Geometric Conditions for Subspace-Sparse Recovery

Chong You CYOU@CIS.JHU.EDU
René Vidal RVIDAL@CIS.JHU.EDU
Center for Imaging Science, Johns Hopkins University, Baltimore, MD, 21218, USA

Abstract

Given a dictionary Π and a signal $\xi = \Pi\mathbf{x}$ generated by a few *linearly independent* columns of Π, classical sparse recovery theory deals with the problem of uniquely recovering the sparse representation $\mathbf{x}$ of ξ. In this work, we consider the more general case where ξ lies in a low-dimensional subspace spanned by a few columns of Π, which are possibly *linearly dependent*. In this case, $\mathbf{x}$ may not unique, and the goal is to recover any subset of the columns of Π that spans the subspace containing ξ. We call such a representation $\mathbf{x}$ *subspace-sparse*. We study conditions under which existing pursuit methods recover a subspace-sparse representation. Such conditions reveal important geometric insights and have implications for the theory of classical sparse recovery as well as subspace clustering.

1. Introduction

Classical sparse recovery theory studies the problem of representing a signal in terms of an over-complete dictionary by using as few dictionary atoms as possible (Baraniuk, 2007; Candès & Wakin, 2008). Since this problem is generally intractable, it is usually approached using approximate algorithms such as Orthogonal Matching Pursuit (OMP) (Pati et al., 1993) and Basis Pursuit (BP) (Chen et al., 1998). The study of these algorithms has been the topic of various works in the past two decades, see e.g., (Tropp, 2004; Candès & Tao, 2005; Donoho et al., 2006; Davenport & Wakin, 2010; Cai & Zhang, 2014). Such work has shown that if the dictionary is *incoherent* or satisfies the so-called *restricted isometry property (RIP)*, the true sparsest solution can be found by these algorithms.

More recently, motivated by applications in subspace classification (Wright et al., 2009) and subspace clustering (Vidal, 2011), sparse recovery theory has been extended to the analysis of *subspace-structured dictionaries* where the dictionary atoms lie in a union of several low-dimensional subspaces of the ambient space. In this case, the data permits a *subspace-sparse* representation, where any atom from a given subspace can be written as a sparse linear combination of other atoms from the same subspace, a concept that is fundamental to several subspace clustering techniques (see below). While in principle a subspace-sparse representation could possibly be recovered via BP or OMP, classical sparse recovery theory typically requires that the dictionary atoms satisfy certain incoherence or RIP properties. For the problem of finding subspace-sparse representations, the data points themselves function as the dictionary elements, so these properties are rarely satisfied. For example, the dictionary could be highly coherent since points from the same subspace could be arbitrarily close. This severely limits the applicability of classical sparse recovery theory to subspace-sparse models.

Following the initial work of (Elhamifar & Vidal, 2009), several recent works (Elhamifar & Vidal, 2010; 2013; Soltanolkotabi & Candès, 2013; Soltanolkotabi et al., 2014; Wang & Xu, 2013; Dyer et al., 2013) have studied this subspace-sparse recovery problem in the context of *subspace clustering*, where the task is to cluster a collection of points lying in a union of subspaces. In this case, the problem is solved by first finding a subspace-sparse representation of each point in terms of a dictionary composed of all other points and then applying spectral clustering to these subspace-sparse representations. Notice, however, that these analyses are specific for the correctness of subspace clustering. In this work we study the more general problem where the dictionary is composed of some "inlier" points lying in a subspace and some arbitrary "outlier" points in the ambient space, while the signal to be represented is an arbitrary point in the inlier subspace. The goal of this paper is to study this more general *subspace-sparse* recovery problem, and to derive conditions on the dictionary under which the OMP and BP algorithms are guaranteed to give subspace-sparse solutions. Furthermore, we obtain new theoretical conditions for classical sparse recovery as well as subspace clustering.

Proceedings of the 32^{nd} *International Conference on Machine Learning*, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

1.1. Problem formulation

Suppose we are given a dictionary Π that can be partitioned into a set of "inliers" and a set of "outliers". The matrix of inliers $\Phi = [\phi_1, \cdots, \phi_M]$ consists of M points that span a low-dimensional subspace of dimension $s := \text{rank}(\Phi) \leq M$. The matrix of outliers $\Psi = [\psi_1, \cdots, \psi_N]$ contains arbitrary points that do not lie in the inlier subspace. For instance, in subspace clustering the outliers lie in a union of several other subspaces. We assume that we do not know which columns of Π are inliers and which are outliers, i.e. the known matrix Π is equal to the matrix $[\Phi, \Psi]$ up to an unknown permutation of its columns. Given an arbitrary point ξ in the inlier subspace $\xi \in \mathcal{R}(\Phi)$, where $\mathcal{R}(\cdot)$ denotes the range, the goal is to find a representation of ξ that is *subspace-preserving*, i.e. a vector $\mathbf{x} \in \mathbb{R}^{M+N}$ such that 1) $\xi = \Pi\mathbf{x}$, and 2) $\mathbf{x}_j = 0$ if the j-th column of Π corresponds to an outlier. We can see that a subspace-preserving representation always exists since $\xi \in \mathcal{R}(\Phi)$, but it may not be unique if $s < M$. As such, our goal is to find any subspace-preserving representation, and we consider all such solutions equally good for our purpose.

Notice that one can always find a subspace-preserving representation $\xi = \Pi\mathbf{x}$ where $\mathbf{x}$ has at most s non-zero entries, and if $s \ll N + M$, then such a representation is *sparse*. This motivates us to search for a sparse representation, i.e.

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0 \text{ s.t. } \xi = \Pi\mathbf{x}, \tag{1}$$

where $\|\cdot\|_0$ counts the number of nonzero entries in $\mathbf{x}$. Since this problem is intractable, it is usually solved in an approximate manner using classical pursuit methods, such as OMP and BP. We call a representation $\mathbf{x}$ found by such sparsity pursuit algorithms *subspace-sparse* if it is subspace-preserving. The main goal of this paper is thus to study the conditions on Φ, Ψ under which these two algorithms give subspace-sparse solutions for any $\xi \in \mathcal{R}(\Phi)$.

