delta-2\sqrt{s\,t_\delta}}\right] \leq \delta\,, \tag{49}$$

and

$$\mathbb{P}\left[\max_i \left\{\|\alpha_i\|_2^2/\|\mathbf{d}_i\|_2^2\right\} \geq \frac{s+2\sqrt{s\,t_\delta}+2t_\delta}{r+2\sqrt{s\,t_\delta}+2t_\delta-2\sqrt{(r-s)t_\delta}}\right] \leq \delta\,, \tag{50}$$

where $t_\delta = \log\frac{2M_2}{\delta}$

According to Lemma 9 and the definition of η_{δ_x},

$$\bar{\Delta}_{\max} \leq \frac{s}{r}\left(\frac{n}{s}+\eta_{\delta s}\right) \tag{51}$$

with probability at least $1-2\delta$ because

$$\mathbb{P}\left[\frac{r}{s}\bar{\Delta}_{\max} > \frac{n}{s}+\eta_{\delta s}\right] \leq \sum_j \mathbb{P}\left[\frac{r}{s}\sup_{\|\mathbf{u}\|=1}\|\mathbf{u}^T\bar{\mathbf{D}}_j\|_p^p > \frac{n}{s}+\eta_{\delta s}\right] \leq \sum_{j=1}^m \frac{2\delta}{m}\,. \tag{52}$$

Similarly,

$$\mathbb{P}\left[\dot{\Delta}_{\max} > \frac{r-s}{r}\left(\frac{n}{r-s}+\eta_{\delta_{r-s}}\right)\right] < 2\delta\,. \tag{53}$$

Therefore, according to (40), (48), (49), (50), (51), and (53), if the sufficient condition of Theorem 8 holds, Requirement 1 is satisfied with probability at $1-9\delta$.

Proof of Theorem 5

We use the same procedure employed in the proof of Theorem 2, i.e., a sufficient condition is derived which guarantees that

$$\|\mathbf{D}_{k_i}^T\mathbf{c}_i^*\|_p^p > \frac{\kappa}{\kappa+m-1}\|\mathbf{D}^T\mathbf{c}_i^*\|_p^p \tag{54}$$

holds for all $1 \leq i \leq M_2$. The lower-bound on the LHS of (54) was derived in (18).

We only need to establish an upper-bound for the RHS of (54). Corresponding to each data point $\mathbf{d}_i$, define $\vec{\mathbf{d}}_i$ as

$$\vec{\mathbf{d}}_i = \frac{\vec{\mathbf{U}}_{k_i}\vec{\mathbf{U}}_{k_i}^T\mathbf{d}_i}{\|\vec{\mathbf{U}}_{k_i}^T\dot{\mathbf{U}}_{k_i}\beta_i\|_2^2}\,. \tag{55}$$

Note that $\vec{\mathbf{d}}_i^T\mathbf{d}_i = 1$ because $\dot{\mathbf{U}}_{k_i}$ and $\vec{\mathbf{U}}_{k_i}$ are orthogonal to $\mathcal{S}$ and

$$\vec{\mathbf{d}}_i^T\mathbf{d}_i = \frac{\mathbf{d}_i^T\vec{\mathbf{U}}_{k_i}\vec{\mathbf{U}}_{k_i}^T\mathbf{d}_i}{\|\vec{\mathbf{U}}_{k_i}^T\dot{\mathbf{U}}_{k_i}\beta_i\|_2^2} = \frac{\beta_i^T\dot{\mathbf{U}}_{k_i}^T\vec{\mathbf{U}}_{k_i}\vec{\mathbf{U}}_{k_i}^T\dot{\mathbf{U}}_{k_i}\beta_i}{\|\vec{\mathbf{U}}_{k_i}^T\dot{\mathbf{U}}_{k_i}\beta_i\|_2^2} = 1\,. \tag{56}$$

Since $\vec{\mathbf{d}}_i^T\mathbf{d}_i = 1$ and $\mathbf{c}_i^*$ is the optimal direction, we conclude that

$$\|\mathbf{D}^T\mathbf{c}_i^*\|_p^p \leq \|\mathbf{D}_{k_i}^T\vec{\mathbf{d}}_i\|_p^p + \|\mathbf{D}_{-k_i}^T\vec{\mathbf{d}}_i\|_p^p\,. \tag{57}$$

