Dual Principal Component Pursuit for Robust Subspace Learning: Theory and Algorithms for a Holistic Approach

Tianyu Ding [1] **Zhihui Zhu** [2] **René Vidal** [3] **Daniel P. Robinson** [4]

Abstract

The Dual Principal Component Pursuit (DPCP) method has been proposed to robustly recover a subspace of high relative dimension from corrupted data. Existing analyses and algorithms of DPCP, however, mainly focus on finding a normal to a single *hyperplane* that contains the inliers. Although these algorithms can be extended to a *subspace* of higher codimension through a recursive approach that sequentially finds a new basis element of the space orthogonal to the subspace, this procedure is computationally expensive and lacks convergence guarantees. In this paper, we consider a DPCP approach for *simultaneously* computing the entire basis of the orthogonal complement subspace (we call this a *holistic approach*) by solving a non-convex non-smooth optimization problem over the Grassmannian. We provide geometric and statistical analyses for the global optimality and prove that it can tolerate as many outliers as the square of the number of inliers, under both noiseless and noisy settings. We then present a Riemannian regularity condition for the problem, which is then used to prove that a Riemannian subgradient method converges linearly to a neighborhood of the orthogonal subspace with error proportional to the noise level.

1. Introduction

Robustly learning a linear subspace from high-dimensional corrupted data has become a fundamental problem in machine learning, pattern recognition, and computer vision. Two forms of corruption commonly appear in real data: outliers and noise. Unlike inliers, which exactly lie in the subspace, *outliers* are far from the subspace and do not exhibit linear structure. Noise in the data means that the inliers are perturbed so that they lie close to or on the subspace, i.e., they are *noisy inliers*. Such corruptions of the dataset cause significant challenges to the subspace recovery task.

In the past decade, many robust subspace recovery (RSR) methods have been proposed (Lerman & Maunu, 2018b) with the assumption that high-dimensional data can be well-approximated by low-dimensional structures. Among them the representatives include robust PCA (Brooks et al., 2013; Markopoulos et al., 2018; Vaswani & Narayanamurthy, 2018), sparse subspace clustering (Elhamifar & Vidal, 2013; You et al., 2016) and low-rank matrix methods (Rahmani & Atia, 2017; Xu et al., 2012), which are normally solved by convex optimization. However, the guarantees for theory and algorithms are developed for a low-dimensional underlying structure with $d \ll D$, where d and D are the subspace dimension and ambient dimension, respectively. This may be violated in the *high relative dimension* regime where $d/D \approx 1$. For example, many computer vision applications involve learning a hyperplane ($d = D - 1$), such as pose estimation in multi-view geometry (Hartley & Zisserman, 2003) and 3D point cloud analysis (Geiger et al., 2013; Silberman et al., 2012). It is also observed that in ImageNet (Deng et al., 2009) the subspace spanned by deep features extracted from images of a single object category is of high relative dimension, making the existing methods for outlier detection theoretically inapplicable to ImageNet.

Dual Principal Component Pursuit (DPCP) (Tsakiris & Vidal, 2018) is an RSR method developed to tackle the high relative dimension regime directly since it estimates a basis for the orthogonal complement of a subspace $\mathcal{S} \subset \mathbb{R}^D$ by solving a non-convex ℓ_1 co-sparse problem on the sphere:

$$\min_{\boldsymbol{b} \in \mathbb{R}^D} \|\widetilde{\boldsymbol{\mathcal{X}}}^\top \boldsymbol{b}\|_1 \ \text{s.t.} \ \ \|\boldsymbol{b}\|_2 = 1, \tag{1}$$

where $\widetilde{\boldsymbol{\mathcal{X}}} \in \mathbb{R}^{D \times L}$ is the dataset contaminated by outliers and noise. In sharp contrast to the existing RSR methods that that can tolerate at best $M = O(N)$ outliers with N inliers ($L = M + N$), DPCP can handle $M = O(N^2)$ outliers even in the noisy case (Ding et al., 2019; Zhu et al., 2018). However, problem (1) can only find a normal to a single

[1]Department of Applied Mathematics & Statistics, Johns Hopkins University, USA [2]School of Electrical and Computer Engineering, University of Denver, USA [3]Mathematical Institute for Data Science, Johns Hopkins University, USA [4]Department of Industrial and Systems Engineering, Lehigh University, USA. Correspondence to: Tianyu Ding <tding1@jhu.edu>, Zhihui Zhu <zhihui.zhu@du.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

hyperplane, while recovering a *subspace* with codimension $c = D - d > 1$ requires recursive applications of (1) for c times, with each time finding a normal to $\mathcal{S}$ that is also orthogonal to previously computed normal vectors. This procedure is computationally expensive and lacks a convergence analysis. Moreover, the error accumulated during the recursion makes its behavior difficult to analyze.

In this paper, we consider *simultaneously* estimating the entire basis of the orthogonal complement subspace $\mathcal{S}^\perp$ by

$$\min_{\boldsymbol{B}\in\mathbb{R}^{D\times c},\ \boldsymbol{B}^\top\boldsymbol{B}=\mathbf{I}} \Big\{ f(\boldsymbol{B}) := \sum_{j=1}^{L} \|\widetilde{\boldsymbol{x}}_j^\top \boldsymbol{B}\|_2 \Big\}. \tag{2}$$

We call problem (2) a *holistic approach* as compared with the recursive approach with problem (1). Note that (2) is a natural extension to (1) that seeks a matrix $\boldsymbol{B}$ with orthonormal columns that are orthogonal to as many data points as possible. Observe that (2) is an optimization problem on the Grassmannian $\mathbb{G}(D, c)$ (Edelman et al., 1998), i.e., the set of c-dimensional subspaces in $\mathbb{R}^D$, and thus is inherently non-convex. Recently Zhu et al. (2019) proposed a Riemannian Subgradient method (RSGM) for solving a general non-convex optimization problem on the Grassmannian, and showed in the noiseless case that the RSGM applied to (2) converges linearly to an orthonormal basis of $\mathcal{S}^\perp$. Nevertheless, it is unclear whether a similar guarantee holds in the noisy setting. Moreover, it is reasonable to ask under what conditions every global solution $\boldsymbol{B}^*$ of (2) is an orthonormal basis of $\mathcal{S}^\perp$ when no noise is present, or how the principal angles between $\mathrm{Span}(\boldsymbol{B}^*)$ and $\mathcal{S}^\perp$ behave as a function of the noise level when the data is noisy.

Contributions. We provide geometric and statistical analyses for the *global* optimality of the non-convex DPCP problem (2) under both noiseless and noisy settings. We show that with noiseless data, under certain conditions, any global solution $\boldsymbol{B}^*$ of (2) is an orthonormal basis of $\mathcal{S}^\perp$. As the dataset is further contaminated with noise, we show that the subspace angle between $\mathrm{Span}(\boldsymbol{B}^*)$ and $\mathcal{S}^\perp$ is upper bounded by an amount that is proportional to the noise level. In both cases, we derive probabilistic arguments showing that the DPCP problem (2) can handle $M = O(N^2)$ outliers, which is superior to other existing RSR methods that can tolerate at best $O(N)$ outliers in theory. Moreover, we prove that the RSGM, with a proper initialization and a geometrically diminishing step size choice, converges linearly to a neighborhood of $\mathcal{S}^\perp$ whose radius is proportional to the noise level, and thus generalizes the result in Zhu et al. (2019). Experiments on synthetic data show that the holistic approach (2) performs favorably against the recursive approach (1) as well as other RSR methods in the high relative dimension regime.

2. Related Work

Principal Component Analysis (PCA) (Jolliffe, 1986) is the conventional method of fitting a linear subspace to data. PCA works well even when the data is noisy, but it is limited when the dataset is corrupted by outliers since the ℓ_2-based loss in PCA is sensitive to outliers. Another classical approach is the Random Sample Consensus (RANSAC) (Fischler & Bolles, 1981), which repeatedly estimates a subspace from d randomly sampled points (d is the dimension of the underlying subspace) within a time budget and then outputs the best result according to the number of points being categorized as inliers. Although RANSAC is popular in practice, it is sensitive to a thresholding parameter and the allocated time budget, and its exponential complexity limits its impact in the high relative dimension regime.

There are numerous RSR methods proposed in recent years, and we refer to Lerman & Maunu (2018b) for a comprehensive review. We limit our scope to approaches based on *least absolute deviations*, which minimize the sum of the distances between all data points and the subspace—exactly the formulation (2) considered in this paper. Most of these existing methods are designed for the low-relative dimension case ($d \ll D$). Parallel to this work, they consider solving non-convex optimization problems over $\mathbb{G}(D, d)$ instead of $\mathbb{G}(D, c)$. For example, Maunu et al. (2019) directly estimates a basis $\boldsymbol{V} \in \mathbb{R}^{D\times d}$ for the underlying subspace $\mathcal{S}$ using a formulation similar to (2), which is solved by a Geodesic Gradient Descent (GGD) method with a guarantee of linear convergence, but only providing a *local* optimality analysis and proving that it can handle $O(N)$ outliers. Similarly, Lerman & Maunu (2018a) take the same optimization problem on $\mathbb{G}(D, d)$ and solve it with Iteratively Reweighted Least Squares (IRLS), but its theoretical guarantee is even weaker than that of GGD. There are also many other methods that pertain to least absolute deviations but rely on convex relaxations. For example, GMS (Zhang & Lerman, 2014) and REAPER (Lerman et al., 2015) have similar objective functions as (2), but their constraints are constructed to be convex. Other examples are McCoy et al. (2011); Xu et al. (2012), which involve convex low-rank optimization but for different data modeling formulations.