1.2. Relation with Sparse Recovery

Classical sparse recovery is a particular case of subspace-sparse recovery. In sparse recovery one is given a dictionary Π and a vector $\xi := \Pi\mathbf{x}$ which is a linear combination of a few, say M, columns of Π, and the goal is to find the M-sparse vector $\mathbf{x}$. In order for this problem to be well posed, a necessary condition is that the M columns of Π corresponding to the M nonzero entries of $\mathbf{x}$ must be linearly independent so that the solution of $\xi = \Pi\mathbf{x}$ is unique. Therefore, the classical sparse recovery problem is a particular case of the subspace-sparse recovery problem where the inlier matrix Φ is composed of the M columns of Π that generate ξ, which must be linearly independent. Therefore, the conditions for guaranteeing subspace-sparse recovery are also applicable for guaranteeing sparse recovery.

1.3. Results and Contributions

This section briefly summarizes the major contributions of the paper.

Theorems 4 and 5 introduce, respectively, the principal recovery condition (PRC) and the dual recovery condition (DRC) for *subspace-sparse* recovery. Both of them are conditions on the dictionary Π under which both OMP and BP give a subspace-sparse solution for every $\xi \in \mathcal{R}(\Phi)$.

The PRC requires that

$$r(\mathcal{K}(\pm\Phi)) > \mu(\Psi, \mathcal{R}(\Phi)), \tag{2}$$

where the left hand side, $r(\mathcal{K}(\pm\Phi))$, is the radius of the largest ball inscribed in the convex hull of the symmetrized columns of the inliers, $\mathcal{K}(\pm\Phi)$. This *inradius* measures how well distributed the inlier points are in the inlier subspace $\mathcal{R}(\Phi)$, and should be relatively large if the points are equally distributed in all directions within the inlier subspace and not skewed in a certain direction. The right hand side, $\mu(\Psi, \mathcal{R}(\Phi))$, is the coherence between all outliers Ψ and all the points in $\mathcal{R}(\Phi)$, defined as the maximum coherence (cosine of acute angle) between any pair of points each taken from one set. The coherence is small when all pairs of points from the two sets are sufficiently separated, so intuitively, the PRC requires the inlier points to be sufficiently well spread-out and the outliers to be sufficiently away from the inlier subspace.

The PRC has the drawback that $\mathcal{R}(\Phi)$ on the right hand side contains infinitely many points, making the condition too strong. We show that a finite subset of the points in $\mathcal{R}(\Phi)$ is sufficient for this purpose, leading to the DRC:

$$r(\mathcal{K}(\pm\Phi)) > \mu(\Psi, \mathcal{D}(\Phi)), \tag{3}$$

where $\mathcal{D}(\Phi)$ is a finite subset of the points in the inlier subspace $\mathcal{R}(\Phi)$, which will be defined in Section 2.2. Hence, the DRC is implied by the PRC, thus it gives a stronger result. That is, the DRC does not require *all* points in the inlier subspace to be incoherent with the outliers, as done by the PRC. Instead, only a finite number of points, the columns of $\mathcal{D}(\Phi)$, are sufficient for all the points in $\mathcal{R}(\Phi)$.

As a corollary, we show that the PRC and DRC are also sufficient conditions for traditional sparse recovery when Φ has full column rank. Moreover, we compare the result with traditional theories of sparse recovery, and establish that PRC is implied by the incoherence condition

$$\mu(\Pi) < \frac{1}{2M - 1}, \tag{4}$$

where $\mu(\cdot)$ is the coherence of a matrix, defined as the maximum absolute inner product between any two columns of the matrix. Thus the PRC provides a stronger and geometrically more interpretable result for sparse recovery.

2. Background

This section presents some background material that will be needed for the main results of the paper. We first briefly introduce the OMP and BP methods, and then define various geometric properties that characterize the dictionary.

2.1. Algorithms

OMP and BP are two methods for approximately solving problem (1).

OMP is a greedy method that sequentially chooses one dictionary atom in a locally optimal manner. It keeps track of a residual η_k at step k, initialized as the input signal ξ, and a set Λ_k that contains the atoms already chosen, initialized as the empty set. At each step, Λ_k is updated to Λ_{k+1} by adding the dictionary atom that has the maximum absolute inner product with η_k. Then, η_k is updated to η_{k+1} by setting it to be the component of ξ that is orthogonal to the space spanned by the atoms indexed by Λ_{k+1}. The process is terminated when a precise representation of ξ is established, i.e. when $\eta_k = 0$ for some k.

BP is a convex relaxation approach. The idea is to use the ℓ_1 norm in lieu of the ℓ_0 norm in (1),

$$P(\Pi, \xi) := \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \ \text{ s.t. } \ \Pi\mathbf{x} = \xi, \tag{5}$$

and has the benefit that (5) is convex and can be solved efficiently. We will denote the objective value of (5) by $p(\Pi, \xi)$, i.e. $p(\Pi, \xi) = \|\mathbf{x}\|_1$ where $\mathbf{x} \in P(\Pi, \xi)$, and by convention, $p(\Pi, \xi) = +\infty$ if the problem is infeasible. The dual of the above optimization program is

$$D(\Pi, \xi) := \arg\max_{\omega} \langle \omega, \xi \rangle \ \text{ s.t. } \ \|\Pi^\top \omega\|_\infty \leq 1. \tag{6}$$

Let $d(\Pi, \xi)$ be the objective value of the dual problem (6). If the primal problem is feasible, then strong duality holds and $p(\Pi, \xi) = d(\Pi, \xi)$.

2.2. Geometric characterization of the dictionary

Our subspace-sparse recovery conditions rely on geometric properties of the dictionary Π that characterize the distribution of the inliers and the separation between inliers and outliers. The distribution of the inliers is characterized by the inradius $r(\mathcal{K}(\pm\Phi))$ of the symmetric convex body $\mathcal{K}(\pm\Phi) = \text{conv}\{\pm\phi_1, \cdots, \pm\phi_M\}$, where conv$\{\cdot\}$ denotes the convex hull. The notions of convex body, symmetric convex body and inradius are defined as follows.

Definition 1 (Symmetric convex body)**.** *A convex set $\mathcal{P}$ that satisfies $\mathcal{P} = -\mathcal{P}$ is called symmetric. A compact convex set with nonempty interior is called a convex body.*

Definition 2 (Inradius)**.** *The (relative) inradius $r(\mathcal{P})$ of a convex body $\mathcal{P}$ is defined as the radius of the largest Euclidean ball in the space span$(\mathcal{P})$ that is inscribed in $\mathcal{P}$.*

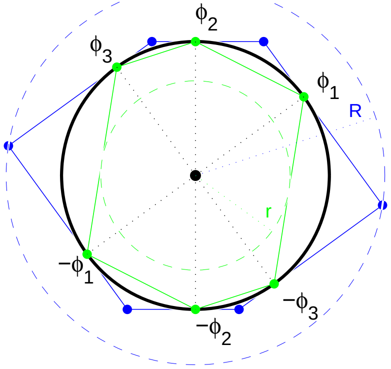

Figure 1. Illustration of inlier characterization. Dictionary atoms are $\{\phi_i\}_{i=1}^3$ that lie on the unit circle (drawn in black) of a two-dimensional subspace. $\mathcal{K}(\pm\Phi)$ and its inradius are illustrated in green. The polar set $\mathcal{K}^o(\pm\Phi)$ and its circumradius are illustrated in blue. The six blue dots are the dual points.