In addition, $\vec{\mathbf{d}}_i$ is orthogonal to $\oplus_{j\neq k_i}\mathcal{S}_i$ which means that $\|\mathbf{D}_{-k_i}^T\vec{\mathbf{d}}_i\|_p^p = 0$. Moreover, since $\vec{\mathbf{d}}_i$ is orthogonal $\mathcal{S}$,

$$\begin{aligned}\|\mathbf{D}_{k_i}^T\vec{\mathbf{d}}_i\|_p^p &= \|\dot{\mathbf{D}}_{k_i}^T\dot{\mathbf{U}}_{k_i}\dot{\mathbf{U}}_{k_i}^T\vec{\mathbf{d}}_i\|_p^p \\ &\leq \|\vec{\mathbf{d}}_i\|_2^p\sup_{\|\mathbf{u}\|=1}\|\mathbf{u}^T\dot{\mathbf{D}}_{k_i}\|_p^p\,.\end{aligned} \tag{58}$$

In addition,

$$\|\vec{\mathbf{d}}_i\|_2 = \frac{1}{\|\vec{\mathbf{U}}_{k_i}^T\dot{\mathbf{U}}_{k_i}\beta_i\|_2}\,. \tag{59}$$

Therefore, if

$$\frac{1}{\|\mathbf{d}_i\|_2^p} \inf_{\substack{\|\mathbf{u}\|=1 \\ \mathbf{u}\in\mathcal{S}_{k_i}}} \|\mathbf{u}^T\mathbf{D}_{k_i}\|_p^p \, \|\vec{\mathbf{U}}_{k_i}^T\dot{\mathbf{U}}_{k_i}\beta_i\|_2^p \geq \frac{\kappa}{\kappa+m-1} \sup_{\|\mathbf{u}\|=1} \|\mathbf{u}^T\dot{\mathbf{D}}_{k_i}\|_p^p \,, \tag{60}$$

then (54) holds. In addition,

$$\|\vec{\mathbf{U}}_{k_i}^T\dot{\mathbf{U}}_{k_i}\beta_i\|_2 \geq \|\beta_i\|_2 \|\vec{\mathbf{U}}_{k_i}^T\dot{\mathbf{U}}_{k_i}\|_m \,, \tag{61}$$

where $\|\vec{\mathbf{U}}_{k_i}^T\dot{\mathbf{U}}_{k_i}\|_m$ denotes the minimum singular value of matrix $\vec{\mathbf{U}}_{k_i}^T\dot{\mathbf{U}}_{k_i}$. Accordingly, if (7) holds, then Requirement 1 is satisfied.

Proof of Theorem 6
According to the proof of Theorem 5, it is enough to guarantee that (60) holds for all the data points. First we establish a lower-bound for

$$\|\vec{\mathbf{U}}_{k_i}\vec{\mathbf{U}}_{k_i}^T\dot{\mathbf{U}}_{k_i}\beta_i\|_2/\|\mathbf{d}_i\|_2^2 \,.$$

Note that $\vec{\mathbf{U}}$ is the span of $\{(\mathbf{I}-\mathbf{P}_{k_i}\mathbf{P}_{k_i}^T)\mathbf{x} \,|\, \mathbf{x}\in\dot{\mathcal{S}}_{k_i}\}$ where $\mathbf{P}_{k_i}$ is an orthonormal basis for $\oplus_{j\neq k_i}\mathcal{S}_j$. Thus, the column space of $\vec{\mathbf{U}}_{k_i}$ is a subset of the column space of $(\mathbf{I}-\mathbf{P}_{k_i}\mathbf{P}_{k_i}^T)$. Therefore, $\vec{\mathbf{U}}_{k_i}\vec{\mathbf{U}}_{k_i}^T\dot{\mathbf{U}}_{k_i}\beta_i = (\mathbf{I}-\mathbf{P}_{k_i}\mathbf{P}_{k_i}^T)\dot{\mathbf{U}}_{k_i}\beta_i$. Since the subspaces $\{\dot{\mathcal{S}}_i\}_{i=1}^m$ and $\mathcal{S}$ are generated independently and uniformly at random and $M_1 > s+(r-s)m$, then $\{\dot{\mathcal{S}}_i\}_{i=1}^m$ and $\mathcal{S}$ are independent subspaces with an overwhelming probability. In other word, the dimension of $(\oplus_{i=1}^m\dot{\mathcal{S}}_{i=1})\oplus\mathcal{S}$ is equal to $s+(r-s)m$ with an overwhelming probability [Vershynin, 2010]. Let us assume that this is true using which we can conclude that the column-space of $(\mathbf{I}-\mathbf{P}_{k_i}\mathbf{P}_{k_i}^T)$ is a random $s+(r-s)(m-1)$ subspace which is generated independently from $\dot{\mathcal{S}}_{k_i}$. Define $\mathbf{H}_{k_i}$ as an orthonormal basis for $(\mathbf{I}-\mathbf{P}_{k_i}\mathbf{P}_{k_i}^T)$. Since the column space of $\mathbf{H}_{k_i}$ and $\dot{\mathcal{S}}_{k_i}$ are generated independently, the distribution of