The theoretical guarantees for the above RSR methods are usually violated in the high relative dimension regime ($d/D \approx 1$), and the algorithms become computationally expensive since optimizing over $\mathbb{G}(D, d)$ is very inefficient. To the best of our knowledge, DPCP is the only method that directly aims at recovering a subspace $\mathcal{S}$ of high relative dimension. Tsakiris & Vidal (2018) first introduced the idea of recursively learning a basis for $\mathcal{S}^\perp$ by solving (1). However, its global optimality analysis is difficult to interpret, thus making it unclear how many outliers it can tolerate. Furthermore, it proposes to solve (1) with IRLS without con-

vergence guarantees, while a provable linear programming based approach is inefficient. Zhu et al. (2018) improves the analysis of (1) with interpretable and tighter geometric quantities, and for the first time it shows that DPCP can handle $O(N^2)$ outliers. Moreover, it proposes an efficient Projected Subgradient Method that converges linearly with proper initialization. Nevertheless, neither Tsakiris & Vidal (2018) nor Zhu et al. (2018) consider the corruption of noise in the dataset. Ding et al. (2019) bridges this gap by extending the theoretical and convergence analysis in Zhu et al. (2018) to noisy data. So far, the previous work on DPCP based on solving (1) computes a normal to a single *hyperplane* that contains the inliers, which is cumbersome and inefficient when recursively applied to finding a new basis element of $\mathcal{S}^\perp$ for a *subspace* of higher codimension. Our work successfully addresses this issue by proposing a holistic approach that simultaneously computes the entire basis of $\mathcal{S}^\perp$, and provides global optimality and convergence theory of (2) under both noiseless and noisy settings.

3. Background

The setting considered in this paper is a unit ℓ_2-norm dataset $\widetilde{\boldsymbol{\mathcal{X}}} = [\boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{E}} \ \boldsymbol{\mathcal{O}}]\boldsymbol{\Gamma} \in \mathbb{R}^{D\times L}$, where $\boldsymbol{\mathcal{X}} = [\boldsymbol{x}_1 \ \cdots \ \boldsymbol{x}_N] \in \mathbb{R}^{D\times N}$ are inlier points within a d-dimensional subspace $\mathcal{S}$ of $\mathbb{R}^D$, $\boldsymbol{\mathcal{E}} = [\boldsymbol{\epsilon}_1 \ \cdots \ \boldsymbol{\epsilon}_N] \in \mathbb{R}^{D\times N}$ are additive noise imposed on inliers, $\boldsymbol{\mathcal{O}} = [\boldsymbol{o}_1 \ \cdots \ \boldsymbol{o}_M] \in \mathbb{R}^{D\times M}$ are outlier points in $\mathbb{R}^D$ that do not exhibit linear structure, and $\boldsymbol{\Gamma}$ is an unknown permutation matrix. Our goal is to recover the underlying subspace $\mathcal{S}$ from the corrupted data $\widetilde{\boldsymbol{\mathcal{X}}}$. We let $c := D - d$ denote the codimension of $\mathcal{S}$. Since we are interested in the high relative dimension regime with $c \ll d$, it is more efficient to find the dual subspace $\mathcal{S}^\perp$ instead of $\mathcal{S}$. Intuitively, in the noiseless case ($\boldsymbol{\mathcal{E}} = \boldsymbol{0}$), if $\boldsymbol{B}$ is an orthonormal basis of $\mathcal{S}^\perp$, then $f(\boldsymbol{B})$ in (2) only depends on the outliers and is insensitive to the choice of $\boldsymbol{B}$ since outliers are unstructured, which motivates our formulation (2).

We parameterize the Grassmannian $\mathbb{G}(D,c)$ with orthonormal matrices in the set $\mathbb{O}(D,c) := \{\boldsymbol{B} \in \mathbb{R}^{D\times c} : \boldsymbol{B}^\top\boldsymbol{B} = \boldsymbol{I}\}$. In particular, when $c = 1$, we also use $\mathbb{S}^{D-1}$, i.e., the unit sphere, as a substitute for $\mathbb{O}(D,1)$. In addition, we denote $\mathbb{O}(c,c)$ by $\mathbb{O}(c)$ for simplicity. Let $\boldsymbol{S}^\perp \in \mathbb{O}(D,c)$ be an orthonormal basis of $\mathcal{S}^\perp$. Since f in (2) is rotational invariant, we consider equivalence classes of matrices. In particular, for $\boldsymbol{U},\boldsymbol{V} \in \mathbb{G}(D,c)$ we say $\boldsymbol{U}$ is equivalent to $\boldsymbol{V}$ if $\mathrm{Span}(\boldsymbol{U}) = \mathrm{Span}(\boldsymbol{V})$, and use $\boldsymbol{U}$ to represent the equivalence class $[\boldsymbol{U}] := \{\boldsymbol{U}\boldsymbol{R} : \boldsymbol{R} \in \mathbb{O}(c)\}$. As the dataset is contaminated with noise, a solution $\boldsymbol{B}$ is expected to be perturbed away from $\boldsymbol{S}^\perp$, which can be measured geometrically by the *principal angles* between two subspaces.

Definition 1 (Knyazev & Zhu (2012)). *Let $\boldsymbol{U},\boldsymbol{V} \in \mathbb{R}^{D\times c}$ be orthonormal matrices. The principal angles between $\mathrm{Span}(\boldsymbol{U})$ and $\mathrm{Span}(\boldsymbol{V})$ are defined as $\theta_i(\boldsymbol{U},\boldsymbol{V}) = \arccos(\sigma_i(\boldsymbol{U}^\top\boldsymbol{V}))$ for all $i \in \{1,2,\dots,c\}$, where $\sigma_i(\cdot)$ denotes the i-th largest singular value. The largest principal angle $\theta_c(\boldsymbol{U},\boldsymbol{V})$ is used to define the* subspace angle *between $\mathrm{Span}(\boldsymbol{U})$ and $\mathrm{Span}(\boldsymbol{V})$.*

With Definition 1, we are able to compute how close $\mathrm{Span}(\boldsymbol{B})$ and $\mathrm{Span}(\boldsymbol{S}^\perp) = \mathcal{S}^\perp$ are to one another. In particular, when $\mathrm{Span}(\boldsymbol{B}) = \mathcal{S}^\perp$, we have $\theta_1(\boldsymbol{B},\boldsymbol{S}^\perp) = \cdots = \theta_c(\boldsymbol{B},\boldsymbol{S}^\perp) = 0$, which means that their subspace angle is zero, thus justifying the definition.

The subdifferential of $\|\boldsymbol{a}\|_2$ for any $\boldsymbol{a} \in \mathbb{R}^c$ is given by

$$\mathrm{Sgn}(\boldsymbol{a}) = \begin{cases} \{\boldsymbol{a}/\|\boldsymbol{a}\|_2\}, & \boldsymbol{a} \neq \boldsymbol{0}, \\ \{\boldsymbol{d} \in \mathbb{R}^c : \|\boldsymbol{d}\| \leq 1\}, & \boldsymbol{a} = \boldsymbol{0}. \end{cases} \tag{3}$$

An element of the set $\mathrm{Sgn}(\boldsymbol{a})$ of particular interest will be

$$\mathrm{sign}(\boldsymbol{a}) = \begin{cases} \boldsymbol{a}/\|\boldsymbol{a}\|_2, & \boldsymbol{a} \neq \boldsymbol{0}, \\ \boldsymbol{0}, & \boldsymbol{a} = \boldsymbol{0}. \end{cases} \tag{4}$$

The subdifferential of f in (2) at $\boldsymbol{B}$ is

$$\partial f(\boldsymbol{B}) = \sum_{j=1}^{L} \widetilde{\boldsymbol{x}}_j \, \mathrm{Sgn}(\widetilde{\boldsymbol{x}}_j^\top \boldsymbol{B}). \tag{5}$$

Since the optimization problem (2) is taken over the Grassmannian $\mathbb{G}(D,c)$, we associate the optimality conditions of (2) with Riemannian geometry (Edelman et al., 1998). Letting $\widetilde{\partial} f$ denote the Riemannian subdifferential of f, it follows from regularity of f and Yang et al. (2014) that $\widetilde{\partial} f(\boldsymbol{B}) = (\boldsymbol{I} - \boldsymbol{B}\boldsymbol{B}^\top)\partial f(\boldsymbol{B})$, i.e., the projection of $\partial f(\boldsymbol{B})$ onto the tangent space of $\mathbb{G}(D,c)$ at $\boldsymbol{B}$. Also, $\boldsymbol{B}$ is a critical point of problem (2) if and only if $\boldsymbol{0} \in \widetilde{\partial} f(\boldsymbol{B})$, which will be used to study global optimality for (2) in the next section.

4. Global Optimality Analysis

In this section, we analyze the non-convex non-smooth DPCP problem (2) in the noiseless setting (Section 4.1) and noisy setting (Section 4.2). Towards that end, we first define the random spherical model considered in this paper.