The inradius $r(\mathcal{K}(\pm\Phi))$ characterizes the distribution of the inliers in $\mathcal{R}(\Phi)$: if the atoms are well distributed across all directions the inradius is large, while if the atoms are skewed towards certain directions the inradius is small (see Figure 1 for an illustration in $\mathbb{R}^2$).

Another characterization of the distribution of the inliers is in terms of the circumradius of the polar set of $\mathcal{K}(\pm\Phi)$:

Definition 3 (Polar Set)**.** *The (relative) polar of a set $\mathcal{P}$ is defined as $\mathcal{P}^o = \{\eta \in span(\mathcal{P}) : \langle \eta, \theta \rangle \leq 1, \forall \theta \in \mathcal{P}\}$.*

Definition 4 (Circumradius)**.** *The circumradius $R(\mathcal{P})$ of a convex body $\mathcal{P}$ is defined as the radius of the smallest ball containing $\mathcal{P}$.*

Notice that the polar set of $\mathcal{K}(\pm\Phi)$ is given by $\mathcal{K}^o(\pm\Phi) = \{\eta \in \mathcal{R}(\Phi) : \|\Phi^\top \eta\|_\infty \leq 1\}$. It is also a symmetric convex body, as the polar of a convex body is also a convex body (Brazitikos et al., 2014).

A subset of the points in $\mathcal{K}^o(\pm\Phi)$ will play a critical role.

Definition 5 (Extreme Point)**.** *A point η in a convex set $\mathcal{P}$ is an extreme point if it cannot be expressed as a strict convex combination of two other points in $\mathcal{P}$, i.e. there are no $\lambda \in (0,1)$, $\eta_1, \eta_2 \in \mathcal{P}$, $\eta_1 \neq \eta_2$, such that $\eta = (1-\lambda)\eta_1 + \lambda\eta_2$.*

Definition 6 (Dual Point)**.** *The set of dual points of the matrix $\Phi = [\phi_1, \cdots, \phi_M]$, denoted by $\mathcal{D}(\Phi)$, is defined as the set of extreme points of the set $\mathcal{K}^o(\pm\Phi)$.*

The following result gives a relationship between the inradius of a set and the circumradius of its polar set.

Theorem 1 (Soltanolkotabi & Candès (2013))**.** *Let $\mathcal{P}$ be a symmetric convex body and $\mathcal{P}^o$ be its polar. Then we have $r(\mathcal{P})R(\mathcal{P}^o) = 1$.*

Applying the above theorem to $\mathcal{K}(\pm\Phi)$ we get

$$r(\mathcal{K}(\pm\Phi)) \cdot R(\mathcal{K}^o(\pm\Phi)) = 1. \tag{7}$$

The following result is well-studied in linear programming. We sketch its proof, as it provides bounds on the size of the dual set, which we will use later.

Theorem 2 (Nocedal & Wright (2006))**.** *The set $\mathcal{D}(\Phi)$ is finite.*

Proof sketch. Consider a linear program with variable η, constraint $\eta \in \mathcal{K}^o(\pm\Phi)$, and arbitrary objective. Since the dual points $\mathcal{D}(\Phi)$ are the extreme points of $\mathcal{K}^o(\pm\Phi)$, they are the same as the basic feasible solutions of the linear program. Assume Φ has M columns and let $s = \dim(\mathcal{R}(\Phi))$. Each basic feasible solution is determined by s linearly independent constraints from the $2M$ constraints of $\|\Phi^\top\eta\|_\infty \leq 1$. Obviously, there are at most $2^s \cdot \binom{M}{s}$ ways to choose such set of constraints. It follows that there are finitely many dual points. □

Moreover, all points in $\mathcal{K}^o(\pm\Phi)$ are convex combinations of these finitely many dual points in $\mathcal{D}(\Phi)$. This is implied by the following stronger result (Brazitikos et al., 2014).

Theorem 3. *The set of the extreme points of a convex body $\mathcal{P}$ is the smallest subset of $\mathcal{P}$ with convex hull $\mathcal{P}$.*

At the end of this section, we present two definitions that characterize the incoherence between the outliers $\{\psi_i\}$ and (a subset of) points in the inlier subspace. Specifically, let

$$\mu(\Psi, \mathcal{R}(\Phi)) := \max_{\eta\in\mathcal{R}(\Phi)\setminus\{0\}} \|\Psi^\top\eta\|_\infty/\|\eta\|_2, \tag{8}$$
$$\mu(\Psi, \mathcal{D}(\Phi)) := \max_{\eta\in\mathcal{D}(\Phi)\setminus\{0\}} \|\Psi^\top\eta\|_\infty/\|\eta\|_2. \tag{9}$$

They measure how close the outliers are from the inlier subspace $\mathcal{R}(\Phi)$ or the dual points $\mathcal{D}(\Phi)$.

3. Subspace-Sparse Recovery

3.1. Major results

Throughout this section, we assume that the inlier and outlier points are normalized to unit Euclidean norm. Let BP(Π, ξ) and OMP(Π, ξ) be the (sets of) solutions given by the two algorithms. We present conditions under which the solutions BP(Π, ξ) and OMP(Π, ξ) are subspace-sparse for all the ξ in the inlier subspace $\mathcal{R}(\Phi)$. Concretely, we identify the following two conditions for our analysis.

Definition 7. *The dictionary $\Pi = [\Phi, \Psi]$ with normalized columns is said to satisfy the principal subspace-sparse recovery condition (PRC) if*

$$r(\mathcal{K}(\pm\Phi)) > \mu(\Psi, \mathcal{R}(\Phi)), \tag{10}$$

where $\mathcal{R}(\cdot)$ is the range of the matrix. It is said to satisfy the dual subspace-sparse recovery condition (DRC) if

$$r(\mathcal{K}(\pm\Phi)) > \mu(\Psi, \mathcal{D}(\Phi)), \tag{11}$$

where $\mathcal{D}(\cdot)$ is the set of dual points.