$$\frac{\|\mathbf{H}_{k_i}^T\dot{\mathbf{U}}_{k_i}\beta_i\|_2^2}{\|\beta_i\|_2^2+\|\alpha_i\|_2^2}$$

is equivalent to the distribution of

$$\frac{\|\mathbf{H}_{k_i}^T\mathbf{a}\|_2^2 \, \|\beta_i\|_2^2}{\|\beta_i\|_2^2+\|\alpha_i\|_2^2}$$

where $\mathbf{a}\in\mathbb{R}^{M_1}$ is a random vector on $\mathbb{S}^{M_1-1}$ which is independent from $\|\beta_i\|_2^2$. In addition, for any $t_1>0$ and $t_2>0$ we can write

$$\mathbb{P}\left[\frac{\|\mathbf{H}_{k_i}^T\mathbf{a}\|_2^2 \, \|\beta_i\|_2^2}{\|\beta_i\|_2^2+\|\alpha_i\|_2^2} \geq t_1 \, t_2\right] \geq \mathbb{P}\left[\frac{\|\beta_i\|_2^2}{\|\beta_i\|_2^2+\|\alpha_i\|_2^2} \geq t_1 \quad \text{and} \quad \|\mathbf{H}_{k_i}^T\mathbf{a}\|_2^2 \geq t_2\right] \,. \tag{62}$$

Similar to (49),

$$\mathbb{P}\left[\frac{\|\beta_i\|_2^2}{\|\beta_i\|_2^2+\|\alpha_i\|_2^2} \geq t_1\right] \geq 1-\delta/M_2$$

where $t_1 = \frac{r-s-2\sqrt{(r-s)\log\frac{2M_2}{\delta}}}{r+2\sqrt{s\log\frac{2M_2}{\delta}}+2\log\frac{2M_2}{\delta}-2\sqrt{(r-s)\log\frac{2M_2}{\delta}}}$.

Since the column-space of $\mathbf{H}_{k_i}$ is a random subspace and $\mathbf{a}$ is a random vector on $\mathbb{S}^{M_1-1}$, the distribution of $\|\mathbf{H}_{k_i}^T\mathbf{a}\|_2$ is equivalent to the distribution of $\|\mathbf{H}_{k_i}^T\mathbf{e}_1\|_2$ where $\mathbf{e}_1$ is the first column of the identity matrix $\mathbf{I}\in\mathbb{R}^{M_1\times M_1}$. The distribution of $\|\mathbf{H}_{k_i}^T\mathbf{e}_1\|_2^2$ is equivalent to the distribution of $\frac{X_{M_1-\vartheta}}{X_{M_1}}$ where X_{M_1} is a chi-squared random variables with M_1 degree of freedom and $\vartheta = M_1-\big(s+(r-s)(m-1)\big)$ [Candès and Recht, 2009]. Therefore, we can bound $\|\mathbf{H}_{k_i}^T\mathbf{e}_1\|_2^2$ similarly

$$\mathbb{P}\left[\|\mathbf{H}_{k_i}^T\mathbf{e}_1\|_2^2 \geq t_2\right] \geq 1-\frac{\delta}{M_2} \,, \tag{63}$$

where $t_2 = \frac{\vartheta-2\sqrt{\vartheta\, t_\delta}}{M_1+2\sqrt{(M_1-\vartheta)t_\delta}+2t_\delta-2\sqrt{\vartheta t_\delta}}$ and $t_\delta = \log\frac{2M_2}{\delta}$. Therefore, according to (40), (42), (60), (62), and (63), if the sufficient condition of Theorem 6 holds, then Requirement 1 is satisfied with probability at least $1-6\delta-\epsilon$ where ϵ is the probability that the rank of $\mathbf{D}$ is smaller than $s+(r-s)m$.