Definition 2 (Random spherical model). *Consider a random spherical model where the columns of $\boldsymbol{\mathcal{O}}$ are drawn uniformly from the sphere $\mathbb{S}^{D-1}$, the columns of noisy inliers $\boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{E}}$ are drawn by first independently generating inliers from $\mathcal{N}(\boldsymbol{0}, \frac{1}{d}\boldsymbol{\mathcal{P}}_{\mathcal{S}})$ and noise from $\mathcal{N}(\boldsymbol{0}, \frac{\sigma^2}{D}\boldsymbol{I}_D)$, and then projecting their sum onto $\mathbb{S}^{D-1}$, where $d = \dim(\mathcal{S})$, $\boldsymbol{\mathcal{P}}_{\mathcal{S}}$ is the ortho-projector onto $\mathcal{S}$, and $\sigma \geq 0$ controls the amount of noise present in the inliers; under this model, the SNR is $\mathbb{E}[\|\boldsymbol{\mathcal{X}}\|_F]/\,\mathbb{E}[\|\boldsymbol{\mathcal{E}}\|_F] = 1/\sigma$. In the following analysis, we always assume $\sigma < 1$.*

4.1. Noiseless Setting

Geometric quantities. We now introduce several geometric quantities that characterize the distributions of the inliers

and outliers in the noiseless setting. For inliers, we have the *permeance statistic* (Lerman et al., 2015):

$$c_{\boldsymbol{\mathcal{X}},\min} := \frac{1}{N} \min_{\boldsymbol{b}\in\mathcal{S}\cap\mathbb{S}^{D-1}} \|\boldsymbol{\mathcal{X}}^\top \boldsymbol{b}\|_1. \tag{6}$$

Well-distributed inliers result in a large value of $c_{\boldsymbol{\mathcal{X}},\min}$ due to the fact that it is difficult to find a single direction $\boldsymbol{b}$ that is orthogonal to most of the points. For outliers, we extend the $\eta_{\boldsymbol{\mathcal{O}}}$ quantity in Zhu et al. (2018) where the codimension $c = 1$, to the more general case of $c \geq 1$ by defining

$$\eta_{\boldsymbol{\mathcal{O}},c} := \frac{1}{M} \max_{\boldsymbol{B}\in\mathbb{O}(D,c)} \left\|(\mathbf{I} - \boldsymbol{B}\boldsymbol{B}^\top) \sum_{i=1}^{M} \boldsymbol{o}_i \operatorname{sign}(\boldsymbol{o}_i^\top \boldsymbol{B})\right\|_F \tag{7}$$

which is the maximum norm of a Riemannian subgradient of $\frac{1}{M}\|\boldsymbol{\mathcal{O}}^\top \boldsymbol{B}\|_{1,2}$. As an analogy to $\eta_{\boldsymbol{\mathcal{O}}}$, the $\eta_{\boldsymbol{\mathcal{O}},c}$ characterizes how well the outliers are distributed in the ambient space, with more uniformly distributed outliers leading to smaller $\eta_{\boldsymbol{\mathcal{O}},c}$. We remark that $\eta_{\boldsymbol{\mathcal{O}},c} \equiv \eta_{\boldsymbol{\mathcal{O}}}$ when $c = 1$. Besides $\eta_{\boldsymbol{\mathcal{O}},c}$, we also use another two quantities to describe the distribution of outliers, namely, we extend the $c_{\boldsymbol{\mathcal{O}},\min}$ and $c_{\boldsymbol{\mathcal{O}},\max}$ in Zhu et al. (2018) for $c = 1$ to the following: $c_{\boldsymbol{\mathcal{O}},\min,c} := \frac{1}{M} \min_{\boldsymbol{B}\in\mathbb{O}(D,c)} \sum_{j=1}^{M} \|\boldsymbol{o}_j^\top \boldsymbol{B}\|_2$, $c_{\boldsymbol{\mathcal{O}},\max,c} := \frac{1}{M} \max_{\boldsymbol{B}\in\mathbb{O}(D,c)} \sum_{j=1}^{M} \|\boldsymbol{o}_j^\top \boldsymbol{B}\|_2$. Well-distributed outliers lead to large $c_{\boldsymbol{\mathcal{O}},\min,c}$ and small $c_{\boldsymbol{\mathcal{O}},\max,c}$, and a small gap between $c_{\boldsymbol{\mathcal{O}},\max,c}$ and $c_{\boldsymbol{\mathcal{O}},\min,c}$.

To better understand the behaviors of the above geometric quantities, we provide their concentration bounds.

Lemma 1. *Consider the random spherical model in Definition 2 with $\sigma = 0$. Then, for any $t > 0$, there exists a constant C_0 independent of N, M, D, d, c and t such that*

$$\begin{aligned}
&\mathbb{P}\left[c_{\boldsymbol{\mathcal{X}},\min} \geq \sqrt{2/(\pi d)} - (2 + t/2)/\sqrt{N}\right] \geq 1 - 2e^{-\frac{t^2}{2}}, \\
&\mathbb{P}\left[\eta_{\boldsymbol{\mathcal{O}},c} \leq C_0(\sqrt{cD}\log D + t)/\sqrt{M}\right] \geq 1 - 2e^{-\frac{t^2}{2}}, \\
&\mathbb{P}\left[c_{\boldsymbol{\mathcal{O}},\max,c} - c_{\boldsymbol{\mathcal{O}},\min,c} \leq (4\sqrt{2c} + t)/\sqrt{M}\right] \geq 1 - 2e^{-\frac{t^2}{2}}.
\end{aligned} \tag{8}$$

One can see that $c_{\boldsymbol{\mathcal{X}},\min}$ scales as $O(1)$ while both $\eta_{\boldsymbol{\mathcal{O}},c}$ and $c_{\boldsymbol{\mathcal{O}},\max,c} - c_{\boldsymbol{\mathcal{O}},\min,c}$ scale as $O(1/\sqrt{M})$. Moreover, the role of c can be seen clearly from (8): both $\eta_{\boldsymbol{\mathcal{O}},c}$ and $c_{\boldsymbol{\mathcal{O}},\max,c} - c_{\boldsymbol{\mathcal{O}},\min,c}$ tend to be larger as c increases.

Using the above geometric quantities, we have the following lemma, which states the geometry of the critical points of (2) in a deterministic sense.

Lemma 2. *Suppose $\boldsymbol{\mathcal{E}} = \mathbf{0}$. Then, any critical point $\boldsymbol{B}^*$ of (2) must either be an orthonormal basis for $\mathcal{S}^\perp$, or span a subspace that has an angle from $\mathcal{S}^\perp$ larger than or equal to $\theta^\diamond := \arccos(M\overline{\eta}_{\boldsymbol{\mathcal{O}},c}/Nc_{\boldsymbol{\mathcal{X}},\min})$ where $\overline{\eta}_{\boldsymbol{\mathcal{O}},c} := \eta_{\boldsymbol{\mathcal{O}},c} + \frac{D}{M}$.*

Lemma 2 generalizes the special case $c = 1$ in Zhu et al. (2018, Lemma 1). It says that, with noiseless data, any critical point of (2) either spans $\mathcal{S}^\perp$ or spans a subspace that is far from $\mathcal{S}^\perp$. Note that for well-distributed inliers and outliers (M/N and c fixed), the geometric location of $\boldsymbol{B}^*$ becomes more restricted. Observe that any critical point $\boldsymbol{B}^*$ such that $\operatorname{Span}(\boldsymbol{B}^*)$ is sufficiently close to $\mathcal{S}^\perp$ (angle smaller than $\theta^\diamond$) must satisfy $\operatorname{Span}(\boldsymbol{B}^*) = \mathcal{S}^\perp$. This motivates the next result on the geometry of global minimizers.

Theorem 1. *Suppose $\boldsymbol{\mathcal{E}} = \mathbf{0}$. Then, any global solution $\boldsymbol{B}^*$ to (2) must be an orthonormal basis for $\mathcal{S}^\perp$ as long as*

$$\frac{M}{N} \cdot \frac{\sqrt{\overline{\eta}_{\boldsymbol{\mathcal{O}},c}^2 + (c_{\boldsymbol{\mathcal{O}},\max,c} - c_{\boldsymbol{\mathcal{O}},\min,c})^2}}{c_{\boldsymbol{\mathcal{X}},\min}} < 1. \tag{9}$$

Theorem 1 is an extension of Zhu et al. (2018, Theorem 1) for the hyperplane case. First note that $c_{\boldsymbol{\mathcal{O}},\max,c} - c_{\boldsymbol{\mathcal{O}},\min,c} \to 0$ as $M \to \infty$ according to (8). Then (9) tells us that, with fixed M/N and c, as long as we have more and more data points that are well-distributed, (9) will be satisfied and thus any global solution to (2) spans $\mathcal{S}^\perp$.[1] Furthermore, combining the global optimality condition (9) and the concentration bounds in (8), one can derive the following probabilistic result that characterizes global optimality with noiseless data in a more interpretable way.

Theorem 2. *Consider the random spherical model in Definition 2 with $\sigma = 0$. Fix any $0 < t < 2\left(\sqrt{\frac{2N}{\pi d}} - 2\right)$. With probability at least $1 - 6e^{-t^2/2}$, any global solution $\boldsymbol{B}^* \in \mathbb{R}^{D\times c}$ to (2) must be an orthonormal basis for $\mathcal{S}^\perp$ if*

$$\begin{aligned}
&M\left((4\sqrt{c} + t)^2 + C_0(\sqrt{cD}\log D + t)^2\right) \\
&\leq N^2\left(\sqrt{2/(\pi d)} - (2 + t/2)/\sqrt{N}\right)^2,
\end{aligned} \tag{10}$$

where C_0 is a universal constant that is independent of N, M, D, d, c and t.