The results for subspace-sparse recovery are as follows.

Theorem 4. *If $\Pi = [\Phi, \Psi]$ satisfies the PRC, then BP(Π, ξ) and OMP(Π, ξ) are both subspace-sparse for all $\xi \in \mathcal{R}(\Phi)$.*

Theorem 5. *If $\Pi = [\Phi, \Psi]$ satisfies the DRC, then BP(Π, ξ) and OMP(Π, ξ) are both subspace-sparse for all $\xi \in \mathcal{R}(\Phi)$.*

As both theorems show, two major factors affect subspace-sparse recovery. The first is that the inlier points should be well populated and spread out across the subspace $\mathcal{R}(\Phi)$, as measured by the inradius on the left hand side of (10) and (11). Specifically, as the inliers get denser, the inradius increases to 1. The second factor is that the outlier points should be incoherent with $\mathcal{R}(\Phi)$ in the case of PRC or $\mathcal{D}(\Phi)$ in the case of DRC. In the extreme case where the outliers are all in the orthogonal complement of $\mathcal{R}(\Phi)$, both two coherences are zero.

Furthermore, note that the incoherence for PRC is measured between outliers and all points in the subspace $\mathcal{R}(\Phi)$. The DRC, however, is a weaker requirement since it only needs the outliers to be incoherent with $\mathcal{D}(\Phi)$, a finite subset of $\mathcal{R}(\Phi)$. Thus, Theorem 4 is implied by Theorem 5.

These two results, alongside with some auxiliary results, are summarized in Figure 2. Each box contains a proposition, and the arrows denote implication relations. The topmost and the bottommost boxes are the properties of subspace-sparse recovery by BP and OMP that we are pursuing. Both of them are implied by the PRC and the DRC. In the following, we discuss in more detail theories of subspace-sparse recovery by BP and OMP.

3.2. Subspace-sparse recovery by BP

We first establish an equivalent condition for subspace-sparse recovery from BP, then show that this condition is implied by PRC and DRC. See the upper half of Figure 2 for an illustration.

An equivalent condition for BP to give subspace-sparse solutions appears in the context of subspace clustering (Elhamifar & Vidal, 2013). We rephrase the result here for our problem and the proof is omitted.

Theorem 6. *(Elhamifar & Vidal, 2013) Given $\Pi = [\Phi, \Psi]$, $BP(\Pi, \xi)$ is subspace-sparse for all $\xi \in \mathcal{R}(\Phi)$ if and only if $p(\Phi, \xi) < p(\Psi, \xi)$ for all $\xi \in \mathcal{R}(\Phi) \setminus \{0\}$.*

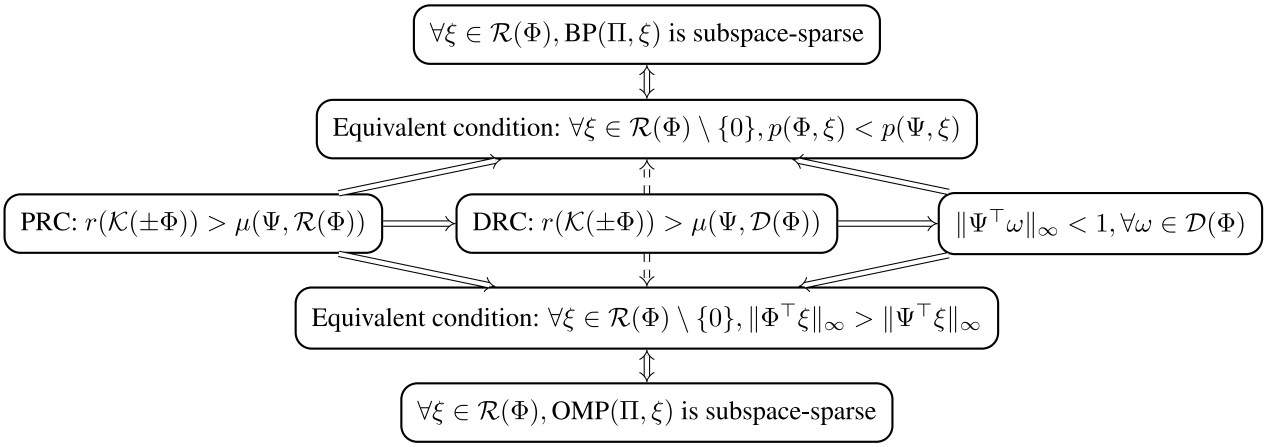
$\forall \xi \in \mathcal{R}(\Phi)$, BP$(\Pi, \xi)$ is subspace-sparse
Equivalent condition: $\forall \xi \in \mathcal{R}(\Phi) \setminus \{0\}, p(\Phi, \xi) < p(\Psi, \xi)$
PRC: $r(\mathcal{K}(\pm\Phi)) > \mu(\Psi, \mathcal{R}(\Phi))$
DRC: $r(\mathcal{K}(\pm\Phi)) > \mu(\Psi, \mathcal{D}(\Phi))$
$\|\Psi^\top \omega\|_\infty < 1, \forall \omega \in \mathcal{D}(\Phi)$
Equivalent condition: $\forall \xi \in \mathcal{R}(\Phi) \setminus \{0\}, \|\Phi^\top \xi\|_\infty > \|\Psi^\top \xi\|_\infty$
$\forall \xi \in \mathcal{R}(\Phi)$, OMP$(\Pi, \xi)$ is subspace-sparse

Figure 2. Summary of the results of subspace-sparse recovery with dictionary $\Pi = [\Phi, \Psi]$. Each box contains a proposition, and arrows denote implications. The topmost (resp., bottommost) box is the property of subspace-sparse recovery by BP (resp., OMP). Two major conditions for subspace-sparse recovery are the PRC and the DRC. In this paper we prove the solid arrows.

In the equivalent condition, it is required that for any $\xi \in \mathcal{R}(\Phi) \setminus \{0\}$, $p(\Phi, \xi)$, which is the objective value of BP for recovering ξ by *inlier* dictionary Φ (see (5)), should be smaller than $p(\Psi, \xi)$, which is the objective value of recovering by *outlier* dictionary Ψ.

We proceed to prove the result of PRC in Theorem 4. One way to do this is by arguing that the PRC implies DRC, so Theorem 4 is implied by Theorem 5. However, we choose to prove the result directly by arguing that the PRC implies the equivalent condition, since this proof reveals certain properties of the problem.