C PROOFS OF THE INTERMEDIATE RESULTS

Proof of Lemma 9
First, we add and subtract the mean of each random component as follows

$$\sup_{\|\mathbf{u}\|=1} \sum_{i=1}^n (\mathbf{u}^T\mathbf{g}_i)^2 \leq \sup_{\|\mathbf{u}\|=1} \sum_{i=1}^n \left[(\mathbf{u}^T\mathbf{g}_i)^2 - \mathbb{E}(\mathbf{u}^T\mathbf{g}_i)^2\right] + \sup_{\|\mathbf{u}\|=1} \sum_{i=1}^n \mathbb{E}(\mathbf{u}^T\mathbf{g}_i)^2. \tag{64}$$

Note that

$$\mathbb{E}(\mathbf{u}^T\mathbf{g}_i)^2 = \frac{1}{N}\mathbb{E}\, r_i(\mathbf{u}^T\mathbf{s}_i)^2 = \frac{1}{N}\mathbb{E}\, r_i(\mathbf{u}^T\mathbf{s}_i)^2 = \frac{1}{N}\mathbb{E}\, r_i\mathbb{E}(\mathbf{u}^T\mathbf{s}_i)^2 \tag{65}$$

where $\mathbf{s}_i\in\mathbb{R}^N$ is a random vector on $\mathbb{S}^{N-1}$ and r_i is a chi-squared random variable with N degree of freedom. Since $\mathbf{s}_i$ is sampled uniformly at random from $\mathbb{S}^{N-1}$,

$$\mathbb{E}(\mathbf{u}^T\mathbf{s}_i)^2 = \mathbb{E}(\mathbf{e}_1^T\mathbf{s}_i)^2 \,, \tag{66}$$

where $\mathbf{e}_i$ is the first column of Identity matrix $\mathbf{I}$. Therefore,

$$\mathbb{E}(\mathbf{u}^T\mathbf{g}_i)^2 = \frac{1}{N}\mathbb{E}\, r_i \;\; \mathbb{E}(\mathbf{e}_1^T\mathbf{s}_i)^2 = \mathbb{E}(\mathbf{e}_1^T\mathbf{g}_i)^2 = \frac{1}{N} \,, \tag{67}$$

and

$$\sup_{\|\mathbf{u}\|=1} \sum_{i=1}^{n} \mathbb{E}(\mathbf{u}^T\mathbf{g}_i)^2 = \frac{n}{N} \,. \tag{68}$$

The first component of the RHS of (82) can be rewritten as

$$\begin{aligned} &\sup_{\|\mathbf{u}\|=1} \sum_{i=1}^{n} \left[(\mathbf{u}^T\mathbf{g}_i)^2 - \mathbb{E}(\mathbf{u}^T\mathbf{g}_i)^2\right] \\ &= \sup_{\|\mathbf{u}\|=1} \mathbf{u}^T \left(\sum_{i=1}^{n} \mathbf{g}_i\mathbf{g}_i^T - \mathbb{E}\left\{\mathbf{g}_i\mathbf{g}_i^T\right\}\right)\mathbf{u} \\ &= \sup_{\|\mathbf{u}\|=1} \mathbf{u}^T \left(\sum_{i=1}^{n} \mathbf{g}_i\mathbf{g}_i^T - \frac{1}{N}\mathbf{I}\right)\mathbf{u}. \end{aligned} \tag{69}$$