Condition (10) interprets the global optimality condition (9) of Theorem 1 with natural quantities such as N, M, D, d and c. It validates that the new formulation (2) of DPCP on the Grassmannian $\mathbb{G}(D, c)$ is still able to tolerate $O(N^2)$ outliers for recovering the entire orthonormal basis of $\mathcal{S}^\perp$. Also, note that for fixed N, M, D, and d, the smaller c becomes, the easier it is for condition (10) to be satisfied. Particularly, in the hyperplane case $c = 1$, Theorem 2 reduces to the result of Zhu et al. (2018, Theorem 2).

4.2. Noisy Setting

We now consider the scenario when inliers $\boldsymbol{\mathcal{X}}$ are further contaminated with noise, i.e., $\sigma > 0$ and $\boldsymbol{\mathcal{E}} \neq \mathbf{0}$ in Definition 2. We decompose the noise term as $\boldsymbol{\mathcal{E}} = \boldsymbol{\mathcal{E}}_{\boldsymbol{s}} + \boldsymbol{\mathcal{E}}_{\boldsymbol{n}}$,

[1] A similar theorem appears in Ding et al. (2020, Proposition 3), although they analyze a group-DPCP formulation different from (2) designed specifically for homography estimation.

where $\boldsymbol{\mathcal{E}_s}$ is the projection of $\boldsymbol{\mathcal{E}}$ onto $\mathcal{S}$ and $\boldsymbol{\mathcal{E}_n}$ is the projection onto $\mathcal{S}^\perp$. Observe that the term $\boldsymbol{\mathcal{E}_s}$ plays the same role as inliers since its columns lie exactly in $\mathcal{S}$, and that the component $\boldsymbol{\mathcal{E}_n}$ is the effective noise that influences the global solution to (2), making it different from the noiseless case. With this in mind, as in Ding et al. (2019), we separate them out and denote $\widehat{\boldsymbol{\mathcal{X}}} := \boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{E}_s}$ with $\operatorname{Span}(\widehat{\boldsymbol{\mathcal{X}}}) \subset \mathcal{S}$ and $\widehat{\boldsymbol{\mathcal{E}}} := \boldsymbol{\mathcal{E}_n}$ with $\operatorname{Span}(\widehat{\boldsymbol{\mathcal{E}}}) \subset \mathcal{S}^\perp$. Obviously, we have $\boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{E}} = \widehat{\boldsymbol{\mathcal{X}}} + \widehat{\boldsymbol{\mathcal{E}}}$, and we can rewrite the objective in (2) as

$$f(\boldsymbol{B}) = \sum_{j=1}^{N} \|(\widehat{\boldsymbol{x}}_j + \widehat{\boldsymbol{\epsilon}}_j)^\top \boldsymbol{B}\|_2 + \sum_{j=1}^{M} \|\boldsymbol{o}_j^\top \boldsymbol{B}\|_2, \quad (11)$$

with $\widehat{\boldsymbol{x}}_j$ and $\widehat{\boldsymbol{\epsilon}}_j$ the j-th column of $\widehat{\boldsymbol{\mathcal{X}}}$ and $\widehat{\boldsymbol{\mathcal{E}}}$, respectively.

Geometric quantities. First note that the previous quantities related to outliers, i.e., $\eta_{\boldsymbol{\mathcal{O}},c}$, $c_{\boldsymbol{\mathcal{O}},\max,c}$ and $c_{\boldsymbol{\mathcal{O}},\min,c}$, remain the same. For noisy inliers, since we have separated out the effective noise, we have the following two extra quantities with respect to $\widehat{\boldsymbol{\mathcal{X}}}$ and $\widehat{\boldsymbol{\mathcal{E}}}$:

$$\begin{aligned} c_{\widehat{\boldsymbol{\mathcal{X}}},\min} &:= \frac{1}{N} \min_{\boldsymbol{b} \in \mathcal{S} \cap \mathbb{S}^{D-1}} \|\widehat{\boldsymbol{\mathcal{X}}}^\top \boldsymbol{b}\|_1, \\ c_{\widehat{\boldsymbol{\mathcal{E}}},\max,c} &:= \frac{1}{N} \max_{\boldsymbol{B} \in \mathbb{O}(D,c)} \sum_{j=1}^{N} \|\widehat{\boldsymbol{\epsilon}}_j^\top \boldsymbol{B}\|_2. \end{aligned} \quad (12)$$

Note that $c_{\widehat{\boldsymbol{\mathcal{X}}},\min}$ is analogous to $c_{\boldsymbol{\mathcal{X}},\min}$ in (6) by replacing $\boldsymbol{\mathcal{X}}$ with $\widehat{\boldsymbol{\mathcal{X}}}$: the more well-distributed $\widehat{\boldsymbol{\mathcal{X}}}$ is, the larger $c_{\widehat{\boldsymbol{\mathcal{X}}},\min}$ becomes. The quantity $c_{\widehat{\boldsymbol{\mathcal{E}}},\max,c}$ generalizes $c_{\widehat{\boldsymbol{\mathcal{E}}},\max}$ defined in Ding et al. (2019) for $c = 1$, and quantifies the effective noise level. Note that $c_{\widehat{\boldsymbol{\mathcal{E}}},\max,c} \le \frac{1}{N}\sum_{j=1}^{N} \|\widehat{\boldsymbol{\epsilon}}_j\|_2$, which is the *total inlier residual* used in Lerman et al. (2015), but $c_{\widehat{\boldsymbol{\mathcal{E}}},\max,c}$ also considers the geometry of the effective noise.

We now give concentration bounds for $c_{\widehat{\boldsymbol{\mathcal{X}}},\min}$ and $c_{\widehat{\boldsymbol{\mathcal{E}}},\max,c}$ when $\sigma \in (0,1)$. To estimate their expectations, one has

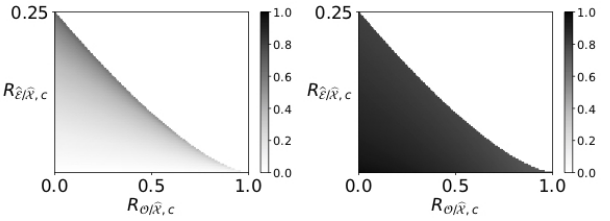
$$\mathbb{E}\left[|\widehat{\boldsymbol{x}}_j^\top \boldsymbol{b}|\right] \ge \sqrt{2/(\pi d)}\rho(\sigma), \quad \mathbb{E}\left[\|\widehat{\boldsymbol{\epsilon}}_j^\top \boldsymbol{B}\|_2\right] \le \delta(\sigma) \quad (13)$$

where $\rho(\sigma) := (1-\sigma)F_{D-d,d}(1/\sigma)$, $\delta(\sigma) := \sqrt{\sigma} + \sqrt{(1-\sigma)F_{d,D-d}(\sigma)}$, and $F_{d_1,d_2}(\cdot)$ is the cumulative density function of the F-distribution with $F_{d_1,d_2}(0) = 0$ and $F_{d_1,d_2}(\infty) = 1$. It has been shown in Ding et al. (2019) that $\delta(\sigma) = O(\sigma^{d/4} + \sqrt{\sigma})$ and $\rho(\sigma) = 1 - O(\sigma + \sigma^{d/2})$.

Lemma 3. *Consider the random spherical model defined in Definition 2 with $\sigma \in (0,1)$. Then for any $t > 0$, we have*

$$\begin{aligned} &\mathbb{P}\left[c_{\widehat{\boldsymbol{\mathcal{X}}},\min} \ge \sqrt{2/(\pi d)}\rho(\sigma) - \tfrac{2+t/2}{\sqrt{N}}\right] \ge 1 - 2e^{-\frac{t^2}{2}}, \\ &\mathbb{P}\left[c_{\widehat{\boldsymbol{\mathcal{E}}},\max,c} \le (1 + \tfrac{2\sqrt{2c}}{\sqrt{N}})\delta(\sigma) + \tfrac{t}{\sqrt{N}}\right] \ge 1 - 2e^{-\frac{t^2}{2}}. \end{aligned} \quad (14)$$

As $\sigma \to 0$, we know that $\rho(\sigma) \to 1$ and $\delta(\sigma) \to 0$. In particular, when $\sigma = 0$ ($\boldsymbol{\mathcal{E}} = \boldsymbol{0}$), the above result for $c_{\widehat{\boldsymbol{\mathcal{X}}},\min}$ is the same as that of $c_{\boldsymbol{\mathcal{X}},\min}$ in (8), but for $c_{\widehat{\boldsymbol{\mathcal{E}}},\max,c}$ it does not immediately imply $c_{\widehat{\boldsymbol{\mathcal{E}}},\max,c} = 0$ due to the existence of the term $t/\sqrt{N}$ (usually very small since N is very large compared with t), which is an artifact of the proof.[2]



(a) Value of t_1 (b) Value of t_2

Figure 1. Plot of t_1 and t_2 in Lemma 4 given $(R_{\boldsymbol{\mathcal{O}}/\widehat{\boldsymbol{\mathcal{X}}},c}, R_{\widehat{\boldsymbol{\mathcal{E}}}/\widehat{\boldsymbol{\mathcal{X}}},c})$ pair such that condition (16) holds true (area below the curve).