In the equivalent condition, $p(\Phi, \xi)$ depends purely on the properties of inliers, while $p(\Psi, \xi)$ depends on a relation between the outliers and the subspace spanned by inliers. This enlightens us to upper bound the former by inliers characterization, and to lower bound the latter by inlier-outlier relations.

Theorem 7. *If PRC:* $r(\mathcal{K}(\pm\Phi)) > \mu(\Psi, \mathcal{R}(\Phi))$ *holds then* $\forall \xi \in \mathcal{R}(\Phi) \setminus \{0\}, p(\Phi, \xi) < p(\Psi, \xi)$.

Proof. We bound the left and right hand sides of the objective inequality separately.

First, we notice that $p(\Phi, \xi) = d(\Phi, \xi) = \langle \omega, \xi \rangle$ by strong duality, in which ω is dual optimal solution. Decompose ω into two orthogonal components $\omega = \omega^\perp + \omega^\parallel$, in which $\omega^\parallel \in \mathcal{R}(\Phi)$. It has $\|\Phi^\top \omega^\parallel\|_2 = \|\Phi^\top \omega\|_2 \leq 1$, then from (7), $\|\omega^\parallel\|_2 \leq 1/r(\mathcal{K}(\pm\Phi))$. Thus,

$$\begin{aligned} p(\Phi, \xi) &= \langle \omega^\parallel, \xi \rangle \\ &\leq \|\xi\|_2 \|\omega^\parallel\|_2 \leq \|\xi\|_2 / r(\mathcal{K}(\pm\Phi)). \end{aligned} \tag{12}$$

On the other hand, consider the optimization problem

$$P(\Psi, \xi) = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \ \text{ s.t. } \ \Psi\mathbf{x} = \xi. \tag{13}$$

If the problem is infeasible, then the objective of the above optimization $p(\Psi, \xi) = +\infty$, the conclusion follows trivially. Otherwise, take any $\mathbf{x}^* \in P(\Psi, \xi)$ to be the optimal solution, we have $\xi = \Psi\mathbf{x}^*$. Left multiply by $\xi^\top$ and manipulate the right hand side we have the following:

$$\begin{aligned} \|\xi\|_2^2 &= \xi^\top \Psi \mathbf{x}^* \leq \|\Psi^\top \xi\|_\infty \|\mathbf{x}^*\|_1 \\ &= \|\Psi^\top \frac{\xi}{\|\xi\|_2}\|_\infty \|\xi\|_2 \cdot p(\Psi, \xi) \\ &\leq \mu(\Psi, \mathcal{R}(\Phi)) \cdot \|\xi\|_2 \cdot p(\Psi, \xi), \end{aligned} \tag{14}$$

so $p(\Psi, \xi) \geq \|\xi\|_2 / \mu(\Psi, \mathcal{R}(\Phi))$.

The conclusion thus follows by combining the above two parts and the condition of PRC. □

While PRC requires all points in $\mathcal{R}(\Phi)$ to be incoherent with the outliers, the DRC shows that a finite subset of $\mathcal{R}(\Phi)$ is in fact sufficient, i.e. we only need $\mathcal{D}(\Phi)$ to be incoherent with the outliers. To prove this claim, we need a statement that is weaker than DRC but is more convenient to work with, see the rightmost box of Figure 2.

Lemma 1. *If DRC:* $r(\mathcal{K}(\pm\Phi)) > \mu(\Psi, \mathcal{D}(\Phi))$ *holds then* $\|\Psi^\top \omega\|_\infty < 1, \forall \omega \in \mathcal{D}(\Phi)$.

Proof. For any $\omega \in \mathcal{D}(\Phi)$, we know that $\omega \in \mathcal{K}^o(\pm\Phi)$, so by (7), $\|\omega\|_2 \leq R(\mathcal{K}^o(\pm\Phi)) = 1/r(\mathcal{K}(\pm\Phi))$, thus

$$\begin{aligned} \|\Psi^\top \omega\|_\infty &= \|\Psi^\top \frac{\omega}{\|\omega\|_2}\|_\infty \|\omega\|_2 \\ &\leq \mu(\Psi, \mathcal{D}(\Phi)) / r(\mathcal{K}(\pm\Phi)) < 1. \end{aligned} \tag{15}$$

□

Theorem 8. *If* $\|\Psi^\top \omega\|_\infty < 1, \forall \omega \in \mathcal{D}(\Phi)$ *holds then* $\forall \xi \in \mathcal{R}(\Phi) \setminus \{0\}$, $p(\Phi, \xi) < p(\Psi, \xi)$.

Proof. To prove the result, we need some basic results from linear programming. Consider the linear program:

$$\arg\max_{w} \langle \omega, \xi \rangle \;\text{ s.t. }\; \|\Phi^\top \omega\|_\infty \le 1, \omega \in \mathcal{R}(\Phi). \tag{16}$$

Note that the feasible region of (16) is $\mathcal{K}^o(\pm\Phi)$, and it is bounded because it is a convex body. By theories of linear programming (e.g., Nocedal & Wright (2006)), there must have a solution to (16) that is an extreme point of $\mathcal{K}^o(\pm\Phi)$. Thus, we can always find a solution of (16) that is in the set of dual points $\mathcal{D}(\Phi)$.

Now let us consider the optimization problem $D(\Phi, \xi)$, rewritten below for convenience:

$$D(\Phi, \xi) := \arg\max_{w} \langle \omega, \xi \rangle \;\text{ s.t. }\; \|\Phi^\top \omega\|_\infty \le 1. \tag{17}$$

Note that this program differs from (16) only in the constraint. The claim is, despite of this change, there is still at least one optimal solution to (17) that is in $\mathcal{D}(\Phi)$. This follows from the fact that both ξ and the columns of Φ are in $\mathcal{R}(\Phi)$, thus any solution ω to (17) can be decomposed into two parts as $\omega = \omega^{\|} + \omega^{\perp}$, in which $\omega^{\|}$ is a solution to (16) and $\omega^{\perp}$ is orthogonal to $\mathcal{R}(\Phi)$.

Prepared with the above discussion, we now go to the proof. The proof is trivial if $p(\Psi, \xi) = +\infty$, since $p(\Phi, \xi)$ always has feasible solutions and thus is finite.