The matrices $\{\mathbf{g}_i\mathbf{g}_i^T - \frac{1}{N}\mathbf{I}\}_{i=1}^n$ are zero mean random matrices. Thus, we use the non-commutative Bernstein inequality to bound the spectral norm of the matrix $\mathbf{M}$ defined as

$$\mathbf{M} = \sum_{i=1}^{n} \left(\mathbf{g}_i\mathbf{g}_i^T - \frac{1}{N}\mathbf{I}\right). \tag{70}$$

Lemma 12. *[Junge et al., 2013] Let $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_L$ be independent zero-mean random matrices of dimension $d_1 \times d_2$. Suppose $\rho_k^2 = \max\{\|\mathbb{E}[\mathbf{X}_k\mathbf{X}_k^T]\|, \|\mathbb{E}[\mathbf{X}_k^T\mathbf{X}_k]\|\}$ and $\|\mathbf{X}_k\| \leq M$ almost surely for all k. Then for any $\tau > 0$*

$$\begin{aligned} &\mathbb{P}\left[\left\|\sum_{k=1}^{L} \mathbf{X}_k\right\| > \tau\right] \leq \\ &\qquad (d_1+d_2)\exp\left(\frac{-\tau^2/2}{\sum_{k=1}^{L}\rho_k^2 + M\tau/3}\right). \end{aligned} \tag{71}$$

To find the parameter M defined in Lemma 12, we compute

$$\|\mathbf{g}_i\mathbf{g}_i^T - \frac{1}{N}\mathbf{I}\| \leq \max(\|\mathbf{g}_i\mathbf{g}_i^T\|, \frac{1}{N}) \tag{72}$$

where we used the fact that $\|\mathbf{H}_1 - \mathbf{H}_2\| \leq \max(\|\mathbf{H}_1\|, \|\mathbf{H}_2\|)$, if $\mathbf{H}_1$ and $\mathbf{H}_2$ are positive definite matrices. Note that,

$$\|\mathbf{g}_i\mathbf{g}_i^T\| = \sup_{\|\mathbf{u}\|=1} \|\mathbf{g}_i\mathbf{g}_i^T\mathbf{u}\|_2 \leq \|\mathbf{g}_i\|_2^2 \,. \tag{73}$$

The distribution of $N\|\mathbf{g}_i\|_2^2$ is equivalent to the distribution of a chi-square distribution with N degrees of freedom. Therefore, according to Lemma 10,

$$\mathbb{P}\left[\|\mathbf{g}_i\mathbf{g}_i^T\| > 1 + 2\sqrt{\frac{1}{N}\log\frac{2n}{\delta}} + \frac{2}{N}\log\frac{2n}{\delta}\right] < \frac{\delta}{n} \,. \tag{74}$$

According to (74) we conclude that M is less than or equal to $1 + 2\sqrt{\frac{1}{N}\log\frac{2n}{\delta}} + \frac{2}{N}\log\frac{2n}{\delta}$ with probability at least $1-\delta$.

For the parameter ρ we have

$$\begin{aligned} &\left\|\mathbb{E}\left[\left(\mathbf{g}_i\mathbf{g}_i^T - \frac{1}{N}\mathbf{I}\right)\left(\mathbf{g}_i\mathbf{g}_i^T - \frac{1}{N}\mathbf{I}\right)\right]\right\| = \\ &\left\|\mathbb{E}\left[\mathbf{g}_i\mathbf{g}_i^T\mathbf{g}_i\mathbf{g}_i^T - \frac{2}{N}\mathbf{g}_i\mathbf{g}_i^T + \frac{1}{N^2}\mathbf{I}\right]\right\| = \\ &\left\|\mathbb{E}\left[\mathbf{g}_i\mathbf{g}_i^T\mathbf{g}_i\mathbf{g}_i^T\right] - \frac{2}{N}\mathbf{g}_i\mathbf{g}_i^T + \frac{1}{N^2}\mathbf{I}\right\| = \\ &\left\|\mathbb{E}\left[\mathbf{g}_i\mathbf{g}_i^T\mathbf{g}_i\mathbf{g}_i^T\right] - \frac{1}{N^2}\mathbf{I}\right\| \leq \max\left(\frac{1}{N^2}, \left\|\mathbb{E}\left[\mathbf{g}_i\mathbf{g}_i^T\mathbf{g}_i\mathbf{g}_i^T\right]\right\|\right) \end{aligned} \tag{75}$$