To simplify the presentation of the remaining analysis, let

$$R_{\boldsymbol{\mathcal{O}}/\widehat{\boldsymbol{\mathcal{X}}},c} := \frac{M}{N}\frac{\overline{\eta}_{\boldsymbol{\mathcal{O}},c}}{c_{\widehat{\boldsymbol{\mathcal{X}}},\min}}, \quad R_{\widehat{\boldsymbol{\mathcal{E}}}/\widehat{\boldsymbol{\mathcal{X}}},c} := \frac{c_{\widehat{\boldsymbol{\mathcal{E}}},\max,c}}{c_{\widehat{\boldsymbol{\mathcal{X}}},\min}}, \quad (15)$$

which can be viewed as outlier-to-inlier and noise-to-inlier type of ratios (Ding et al., 2019), respectively. Now we are ready to characterize the distribution of the critical points of (2) when the dataset is also contaminated with noise.

Lemma 4. *Assume $R_{\boldsymbol{\mathcal{O}}/\widehat{\boldsymbol{\mathcal{X}}},c} < 1$ and*

$$\begin{aligned} R_{\widehat{\boldsymbol{\mathcal{E}}}/\widehat{\boldsymbol{\mathcal{X}}},c} < &\frac{1}{32}\left(\sqrt{R_{\boldsymbol{\mathcal{O}}/\widehat{\boldsymbol{\mathcal{X}}},c}^2 + 8} - 3R_{\boldsymbol{\mathcal{O}}/\widehat{\boldsymbol{\mathcal{X}}},c}\right)^{\frac{3}{2}} \\ &\cdot \left(\sqrt{R_{\boldsymbol{\mathcal{O}}/\widehat{\boldsymbol{\mathcal{X}}},c}^2 + 8} + R_{\boldsymbol{\mathcal{O}}/\widehat{\boldsymbol{\mathcal{X}}},c}\right)^{\frac{1}{2}}. \end{aligned} \quad (16)$$

Any critical point $\boldsymbol{B}^*$ of problem (2) spans a subspace that has an angle θ_c^* from $\mathcal{S}^\perp$ satisfying

$$\theta_c^* \le \sin^{-1}(t_1) \quad \text{or} \quad \theta_c^* \ge \sin^{-1}(t_2) \quad (17)$$

where $0 \le t_1 \le t_2 \le 1$ with

$$t_2 := \sqrt{1 - \frac{1}{4}\left(R_{\boldsymbol{\mathcal{O}}/\widehat{\boldsymbol{\mathcal{X}}},c} + \sqrt{R_{\boldsymbol{\mathcal{O}}/\widehat{\boldsymbol{\mathcal{X}}},c}^2 + 8R_{\widehat{\boldsymbol{\mathcal{E}}}/\widehat{\boldsymbol{\mathcal{X}}},c}}\right)^2}, \quad (18)$$

and t_1 being the smallest nonnegative root of

$$t^4 + (R_{\boldsymbol{\mathcal{O}}/\widehat{\boldsymbol{\mathcal{X}}},c}^2 - 1)t^2 + 4R_{\boldsymbol{\mathcal{O}}/\widehat{\boldsymbol{\mathcal{X}}},c}R_{\widehat{\boldsymbol{\mathcal{E}}}/\widehat{\boldsymbol{\mathcal{X}}},c}t + 4R_{\widehat{\boldsymbol{\mathcal{E}}}/\widehat{\boldsymbol{\mathcal{X}}},c}^2 = 0. \quad (19)$$

The feasible region for $(R_{\boldsymbol{\mathcal{O}}/\widehat{\boldsymbol{\mathcal{X}}},c}, R_{\widehat{\boldsymbol{\mathcal{E}}}/\widehat{\boldsymbol{\mathcal{X}}},c})$ with condition (16) satisfied is shown as the area under the curve in Figure 1, which implies that the outlier-to-inlier ratio and the noise-to-inlier ratio cannot be very large at the same time. In other words, larger noise levels restrict the number of

[2] We also provide another concentration bound for $c_{\widehat{\boldsymbol{\mathcal{E}}},\max,c}$ in the supplemental material that is completely proportional to $\delta(\sigma)$.

outliers that the DPCP problem (2) can tolerate. Next, one can show that the quartic equation (19) must have two nonnegative roots (with t_1 the smaller one), and condition (16) ensures that $t_1 \leq t_2$. Then, (17) indicates that any critical point $\boldsymbol{B}^*$ of the noisy problem (2) spans a subspace that is close to either $\mathcal{S}^\perp$ or $\mathcal{S}$. Figure 1 provides a better understanding of t_1 and t_2: with smaller outlier-to-inlier ratio and noise-to-inlier ratio, t_1 is closer to 0 (lighter) and t_2 is closer to 1 (darker), making the geometric location of $\boldsymbol{B}^*$ more restricted. Compared with Lemma 2 in the noiseless case where $\boldsymbol{B}^*$ is an exact orthonormal basis of $\mathcal{S}^\perp$ if it is sufficiently far from $\mathcal{S}$, here we can only guarantee that it lies in a neighborhood of $\mathcal{S}^\perp$, i.e., $\theta_c^* \leq \sin^{-1}(t_1)$, due to the noise. One can further bound t_1 (Ding et al., 2019) by

$$t_1 \leq 25 R_{\widehat{\boldsymbol{\mathcal{E}}}/\widehat{\boldsymbol{\mathcal{X}}},c}/(1 - R_{\boldsymbol{\mathcal{O}}/\widehat{\boldsymbol{\mathcal{X}}},c})^2. \tag{20}$$

When there is no noise, from (20) we have $t_1 = 0$, and from (18) we have $t_2 = \sqrt{1 - R^2_{\boldsymbol{\mathcal{O}}/\widehat{\boldsymbol{\mathcal{X}}},c}}$, which is consistent with Lemma 2. Moreover, (20) shows that t_1 is small with small outlier-to-inlier ratio and noise-to-inlier ratio, and is proportional to the effective noise level. Finally, compared with the critical point analysis for noisy problem (1) with $c = 1$ in Ding et al. (2019), the proof technique used here is different since problem (2) is defined over the Grassmannian, which requires consideration of the geometry of subspaces in $\mathbb{G}(D, c)$. In particular, in Ding et al. (2019) both t_1 and t_2 are defined by the nonnegative roots of (19), while in this generalized analysis t_2 is decoupled from (19) (see (18)).

Using Lemma 4, we may now characterize the global solution of the noisy DPCP problem (2).

Theorem 3. *If $R_{\boldsymbol{\mathcal{O}}/\widehat{\boldsymbol{\mathcal{X}}},c} < 1$ and (16) holds, and*

$$R^2_{\boldsymbol{\mathcal{O}}/\widehat{\boldsymbol{\mathcal{X}}},c} + \Big(\frac{M}{N}\frac{c_{\boldsymbol{\mathcal{O}},\max,c} - c_{\boldsymbol{\mathcal{O}},\min,c}}{c_{\widehat{\boldsymbol{\mathcal{X}}},\min}} + 2R_{\widehat{\boldsymbol{\mathcal{E}}}/\widehat{\boldsymbol{\mathcal{X}}},c}\Big)^2 + 8R_{\widehat{\boldsymbol{\mathcal{E}}}/\widehat{\boldsymbol{\mathcal{X}}},c} < 1, \tag{21}$$

then any global solution $\boldsymbol{B}^*$ of (2) must span a subspace that has an angle θ_c^* from $\mathcal{S}^\perp$ satisfying $\theta_c^* \leq \sin^{-1}(t_1)$, where $0 \leq t_1 \leq 1$ is the smallest nonnegative root of (19).

Condition (21) is sufficient to ensure that global solutions of (2) span a subspace that is close to $\mathcal{S}^\perp$. We interpret (21) as follows: with fixed M/N, as data points are increasing ($c_{\boldsymbol{\mathcal{O}},\max,c} - c_{\boldsymbol{\mathcal{O}},\min,c} \to 0$) and well-distributed (large $c_{\widehat{\boldsymbol{\mathcal{X}}},\min}$, small $R_{\boldsymbol{\mathcal{O}}/\widehat{\boldsymbol{\mathcal{X}}},c}$), and the effective noise is mild (small $R_{\widehat{\boldsymbol{\mathcal{E}}}/\widehat{\boldsymbol{\mathcal{X}}},c}$), (21) will be satisfied and global solutions of (2) must be close to $\mathcal{S}^\perp$. Note that in the noiseless case condition (21) is equivalent to condition (9) and $t_1 = 0$, which means Theorem 3 is precisely Theorem 1. Next, we give its probabilistic characterization.