Otherwise, take any $\mathbf{x}^* \in P(\Psi, \xi)$ to be a primal optimal solution. It has that $\xi = \Psi\mathbf{x}^*$. On the other hand, we have shown that there exists an optimal dual solution $\omega^* \in D(\Phi, \xi)$ that is in $\mathcal{D}(\Phi)$. Thus,

$$\begin{aligned} p(\Phi, \xi) = d(\Phi, \xi) = \langle \omega^*, \xi \rangle = \langle \omega^*, \Psi\mathbf{x}^* \rangle \\ \le \|\Psi^\top \omega^*\|_\infty \cdot \|\mathbf{x}^*\|_1 < p(\Psi, \xi), \end{aligned} \tag{18}$$

in which $\|\Psi^\top \omega^*\|_\infty < 1$ by assumption, and $\|\mathbf{x}^*\|_1 = p(\Psi, \xi)$ since $\mathbf{x}^*$ is an optimal solution. □

3.3. Subspace-sparse recovery by OMP

The lower half of Figure 2 summarizes the results for sparse recovery by OMP. The results surprisingly have a symmetric structure as that of BP. First, we show an equivalent condition for subspace-sparse recovery by OMP. Then we show that this condition is implied by PRC and DRC.

Theorem 9. *$\forall \xi \in \mathcal{R}(\Phi)$, OMP$(\Pi, \xi)$ is subspace-sparse if and only if $\forall \xi \in \mathcal{R}(\Phi) \setminus \{0\}, \|\Phi^\top \xi\|_\infty > \|\Psi^\top \xi\|_\infty$.*

Proof. The "only if" part is straight forward because if $\|\Phi^\top \xi\|_\infty \le \|\Psi^\top \xi\|_\infty$, then this specific ξ will pick an outlier in the first step of the OMP(Π, ξ).

The other direction is also easily seen in an inductive way if we consider the procedure of the OMP algorithm. Specifically, for any given $\xi \in \mathcal{R}(\Phi)$, the first step of OMP$(\Pi, \xi)$ chooses an entry from columns of Φ, and this gives a residual that is again in $\mathcal{R}(\Phi)$, which then guarantees that the next step of OMP(Π, ξ) also chooses an entry from the columns of Φ. □

Thus, the equivalent condition requires that for any point $\xi \in \mathcal{R}(\Phi) \setminus \{0\}$, the closest point to it in the entire dictionary Π should be an inlier point.

We now show that this equivalent condition is further implied by the PRC. Similar to the discussion for BP, the term $\|\Phi^\top \xi\|_\infty$ in the equivalent condition depends on inliers and can be lower bounded by means of an inradius characterization, and the term $\|\Psi^\top \xi\|_\infty$ depends on inlier-outlier relation and can be upper bounded by the coherence.

Theorem 10. *If PRC: $r(\mathcal{K}(\pm\Phi)) > \mu(\Psi, \mathcal{R}(\Phi))$ holds then $\forall \xi \in \mathcal{R}(\Phi) \setminus \{0\}, \|\Phi^\top \xi\|_\infty > \|\Psi^\top \xi\|_\infty$.*

Proof. We prove this by bounding each side of the objective inequality.

For the right hand side, we have that

$$\|\Psi^\top \xi\|_\infty = \|\Psi^\top \frac{\xi}{\|\xi\|_2}\|_\infty \cdot \|\xi\|_2 \le \mu(\Psi, \mathcal{R}(\Phi)) \cdot \|\xi\|_2.$$

For the left hand side, we will prove that $\|\Phi^\top \xi\|_\infty \ge r(\mathcal{K}(\pm\Phi))\|\xi\|_2$. Notice that

$$\begin{aligned} r(\mathcal{K}(\pm\Phi))\|\xi\|_2 &= \|\xi\|_2 / R(\mathcal{K}^o(\pm\Phi)) \\ &= \frac{\|\xi\|_2}{\max_\eta \|\eta\|_2 \;\text{ s.t. }\; \|\Phi^\top \eta\|_\infty \le 1}. \end{aligned}$$

The optimization program in the denominator could be lower bounded by taking $\eta = \xi / \|\Phi^\top \xi\|_\infty$, thus

$$r(\mathcal{K}(\pm\Phi))\|\xi\|_2 \le \frac{\|\xi\|_2}{\|\xi\|_2 / \|\Phi^\top \xi\|_\infty} = \|\Phi^\top \xi\|_\infty. \tag{19}$$

The conclusion thus follows by concatenating the bounds for both sides above with the PRC. □

Finally, we prove the result for DRC, by showing that the statement in the rightmost box of Figure 2 guarantees the equivalent condition for OMP.

Theorem 11. *If $\|\Psi^\top \omega\|_\infty < 1, \forall \omega \in \mathcal{D}(\Phi)$ holds then $\forall \xi \in \mathcal{R}(\Phi) \setminus \{0\}, \|\Phi^\top \xi\|_\infty > \|\Psi^\top \xi\|_\infty$.*

To prove this theorem, we use the result that the polar set $\mathcal{K}^o(\pm\Phi)$ induces a norm on the space $\mathcal{R}(\Phi)$, by means of the so-called Minkowski functional.

Definition 8. *The Minkowski functional of a set $\mathcal{K}$ is defined on span$(\mathcal{K})$ as*

$$\|\eta\|_{\mathcal{K}} = \inf\{t > 0 : \frac{\eta}{t} \in \mathcal{K}\}. \tag{20}$$

Theorem 12. *(Vershynin, 2009) If $\mathcal{K}$ is a symmetric convex body, then $\|\cdot\|_{\mathcal{K}}$ is a norm on span$(\mathcal{K})$ with $\mathcal{K}$ being the unit ball.*

Proof of Theorem 11. It suffices to prove the result for every $\xi \in \mathcal{R}(\Phi) \setminus \{0\}$ that has a unit norm, by using any norm defined on $\mathcal{R}(\Phi)$. Here we take the norm $\|\cdot\|_{\mathcal{K}^o(\pm\Phi)}$, then we need to prove that $\|\Phi^\top \xi\|_\infty > \|\Psi^\top \xi\|_\infty$ for all $\xi \in \mathcal{R}(\Phi)$ such that $\|\xi\|_{\mathcal{K}^o(\pm\Phi)} = 1$.