Therefore, we only need to compute $\left\|\mathbb{E}\left[\mathbf{g}_i\mathbf{g}_i^T\mathbf{g}_i\mathbf{g}_i^T\right]\right\|$. Define $\mathbf{G} = N^2\mathbf{g}_i\mathbf{g}_i^T\mathbf{g}_i\mathbf{g}_i^T$. The distribution of diagonal values of $\mathbf{G}$ is equivalent to the distribution of X_1^2 where X_1 a chi squared random variable with 1 degree of freedom. Thus, $\mathbb{E}\left[\mathbf{G}(i,i)\right] = 3$. The distribution of off-diagonal elements of $\mathbf{G}$ is equivalent to X_1Y_1 where X_1 and Y_1 are independent chi squared random variables with 1 degree of freedom. Thus, $\mathbb{E}\left[\mathbf{G}(i,i)\right] = 1$ when $i \neq j$. Accordingly,

$$N^2\mathbb{E}\left[\mathbf{g}_i\mathbf{g}_i^T\mathbf{g}_i\mathbf{g}_i^T\right] = 2\mathbf{I} + \mathbf{1}\mathbf{1}^T \,, \tag{76}$$

where $\mathbf{1} \in \mathbb{R}^N$ is a vector whose all elements are equal to 1. Therefore, according to $\|\mathbb{E}\left[\mathbf{g}_i\mathbf{g}_i^T\mathbf{g}_i\mathbf{g}_i^T\right]\| = \frac{2+N}{N^2}$ and (75),

$$\left\|\mathbb{E}\left[\left(\mathbf{g}_i\mathbf{g}_i^T - \frac{1}{N}\mathbf{I}\right)\left(\mathbf{g}_i\mathbf{g}_i^T - \frac{1}{N}\mathbf{I}\right)\right]\right\| \leq \frac{3+N}{N^2} \,.$$

According to Lemma 12 and assuming that $\|\mathbf{g}_i\mathbf{g}_i^T - \frac{1}{N}\mathbf{I}\| \leq z_\delta$,

$$\mathbb{P}\left[\|\mathbf{M}\| > \tau\right] \leq 2N\exp\left(\frac{-\tau^2/2}{n\frac{3+N}{N^2} + z_\delta\tau/3}\right) \,. \tag{77}$$

In addition,

$$\exp\left(\frac{-\tau^2/2}{n\frac{3+N}{N^2} + z_\delta\tau/3}\right) \leq \left(\frac{-\tau^2/2}{2\max(n\frac{3+N}{N^2}, z_\delta\tau/3)}\right) \tag{78}$$

Thus,

$$\mathbb{P}\left[\|\mathbf{M}\| > \eta_\delta\right] \leq 2\delta \tag{79}$$

where

$$\begin{aligned} \eta_\delta &= \max\left(\frac{4z_\delta}{3}\log\frac{2N}{\delta}, \sqrt{4\frac{n(3+N)}{N^2}\log\frac{2N}{\delta}}\right) \\ z_\delta &= 1 + 2\sqrt{\frac{1}{N}\log\frac{2n}{\delta}} + \frac{2}{N}\log\frac{2n}{\delta} \,. \end{aligned} \tag{80}$$

According to (68) and (79),

$$\sup_{\|\mathbf{u}\|=1} \sum_{i=1}^{n} (\mathbf{u}^T\mathbf{g}_i)^2 < \frac{n}{N} + \eta_\delta \tag{81}$$

with probability at least $1-2\delta$.

We use similar techniques to prove the second inequality in (38). First, we add and subtract the mean of each random component as follows

$$\inf_{\|\mathbf{u}\|=1} \sum_{i=1}^{n} (\mathbf{u}^T \mathbf{g}_i)^2 \geq \inf_{\|\mathbf{u}\|=1} \sum_{i=1}^{n} \mathbb{E}(\mathbf{u}^T \mathbf{g}_i)^2 + \inf_{\|\mathbf{u}\|=1} \sum_{i=1}^{n} \left[(\mathbf{u}^T \mathbf{g}_i)^2 - \mathbb{E}(\mathbf{u}^T \mathbf{g}_i)^2 \right] . \tag{82}$$