Theorem 4. *Consider the random spherical model in Definition 2. Assume $N > c$. Then for any positive $t < 2\big(\sqrt{\frac{2N}{\pi d}}\rho(\sigma) - 2\big)$, any global solution $\boldsymbol{B}^*$ of (2) must span a subspace that has an angle θ_c^* from $\mathcal{S}^\perp$ satisfying*

$$\sin(\theta_c^*) \leq \frac{C_1\delta(\sigma) + \frac{25t}{\sqrt{N}}}{\sqrt{\frac{2}{\pi d}}\rho(\sigma) - C_2\frac{t\sqrt{M}+\sqrt{cDM}\log D}{N} - \frac{4+t}{2\sqrt{N}}} \tag{22}$$

with probability at least $1 - 10e^{-t^2/2}$, as long as

$$M\left((8\sqrt{2c} + 2t)^2 + C_3(\sqrt{cD}\log D + t)^2\right) \leq N^2\left[\left(\sqrt{\frac{2}{\pi d}}\rho(\sigma) - \frac{4+t}{2\sqrt{N}}\right)^2 - C_4\delta(\sigma) - \frac{16t^2}{N} - \frac{8t}{\sqrt{dN}}\right], \tag{23}$$

where C_1, C_2, C_3, C_4 are universal constants that are independent of N, M, D, d, c, t and σ.

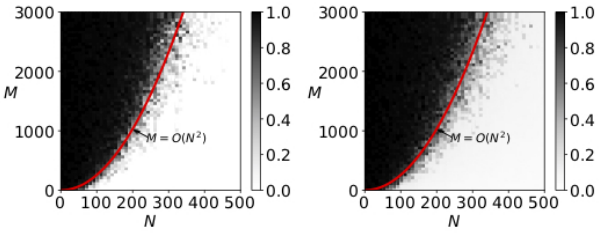


Figure 2. Plot of the subspace angle between $\mathrm{Span}(\boldsymbol{B}^*)$ and $\mathcal{S}^\perp$ with $\boldsymbol{B}^*$ obtained from Algorithm 1 for (Left) noiseless case $\sigma = 0$ and (Right) noisy case $\sigma = 0.1$. Here we fix $D = 30$ and $c = 5$.

Towards interpreting Theorem 4, first recall that $\delta(\sigma) \to 0$ and $\rho(\sigma) \to 1$ as $\sigma \to 0$. Then, (22) indicates that the angle θ_c^* between $\mathcal{S}^\perp$ and the subspace spanned by a global solution $\boldsymbol{B}^*$ of (2) is close[3] to zero as $\sigma \to 0$, and $\sin(\theta_c^*) = O(\sigma^{d/4} + \sqrt{\sigma})$ which is on the same order of $\delta(\sigma)$. Furthermore, the sufficient condition (23) implies that problem (2) can also tolerate $O(N^2)$ outliers for learning the entire orthonormal basis for $\mathcal{S}^\perp$ with noisy data, as illustrated in Figure 2. Finally, we remark that condition (23) does not necessarily have the same form of condition (10) when $\sigma = 0$ or the condition in Ding et al. (2019, Theorem 2) when $c = 1$ because the technical proof details are different; however, they all reveal that the DPCP problems (both (1) and (2)) can roughly handle $O(\frac{1}{cdD\log^2 D}N^2)$ outliers, which is an apparent advantage over other RSR methods that can only deal with $O(N)$ outliers in theory.

Note that we focus on learning a subspace of high relative dimension, where $d/D \approx 1$ and the codimension $c = D - d$ is very small. As a result, the theoretical guarantees derived in this section are well-suited for the regime of $c = O(1)$. In particular, when $c = 1$, the DPCP problem reduces to

[3] Note that the numerator of (22) has a small term $\frac{25t}{\sqrt{N}}$ due to the proof artifact of the concentration bound on $c_{\widehat{\boldsymbol{\mathcal{E}}},\max,c}$, which can be improved to be proportional to $\delta(\sigma)$ (see Footnote 2).

Algorithm 1 Projected Riemannian Subgradient Method

Initialization: $\boldsymbol{B}_0 \in \mathbb{O}(D,c)$ and $\mu_0 \in (0,1)$;

1: **for** $k = 0, 1, \ldots$ **do**
2: Compute a Riemannian subgradient: $\mathcal{G}(\boldsymbol{B}_k) = (\mathbf{I} - \boldsymbol{B}_k\boldsymbol{B}_k^\top)\big(\sum_{j=1}^{L} \widetilde{\boldsymbol{x}}_j \operatorname{sign}(\widetilde{\boldsymbol{x}}_j^\top \boldsymbol{B}_k)\big)$;
3: Compute the step size μ_k according to a certain rule;
4: Update the iterate: $\widehat{\boldsymbol{B}}_{k+1} \leftarrow \boldsymbol{B}_k - \mu_k \mathcal{G}(\boldsymbol{B}_k)$, $\boldsymbol{B}_{k+1} \leftarrow \operatorname{orth}(\widehat{\boldsymbol{B}}_{k+1})$;
5: **end for**

an optimization problem over the sphere. In general, the bound $M = O(\frac{1}{cdD\log^2 D}N^2)$ indicates that for very large c, e.g., $c = O(D)$, the DPCP approach can only handle a small number of outliers. In fact, since the subspace is now low-dimensional, methods designed for low-dimensional subspaces are more appropriate.

5. Convergence Analysis of a Projected Riemannian Subgradient Method

In this section, we study the projected Riemannian Subgradient Method (RSGM) given by Algorithm 1 ($\operatorname{orth}(\boldsymbol{A})$ denotes an orthonormal basis for $\operatorname{Span}(\boldsymbol{A})$) for solving the DPCP problem (2) over the Grassmannian $\mathbb{G}(D,c)$. It has been shown in Zhu et al. (2019) that the RSGM applied to (2) with *noiseless* data converges linearly to an orthonormal basis, say $\boldsymbol{S}^\perp$, of $\mathcal{S}^\perp$. However, the analytical result cannot be immediately generalized to the noisy case, in which one can only expect that it at best converges to a neighborhood of $\boldsymbol{S}^\perp$ as suggested by the noisy analyses in Section 4.2. Note that the convergence analysis of RSGM for problem (2) with noiseless data is built upon a Riemannian Regularity Condition (RRC) (Zhu et al., 2019), which is a local geometric property of problem (2) relative to a point of interest, e.g., $\boldsymbol{S}^\perp$ in our case. We will show that when data is corrupted by noise, the RRC for (2) only holds outside a neighborhood of $\boldsymbol{S}^\perp$ with a radius proportional to the effective noise level, which is then used to show that the RSGM converges linearly to that neighborhood of $\boldsymbol{S}^\perp$.

Define the distance between any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{O}(D,c)$ as

$$\operatorname{dist}(\boldsymbol{A},\boldsymbol{B}) := \min_{\boldsymbol{Q}\in\mathbb{O}(c)} \|\boldsymbol{B} - \boldsymbol{A}\boldsymbol{Q}\|_F. \tag{24}$$

It follows Higham & Papadimitriou (1995) that the optimum value is $\sqrt{2\sum_{i=1}^{c}(1-\cos(\theta_i(\boldsymbol{A},\boldsymbol{B})))}$ since the optimal rotation matrix $\boldsymbol{Q}$ for (24) is $\boldsymbol{Q}^* = \boldsymbol{U}\boldsymbol{V}^\top$, where $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is the SVD of $\boldsymbol{A}^\top\boldsymbol{B}$. Then we define the projection of $\boldsymbol{B}$ onto $[\boldsymbol{A}]$ as $\mathcal{P}_{\boldsymbol{A}}(\boldsymbol{B}) = \boldsymbol{A}\boldsymbol{Q}^*$, where $\boldsymbol{Q}^* = \arg\min_{\boldsymbol{Q}\in\mathbb{O}(c)} \|\boldsymbol{B} - \boldsymbol{A}\boldsymbol{Q}\|_F$.

Proposition 1. *The definition (24) of* $\operatorname{dist}(\boldsymbol{A},\boldsymbol{B})$ *is equivalent to the subspace angle* $\theta_c(\boldsymbol{A},\boldsymbol{B})$ *in measuring the similarity between* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *in the following sense:*

$$\sin(\theta_c(\boldsymbol{A},\boldsymbol{B})) \le \operatorname{dist}(\boldsymbol{A},\boldsymbol{B}) \le \sqrt{2c}\cdot\sin(\theta_c(\boldsymbol{A},\boldsymbol{B})). \tag{25}$$

Proposition 1 implies that $\theta_c(\boldsymbol{A},\boldsymbol{B})$ and $\operatorname{dist}(\boldsymbol{A},\boldsymbol{B})$ are equivalent in characterizing how close $\boldsymbol{A}$ and $\boldsymbol{B}$ are to each other, while the latter is convenient for our convergence analysis. Letting $\boldsymbol{S}^\perp$ be an orthonormal basis for $\mathcal{S}^\perp$, we show that the DPCP problem (2) satisfies a particular RRC in a ring-like neighborhood of $\boldsymbol{S}^\perp$.