Since $\|\xi\|_{\mathcal{K}^o(\pm\Phi)} = 1$, it has $\xi \in \mathcal{K}^o(\pm\Phi)$, thus ξ could be written as a convex combination of the dual points, i.e. one can write $\xi = \sum_i \omega_i x_i$ in which $\omega_i \in \mathcal{D}(\Phi), x_i \in [0,1]$ for all i and $\sum_i x_i = 1$. Thus,

$$\|\Psi^\top \xi\|_\infty = \|\Psi^\top \sum_i \omega_i \cdot x_i\|_\infty \leq \sum_i \|\Psi^\top \omega_i \cdot x_i\|_\infty < \sum_i x_i = 1 = \|\Phi^\top \xi\|_\infty.$$

The last equality follows from $\|\xi\|_{\mathcal{K}^o(\pm\Phi)} = 1$. □

4. Application to Sparse Recovery

The results of subspace-sparse recovery in the previous section can also be applied to the study of sparse recovery. In sparse recovery, the task is to reconstruct an M-sparse signal $\mathbf{x}$ (i.e. $\mathbf{x}$ has at most M nonzero entries) from the observation $\xi = \Pi\mathbf{x}$ for some dictionary Π. By taking the inliers Φ to be the M columns corresponding to the nonzero entries of $\mathbf{x}$, if $s :=$ rank(Φ) is equal to M, then the subspace-sparse solution is unique and is exactly $\mathbf{x}$. In this case, if the dictionary $\Pi = [\Phi, \Psi]$ satisfies PRC or DRC, then an M-sparse recovery of $\mathbf{x}$ can be achieved. Formally,

Theorem 13. *Given a dictionary Π, any M-sparse vector $\mathbf{x}$ can be recovered from the observation $\xi := \Pi\mathbf{x}$ by BP and OMP if for any partition of Π into Φ and Ψ where Φ has M columns, it has $s :=$ rank(Φ) is equal to M and that PRC (respectively, DRC) holds for such partition.*

This result serves as a new condition for guaranteeing reconstruction of sparse signals. Its geometric interpretation is the same as that of PRC and DRC for the subspace-sparse recovery, i.e., for any M atoms of the dictionary, they should be well distributed in their span, while all other atoms should be sufficiently away from this span (by PRC) or from a subset of the span (by DRC).

For the purpose of checking the conditions of the theorem, if for a partition of Π into $[\Phi, \Psi]$ it is true that rank$(\Phi) = M$, then subsequent checking of the PRC and DRC is easy, as explained below. First, the dual points $\mathcal{D}(\Phi)$ can be written out explicitly:

Lemma 2. *If the inlier dictionary $\Phi \in \mathbb{R}^{n \times M}$ has full column rank, then the set of dual points, $\mathcal{D}(\Phi)$, contains exactly 2^M points specified by $\{\Phi(\Phi^\top\Phi)^{-1} \cdot \mathbf{u}, \mathbf{u} \in U_M\}$, where $U_M := \{[u_1, \cdots, u_M], u_i = \pm 1, i = 1, \cdots, M\}$.*

The proof is in the appendix. With the dual points, one can then compute the coherence on the RHS of DRC. Moreover, $R(\mathcal{K}^o(\pm\Phi))$ can be computed as the maximum ℓ_2 norm of the dual points, and the inradius $r(\mathcal{K}(\pm\Phi))$ can be acquired as the reciprocal of $R(\mathcal{K}^o(\pm\Phi))$ (see (7)). Thus, all terms in PRC and DRC can be computed.

At the end of this section, we point out that the result of Theorem 13 can be compared with traditional sparse recovery results. Specifically, we compare it with the result that uses mutual coherence, $\mu(\Pi)$, which is defined as the largest absolute inner product between columns of Π. It is known that $\mu(\Pi) < \frac{1}{2M-1}$ is a sufficient condition for OMP and BP (Donoho & Elad, 2003; Tropp, 2004) to recover M-sparse signals. We show that this is a stronger requirement than that of Theorem 13.

Theorem 14. *Given a dictionary Π. If $\mu(\Pi) < \frac{1}{2M-1}$, then for any partition of Π into Φ and Ψ where Φ has M columns, it has rank$(\Phi) = M$ and that PRC and DRC hold.*

The proof is in the appendix. This result shows that the PRC/DRC conditions in Theorem 13 are implied by the condition of mutual coherence. While the mutual coherence condition requires all columns of Π to be incoherent from each other, the PRC and DRC provide more detailed requirements, in terms of the inlier distribution as well as inlier-outlier relations.

5. Application to Subspace Clustering

Let $\{\xi_j\}_{j=1}^N$ be a set of points drawn from a union of unknown subspaces $\{S_i\}_{i=1}^n$. Subspace clustering addresses the problem of clustering these points into their respective subspaces, without knowing their membership *a priori*.

Sparse Subspace Clustering (SSC) is one of the state-of-the-art approaches for this task. In this approach, subspace-sparse recovery is performed for each ξ_j by using BP (Elhamifar & Vidal, 2009) or OMP (Dyer et al., 2013). More specifically, for each ξ_j that is in one of the subspaces S_i, let the inlier matrix Φ_j associated with ξ_j contain in its columns all sample points in S_i except ξ_j itself, and let the outlier matrix Ψ_j contain in its columns all sample points in all subspaces except S_i. If the recovery of $\mathbf{x}_j$ using BP or OMP is subspace-sparse by the dictionary $\Pi_j := [\Phi_j, \Psi_j]$, then the nonzero entries of the solution identify some points that are also in subspace S_i. One can do this for all points $\{\xi_j\}_{j=1}^N$, and if all of them give subspace-sparse solutions, then connections are built only between points that are from the same subspace. Consequently, one can find clusters by extracting the connected components from the graph of connections.

Depending on whether BP or OMP is used for subspace-sparse recovery, there can be two different versions of SSC, which will be referred to as SSC-BP and SSC-OMP. In this section, we discuss the conditions under which SSC-BP and SSC-OMP can achieve subspace-sparse recovery for all of the sample points. In the following, we first show that the result of DRC can be applied for such a purpose, and we then compare our results to prior work.

5.1. Subspace-sparse recovery for SSC

We assume that all the points $\{\xi_j\}_{j=1}^N$ are normalized to have unit ℓ_2 norm. If a dictionary $\Pi_j = [\Phi_j, \Psi_j]$ satisfies the DRC, then Theorem 5 guarantees the subspace-sparse recovery of all points in S_i, including ξ_j, by BP and OMP. Thus, if the DRC holds for all dictionaries $\{\Pi_j\}_{j=1}^N$, then the correctness of SSC-BP and SSC-OMP is guaranteed.