In addition, $\inf_{\|\mathbf{u}\|=1} \sum_{i=1}^{n} \left[(\mathbf{u}^T \mathbf{g}_i)^2 - \mathbb{E}(\mathbf{u}^T \mathbf{g}_i)^2 \right] = -1 \times \sup_{\|\mathbf{u}\|=1} \sum_{i=1}^{n} \left[\mathbb{E}(\mathbf{u}^T \mathbf{g}_i)^2 - (\mathbf{u}^T \mathbf{g}_i)^2 \right]$ which can be rewritten as

$$\sup_{\|\mathbf{u}\|=1} \sum_{i=1}^{n} \left[\mathbb{E}(\mathbf{u}^T \mathbf{g}_i)^2 - (\mathbf{u}^T \mathbf{g}_i)^2 \right] = \left\| \sum_{i=1}^{n} \left(\mathbf{g}_i \mathbf{g}_i^T - \frac{1}{N} \mathbf{I} \right) \right\| . \tag{83}$$

Using similar techniques which was used to bound $\|\mathbf{M}\|$, we can conclude that

$$\sup_{\|\mathbf{u}\|=1} \sum_{i=1}^{n} \left[\mathbb{E}(\mathbf{u}^T \mathbf{g}_i)^2 - (\mathbf{u}^T \mathbf{g}_i)^2 \right] \leq \eta_\delta \tag{84}$$

with probability at least $1 - 2\delta$. Thus, according to (82) and (84),

$$\inf_{\|\mathbf{u}\|=1} \sum_{i=1}^{n} (\mathbf{u}^T \mathbf{g}_i)^2 > \frac{n}{N} - \eta_\delta \tag{85}$$

with probability at least $1 - 2\delta$.

Proof of Lemma 11

First we review the following lemma.

Lemma 13. *[Park et al., 2014, Ledoux, 2005] Let the columns of $\mathbf{U} \in \mathbb{R}^{N \times r}$ be an orthonormal basis for an r-dimensional random subspace drawn uniformly at random in an ambient N-dimensional space. For a unit ℓ_2-norm vector $\mathbf{c} \in \mathbb{R}^{N \times 1}$*

$$\mathbb{P}\left[\|\mathbf{c}^T \mathbf{U}\|_2 > \sqrt{\frac{c_\delta r}{N}} \right] \leq \delta ,$$

where $\sqrt{c_\delta} = 3 \max\left(1, \sqrt{\frac{8N\pi}{(N-1)r}}, \sqrt{\frac{8N \log 1/\delta}{(N-1)r}}\right)$.

The spectral norm of matrix $\mathbf{U}_1^T \mathbf{U}_2$ is equal to its largest singular value. Define $\mathbf{s} \in \mathbb{R}^r$ as the vector of singular values of $\mathbf{U}_1^T \mathbf{U}_2$ where $\max_i \mathbf{s}(i) = \mathbf{s}(1)$. Note that

$$\|\mathbf{U}_1^T \mathbf{U}_2\| = \mathbf{s}(1) \leq \|\mathbf{s}\|_2 = \|\mathbf{U}_1^T \mathbf{U}_2\|_F \tag{86}$$

where $\|\mathbf{U}_1^T \mathbf{U}_2\|_F$ is the Frobenius norm of $\mathbf{U}_1^T \mathbf{U}_2$. The Frobenius norm can be expanded as

$$\|\mathbf{U}_1^T \mathbf{U}_2\|_F^2 = \sum_{i=1}^{r} \|{\mathbf{u}_1}_i^T \mathbf{U}_2\|_2^2 , \tag{87}$$

where ${\mathbf{u}_1}_i$ is the ith column of $\mathbf{U}_1$. According to Lemma 13,

$$\mathbb{P}\left[\|{\mathbf{u}_1}_i^T \mathbf{U}_2\|_2^2 > \frac{r\, c_{\frac{\delta}{r}}}{N} \right] \leq \frac{\delta}{r} . \tag{88}$$

Therefore,

$$\|\mathbf{U}_1^T \mathbf{U}_2\|_F^2 \leq \frac{c_{\frac{\delta}{r}}\, r^2}{N} \tag{89}$$

with probability at least $1 - \delta$.