Lemma 5 ($(\alpha,\tau,\boldsymbol{S}^\perp)$-RRC)**.** *For any* $\tau > 0$ *satisfying*

$$\tau(1 - R_{\boldsymbol{\mathcal{O}}/\widehat{\boldsymbol{\mathcal{X}}},c} - \tau^2/2) \ge 4\sqrt{2c}R_{\widehat{\boldsymbol{\mathcal{E}}}/\widehat{\boldsymbol{\mathcal{X}}},c}, \tag{26}$$

let $\alpha := Nc_{\widetilde{\boldsymbol{\mathcal{X}}},\min}((1-\tau^2/2) - R_{\boldsymbol{\mathcal{O}}/\widehat{\boldsymbol{\mathcal{X}}},c})/(2\sqrt{2c})$. *Then for any* $\boldsymbol{B} \in \mathbb{O}(D,c)$ *satisfying*

$$\tau \ge \operatorname{dist}(\boldsymbol{B},\boldsymbol{S}^\perp) \ge \omega := (2/\alpha)Nc_{\widehat{\boldsymbol{\mathcal{E}}},\max,c}, \tag{27}$$

there exists a Riemannian subgradient $\mathcal{G}(\boldsymbol{B}) \in \widetilde{\partial} f(\boldsymbol{B})$ *that*

$$\langle -\mathcal{G}(\boldsymbol{B}), \mathcal{P}_{\boldsymbol{S}^\perp}(\boldsymbol{B}) - \boldsymbol{B}\rangle \ge \alpha \operatorname{dist}(\boldsymbol{B},\boldsymbol{S}^\perp). \tag{28}$$

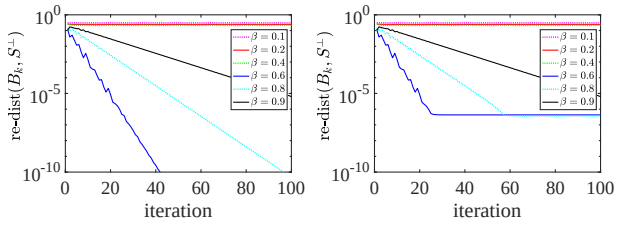
Also, for any $\boldsymbol{B} \in \mathbb{O}(D,c)$*, we have*

$$\|\mathcal{G}(\boldsymbol{B})\|_F \le \xi := \sqrt{N}\|\boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{E}}\|_2 + M\eta_{\boldsymbol{\mathcal{O}},c}. \tag{29}$$

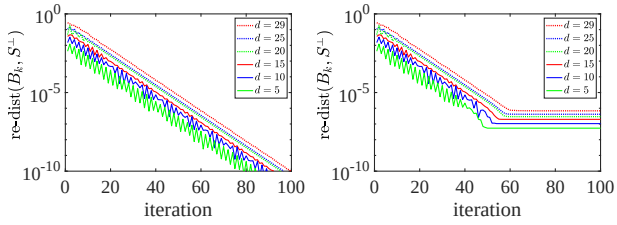
First, condition (27) specifies both an upper bound and a lower bound that $\operatorname{dist}(\boldsymbol{B},\boldsymbol{S}^\perp)$ needs to satisfy: the upper bound τ indicates that the RRC is a local geometric property around $\boldsymbol{S}^\perp$, while the lower bound ω implies the RRC may not hold within a small radius of $\boldsymbol{S}^\perp$ due to the existence of noise. Note that the lower bound ω for $\operatorname{dist}(\boldsymbol{B},\boldsymbol{S}^\perp)$ leads to a region around $\mathcal{S}^\perp$ inside which the RRC is not guaranteed and its radius ω is proportional to the effective noise level (vanishing as $\boldsymbol{\mathcal{E}} \to \mathbf{0}$), making the entire lemma reduce to the noiseless case as stated in Zhu et al. (2019). We remark that (26) gives a valid range for τ and thus ensures the validity of (27). Given $\operatorname{dist}(\boldsymbol{B},\boldsymbol{S}^\perp) \in [\omega,\tau]$, the RRC condition (28) states that a negative Riemannian subgradient $-\mathcal{G}(\boldsymbol{B})$ has a small angle with the direction pointing towards $\boldsymbol{S}^\perp$ at $\boldsymbol{B}$. By Cauchy-Schwarz inequality $\langle \mathcal{G}(\boldsymbol{B}), \boldsymbol{B} - \mathcal{P}_{\boldsymbol{S}^\perp}(\boldsymbol{B})\rangle \le \|\mathcal{G}(\boldsymbol{B})\|_F \operatorname{dist}(\boldsymbol{B},\boldsymbol{S}^\perp)$, and the RRC condition gives $\|\mathcal{G}(\boldsymbol{B})\|_F \ge \alpha$, which implies $\xi \ge \alpha$.

With the RRC for problem (2) stated in Lemma 5, we provide a convergence analysis for RSGM (Algorithm 1) with two different strategies of updating the step size: constant step size and geometrically diminishing step size.

Proposition 2. *Let* α,τ,ω,ξ *be defined in Lemma 5. Suppose the initialization* $\boldsymbol{B}_0$ *satisfies* $\operatorname{dist}(\boldsymbol{B}_0,\boldsymbol{S}^\perp) \le \tau$*, and let* $\{\boldsymbol{B}_k\}$ *be the iterates generated with constant step size* $\mu_k \equiv \mu$ *satisfying* $\mu \le \alpha(\tau-\omega)/\xi^2$*. Then* $\operatorname{dist}(\boldsymbol{B}_k,\boldsymbol{S}^\perp) \le \max\big\{\operatorname{dist}(\boldsymbol{B}_0,\boldsymbol{S}^\perp) - \frac{k\alpha\mu}{2}, \frac{\mu\xi^2}{\alpha} + \omega\big\}$.

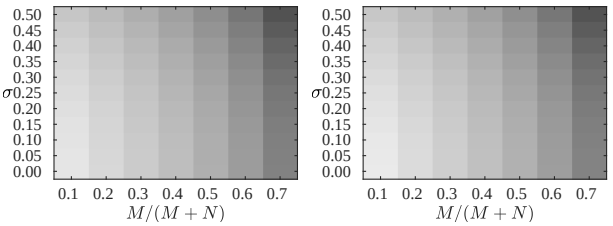


(a) $d = 25, \sigma = 0$ (b) $d = 25, \sigma = 10^{-6}$

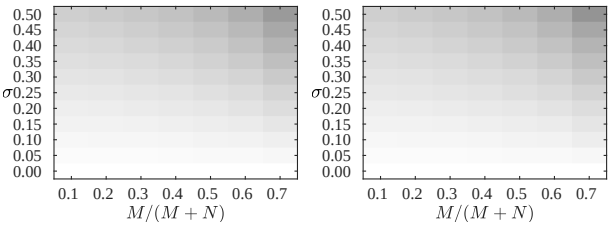


(c) $\beta = 0.8, \sigma = 0$ (d) $\beta = 0.8, \sigma = 10^{-6}$

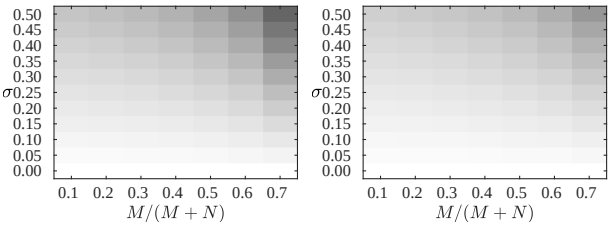
Figure 3. Convergence of Algorithm 1 for the noisy DPCP problem (2). For all the cases, we fix $D = 30, N = 500$ and outlier ratio $M/(M+N) = 0.7$. We use a spectral initialization and compute μ_0 by a backtracking line search method. The relative distance re-dist$(\boldsymbol{B}_k, \boldsymbol{S}^\perp)$ is defined by $\text{dist}(\boldsymbol{B}_k, \boldsymbol{S}^\perp)/\sqrt{c}$.



(a) PCA (0.65s) (b) REAPER (5.39s)



(c) GGD (24.78s) (d) GGD-dual (2.53s)



(e) DPCP-recursive (79.74s) (f) DPCP-holistic (1.78s)

Figure 4. Phase transition of the distance between the ground-truth basis for the (dual) subspace and the computed basis by different methods when varying the outlier ratio $M/(M+N)$ and σ. The lighter the color, the smaller the distance. The mean running time for each method is also recorded. Here we fix $D = 1000, c = 50, N = 10D$, and the results are averaged over 100 experiments.

Proposition 2 shows that with a constant step size, Algorithm 1 ensures convergence to a neighborhood of $\boldsymbol{S}^\perp$ if properly initialized. If $\text{dist}(\boldsymbol{B}_0, \boldsymbol{S}^\perp) > \mu\xi^2/\alpha + \omega$, then $\{\boldsymbol{B}_k\}$ will get closer to $\boldsymbol{S}^\perp$ until the iterates enter the region where $\text{dist}(\boldsymbol{B}_k, \boldsymbol{S}^\perp) \le \mu\xi^2/\alpha + \omega$, after which no further decay is guaranteed. Also, a larger step size μ results in faster convergence of $\boldsymbol{B}_k$ to a larger neighborhood of $\boldsymbol{S}^\perp$.

We now consider diminishing step sizes.

Theorem 5. *Consider α, τ, ω and ξ defined in Lemma 5. Suppose the initialization $\boldsymbol{B}_0$ of Algorithm 1 satisfies $\text{dist}(\boldsymbol{B}_0, \boldsymbol{S}^\perp) \le \tau$, and let $\{\boldsymbol{B}_k\}$ be the iterates generated with step size $\mu_k = \mu_0\beta^k$ satisfying*

$$\mu_0 \le (\alpha/\xi^2)\min\left\{\text{dist}(\boldsymbol{B}_0, \boldsymbol{S}^\perp)/2, \tau - \omega\right\} \quad \textit{and}$$
$$\sqrt{1 - 2\frac{\alpha\mu_0}{\text{dist}(\boldsymbol{B}_0, \boldsymbol{S}^\perp)} + \frac{\mu_0^2\xi^2}{\text{dist}^2(\boldsymbol{B}_0, \boldsymbol{S}^\perp)}} =: \underline{\beta} \le \beta < 1. \tag{30}$$

Then it holds that $\text{dist}(\boldsymbol{B}_k, \boldsymbol{S}^\perp) \le \text{dist}(\boldsymbol{B}_0, \boldsymbol{S}^\perp)\beta^k + \omega$.