We can rephrase this result in terms of properties of the subspaces and make it more interpretable. For S_i, we denote the minimum leave-one-out inradius of sample points in S_i as

$$r_i := \min_{j:\xi_j \in S_i} r(\mathcal{K}(\pm\Phi_j)), \tag{21}$$

where each inradius on the RHS is taken for sample points in S_i with ξ_j excluded. Note that r_i will be relatively large if sample points from S_i are well distributed. Define the dual points

$$\mathcal{D}_i = \cup_{j:\xi_j \in S_i} \mathcal{D}(\Phi_j), \tag{22}$$

which is the union of the leave-one-out duals of sample points in S_i. Finally, notice that for all j such that $\xi_j \in S_i$, the outlier dictionary Ψ_j is the same, and is composed of all sample points not in S_i. With a slight abuse of notation, we denote the common dictionary by Ψ_i.

Theorem 15. *Given $\{\xi_j\}_{j=1}^N$ that lie in a union of subspaces $\{S_i\}_{i=1}^n$, SSC-BP and SSC-OMP both give subspace-sparse solutions if*

$$\forall i = 1, \cdots, n, \quad r_i > \mu(\Psi_i, \mathcal{D}_i). \tag{23}$$

This theorem states that for each subspace, in order for SSC-BP and SSC-OMP to succeed, the sample points should be well distributed so that r_i is large enough, and the set of dual points $\mathcal{D}_i$, which all lie in S_i, should be incoherent with points from all other subspaces Ψ_i.

5.2. Comparison with prior work

While our result in Theorem 15 analyzes SSC-BP and SSC-OMP jointly, prior work has addressed the two problems separately. In particular, a closely related result to our work is given by (Soltanolkotabi & Candès, 2013), which gives the following sufficient condition for guaranteeing the correctness of SSC-BP:

$$\forall i = 1, \cdots, n, \quad r_i > \mu(\Psi_i, \mathcal{V}_i), \tag{24}$$

Here, for each i, $\mathcal{V}_i$ contains a set of points in S_i that are referred to as the dual directions. By comparing (24) with (23), we can see that the only difference is that $\mathcal{D}_i$ is replaced by $\mathcal{V}_i$ in the RHS. Since $\mathcal{V}_i$ is a subset of $\mathcal{D}_i$, we have that $\mu(\Psi_i, \mathcal{V}_i) \leq \mu(\Psi_i, \mathcal{D}_i)$. Consequently, condition (24) is easier to be satisfied, and thus is a better result for the analysis of SSC-BP.

However, we also point out that the conditions (23) and (24) are not entirely comparable, because the former has broader implications. First, note that (23) is also a sufficient condition for SSC-OMP to give subspace-sparse solutions, while (24) is valid for SSC-BP only. Second, if (23) holds, then for any vector ξ (not necessarily one of the ξ_j's) that is in the union of subspaces $\{S_i\}_{i=1}^n$, OMP and BP can give subspace-sparse solution with the dictionary composed of sample points $\{\xi_j\}_{j=1}^N$. As a result, the condition (23) can also be used for the analysis of other related algorithms such as subspace classification (Wright et al., 2009) and large scale SSC (Peng et al., 2013). Such topics are beyond the subject of this paper and are deferred for future work.

6. Conclusion

In this work, we have studied the properties of OMP and BP algorithms for the task of subspace-sparse recovery and have identified the PRC and DRC as two sufficient conditions for guaranteeing subspace-sparse recovery. This result provides new understanding of the performance of sparse recovery based classification and clustering techniques. Moreover, we also have established that the PRC and DRC are sufficient conditions for traditional sparse recovery. We believe that these results serve as new perspectives into the traditional sparse recovery problem.

Acknowledgments

The authors wish to thank Manolis Tsakiris and Benjamin Haeffele for their helpful comments. The authors also thank the support of NSF BIGDATA grant 1447822.

References

Baraniuk, Richard. Compressive sensing. *IEEE Signal Processing Magazine*, 24(4):118–121, 2007.

Brazitikos, S., Giannopoulos, A., Valettas, P., and Vritsiou, B.H. *Geometry of Isotropic Convex Bodies:*. Mathematical Surveys and Monographs. American Mathematical Society, 2014.

Cai, T. Tony and Zhang, Anru. Sparse representation of a polytope and recovery of sparse signals and low-rank matrices. *IEEE Transactions on Information Theory*, 60 (1):122–132, 2014.

Candès, E. and Wakin, M. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, Mar. 2008.

Candès, Emmanuel and Tao, Terence. Decoding by linear programming. *IEEE Trans. on Information Theory*, 51 (12):4203–4215, 2005.

Chen, S. S., Donoho, D. L., and Saunders, M. A. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20:33–61, 1998.

Davenport, M. A. and Wakin, M. B. Analysis of orthogonal matching pursuit using the restricted isometry property. *IEEE Transactions on Information Theory*, 56(9):4395–4401, 2010.

Donoho, D. L. and Elad, M. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of National Academy of Sciences*, 100(5):2197–2202, 2003.

Donoho, D. L., Elad, M., and Temlyakov, V. N. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. on Information Theory*, 52(1):6–18, Jan. 2006.

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

Elhamifar, E. and Vidal, R. Sparse subspace clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Elhamifar, E. and Vidal, R. Clustering disjoint subspaces via sparse representation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2010.

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

Nocedal, Jorge and Wright, Stephen J. *Numerical Optimization, second edition*. World Scientific, 2006.

Pati, Y., Rezaiifar, R., and Krishnaprasad, P. Orthogonal matching pursuit: recursive function approximation with application to wavelet decomposition. In *Asilomar Conference on Signals, Systems and Computation*, 1993.

Peng, Xi, Zhang, Lei, and Yi, Zhang. Scalable sparse subspace clustering. pp. 430–437, 2013.

Soltanolkotabi, M. and Candès, E. J. A geometric analysis of subspace clustering with outliers. *Annals of Statistics*, 2013.

Soltanolkotabi, Mahdi, Elhamifar, Ehsan, and Candès, Emmanuel J. Robust subspace clustering. *Annals of Statistics*, 42(2):669–699, 2014.

Tropp, J. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, Oct. 2004.

Vershynin, Roman. Lectures in geometric functional analysis. 2009.

Vidal, R. Subspace clustering. *IEEE Signal Processing Magazine*, 28(3):52–68, March 2011.

Wang, Yu-Xiang and Xu, Huan. Noisy sparse subspace clustering. In *Proceedings of International Conference on Machine Learning*, 2013.

Wright, J., Yang, A., Ganesh, A., Sastry, S., and Ma, Y. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, Feb. 2009.