With a strategy of geometrically diminishing step size in Algorithm 1, Theorem 5 implies that the RSGM with proper initialization converges to a neighborhood of $\boldsymbol{S}^\perp$ at a linear rate, whose radius ω is proportional to the effective noise level. We note that the decaying rate of $\text{dist}(\boldsymbol{B}_k, \boldsymbol{S}^\perp)$ is determined by the diminishing factor β, which is well-defined in (30). A large β may lead to a slow convergence rate while a small β, e.g., smaller than $\underline{\beta}$, may lead to divergence. Moreover, if no noise is present, we have $\omega = 0$, which implies a linear convergence to $\boldsymbol{S}^\perp$, and is consistent with Zhu et al. (2019, Theorem 1). The above discussion is illustrated in Figures 3a and 3b. We also demonstrate the effect of codimension $c = D - d$ on the RSGM when applied to problem (2). As in Figures 3c and 3d, the RSGM exhibits a similar pattern for various c: it converges linearly to $\boldsymbol{S}^\perp$ with noiseless data, and converges to a neighborhood of $\boldsymbol{S}^\perp$ when the noise level is moderate.

Simulations. We generate data from the random spherical model in Definition 2. All results are obtained on a 64-bit machine with 2.3GHz Intel Xeon Gold 5218 CPU. We compare the performance of the proposed holistic approach (2) with the recursive approach (1) as well as other methods that include PCA, REAPER (Lerman et al., 2015), and GGD (Maunu et al., 2019). Both REAPER and GGD are primarily designed for learning a low-dimensional subspace. However, since the objective problem of GGD is similar to (2) except that it learns a basis for $\mathcal{S}$ instead of

$\mathcal{S}^\perp$, we also apply GGD to learn a basis of $\mathcal{S}^\perp$, and call it GGD-dual. We conduct the experiments with $D = 1000$, $c = 50$ and $N = 10D$ and plot the phase transition of the distance between the ground-truth basis for the (dual) subspace and the basis computed by different methods when varying the outlier ratio $\frac{M}{M+N}$ and noise level σ.

As demonstrated in Figure 4, PCA and REAPER are the least competitive methods in the test. We conjecture that REAPER does not perform well as an RSR method because it needs more inlier points for the underlying convex relaxation to be effective (in contrast to the non-convex approaches used by GGD and DPCP). Next, GGD, GGD-dual and DPCP-holistic perform similarly well in terms of accurately estimating a ground-truth basis. However, GGD takes significantly longer since it optimizes over $\mathbb{G}(D, d)$, which is inefficient in the high relative dimension regime. We see that applying GGD to learn the dual subspace in $\mathbb{G}(D, c)$, i.e., GGD-dual, is much faster, although not as fast as our holistic DPCP approach that solves (2) with RSGM. Finally, we note that the recursive DPCP approach based on solving (1) with RSGM is slow due to its computational cost; moreover, as the outlier ratio and noise level increase, its estimation of the underlying subspace becomes less accurate since the error tends to accumulate during the recursive procedure. We conclude that the proposed holistic DPCP approach performs favorably against the competitors in the high relative dimension regime.

In practice, the codimension c of the underlying subspace is usually unknown. Nevertheless, we can still apply the proposed holistic DPCP approach with an estimated codimension c^+. In particular, the theoretical results in Section 4 can be naturally extended to the case of $c^+ \leq c$, and guarantee that the holistic DPCP approach finds c^+ normal vectors to the underlying subspace. When $c^+ > c$, one may empirically observe that the holistic approach finds a solution $\boldsymbol{B} \in \mathbb{R}^{D\times c^+}$ consisting of c vectors (approximately) orthogonal to $\mathcal{S}$. We can then further obtain an estimate of c by computing $\{\|\widetilde{\boldsymbol{\mathcal{X}}}^\top \boldsymbol{b}_j\|_2\}_{j=1}^{c^+}$ with $\boldsymbol{b}_j$ the j-th column of $\boldsymbol{B}$ and counting the number of columns that have relatively small values.

6. Conclusions

We considered a holistic Dual Principal Component Pursuit (DPCP) approach for robust subspace learning in the high relative dimension regime, which involves non-convex optimization on the Grassmannian and simultaneously estimating the entire basis of the orthogonal complement subspace. We provided global optimality analyses, and showed it can handle $O((\#\text{inliers})^2)$ outliers, in both noiseless and noisy settings. We also proved that an RSGM method converges linearly to a neighborhood of the orthogonal complement subspace, whose region is proportional to the noise level. Extending the holistic approach to multiple subspaces can be the subject of future work.

Acknowledgements

This research is supported in part by NSF grants 1704458 and 2008460.

References

Brooks, J. P., Dulá, J. H., and Boone, E. L. A pure l1-norm principal component analysis. *Computational statistics & data analysis*, 61:83–98, 2013.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Ding, T., Zhu, Z., Ding, T., Yang, Y., Robinson, D., Vidal, R., and Tsakiris, M. Noisy dual principal component pursuit. In *Proceedings of the International Conference on Machine learning*, 2019.

Ding, T., Yang, Y., Zhu, Z., Robinson, D. P., Vidal, R., Kneip, L., and Tsakiris, M. C. Robust homography estimation via dual principal component pursuit. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6080–6089, 2020.

Ding, T., Zhu, Z., Tsakiris, M., Vidal, R., and Robinson, D. Dual principal component pursuit for learning a union of hyperplanes: Theory and algorithms. In *International Conference on Artificial Intelligence and Statistics*, pp. 2944–2952. PMLR, 2021.

Edelman, A., Arias, T. A., and Smith, S. T. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

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.

Fischler, M. A. and Bolles, R. C. 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.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

Hartley, R. and Zisserman, A. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

Higham, N. and Papadimitriou, P. Matrix procrustes problems. *Rapport technique, University of Manchester*, 1995.

Jolliffe, I. T. Principal components in regression analysis. In *Principal component analysis*, pp. 129–155. Springer, 1986.

Knyazev, A. V. and Zhu, P. Principal angles between subspaces and their tangents. *arXiv preprint arXiv:1209.0523*, 2012.

Lerman, G. and Maunu, T. Fast, robust and non-convex subspace recovery. *Information and Inference: A Journal of the IMA*, 7(2):277–336, 2018a.

Lerman, G. and Maunu, T. An overview of robust subspace recovery. *Proceedings of the IEEE*, 106(8):1380–1410, 2018b.

Lerman, G., McCoy, M. B., Tropp, J. A., and Zhang, T. Robust computation of linear models by convex relaxation. *Foundations of Computational Mathematics*, 15 (2):363–410, 2015.

Markopoulos, P. P., Kundu, S., Chamadia, S., Tsagkarakis, N., and Pados, D. A. Outlier-resistant data processing with l1-norm principal component analysis. In *Advances in Principal Component Analysis*, pp. 121–135. Springer, 2018.

Maunu, T., Zhang, T., and Lerman, G. A well-tempered landscape for non-convex robust subspace recovery. *Journal of Machine Learning Research*, 20(37):1–59, 2019.

McCoy, M., Tropp, J. A., et al. Two proposals for robust pca using semidefinite programming. *Electronic Journal of Statistics*, 5:1123–1160, 2011.

Rahmani, M. and Atia, G. K. Coherence pursuit: Fast, simple, and robust principal component analysis. *IEEE Transactions on Signal Processing*, 65(23):6260–6275, 2017.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

Tsakiris, M. C. and Vidal, R. Hyperplane clustering via dual principal component pursuit. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3472–3481. JMLR. org, 2017.

Tsakiris, M. C. and Vidal, R. Dual principal component pursuit. *The Journal of Machine Learning Research*, 19 (1):684–732, 2018.

Vaswani, N. and Narayanamurthy, P. Static and dynamic robust pca and matrix completion: A review. *Proceedings of the IEEE*, 106(8):1359–1379, 2018.

Xu, H., Caramanis, C., and Sanghavi, S. Robust pca via outlier pursuit. *IEEE transactions on information theory*, 58(5):3047–3064, 2012.

Yang, W. H., Zhang, L.-H., and Song, R. Optimality conditions for the nonlinear programming problems on riemannian manifolds. *Pacific Journal of Optimization*, 10(2): 415–434, 2014.

You, C., Li, C.-G., Robinson, D. P., and Vidal, R. Oracle based active set algorithm for scalable elastic net subspace clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3928–3937, 2016.

Zhang, T. and Lerman, G. A novel m-estimator for robust pca. *The Journal of Machine Learning Research*, 15(1): 749–808, 2014.

Zhu, Z., Wang, Y., Robinson, D., Naiman, D., Vidal, R., and Tsakiris, M. Dual principal component pursuit: Improved analysis and efficient algorithms. *Neural Information Processing Systems*, 2018.

Zhu, Z., Ding, T., Robinson, D., Tsakiris, M., and Vidal, R. A linearly convergent method for non-smooth non-convex optimization on the grassmannian with applications to robust subspace and dictionary learning. In *Advances in Neural Information Processing Systems*, pp. 9437–9447, 2019.