Dual Principal Component Pursuit for Learning a Union of Hyperplanes: Theory and Algorithms

Tianyu Ding[1] **Zhihui Zhu**[2] **Manolis Tsakiris**[3] **René Vidal**[1] **Daniel Robinson**[4]
[1]Johns Hopkins University [2]University of Denver [3]ShanghaiTech University [4]Lehigh University

Abstract

State-of-the-art subspace clustering methods are based on convex formulations whose theoretical guarantees require the subspaces to be low-dimensional. *Dual Principal Component Pursuit (DPCP)* is a non-convex method that is specifically designed for learning high-dimensional subspaces, such as hyperplanes. However, existing analyses of DPCP in the multi-hyperplane case lack a precise characterization of the distribution of the data and involve quantities that are difficult to interpret. Moreover, the provable algorithm based on recursive linear programming is not efficient. In this paper, we introduce a new notion of *geometric dominance*, which explicitly captures the distribution of the data, and derive both geometric and probabilistic conditions under which a global solution to DPCP is a normal vector to a geometrically dominant hyperplane. We then prove that the DPCP problem for a *union of hyperplanes* satisfies a Riemannian regularity condition, and use this result to show that a scalable Riemannian subgradient method exhibits (local) linear convergence to the normal vector of the geometrically dominant hyperplane. Finally, we show that integrating DPCP into popular subspace clustering schemes, such as K-ensembles, leads to superior or competitive performance over the state-of-the-art in clustering hyperplanes.

1 INTRODUCTION

Subspace clustering (SC) (Vidal, 2011) assumes data points are drawn from a union of subspaces, and the goal is to estimate the subspaces and cluster the data points according to their membership. Typically, existing SC methods require the underlying subspaces to be of *low-relative dimension* compared to the ambient space in order to enjoy strong theoretical guarantees together with efficient implementations, which have been heavily researched in the past decade. For example, the self-expressive approaches (Elhamifar and Vidal, 2009, 2013; Liu et al., 2010; Lu et al., 2012; Vidal and Favaro, 2014; You et al., 2016a,b) assume each data point can be expressed as a sparse linear combination of other data points from the same subspace.

On the other hand, clustering subspaces of *high-relative dimension* is less studied, with one of the most interesting cases being hyperplane clustering (HC). Many applications in computer vision and machine learning can be reduced to HC problems, such as motion segmentation (Tron and Vidal, 2007; Vidal et al., 2006, 2008), hybrid system identification (Bako, 2011; Vidal et al., 2003), and sparse component analysis (Georgiev et al., 2005; He and Cichocki, 2007; Xu et al., 2018). However, simply applying SC methods that are designed for the *low-relative dimension* to HC is ineffective because the theory and algorithms do not apply to a union of hyperplanes (UoH) setting.

There are several mainstream methods for HC. First, the Random Sampling and Consensus (RANSAC) (Fischler and Bolles, 1981) is a popular heuristic based on fitting one hyperplane at a time using principal component analysis (PCA) from many randomly sampled points; this process is repeating after the points identified as belonging to the previously selected hyperplanes are removed. However, it suffers from an exponential complexity as the number of hyperplanes grows. Second, Algebraic Subspace Clustering (ASC) enjoys strong theoretical guarantees for hyperplanes (Tsakiris and Vidal, 2017a,b; Vidal et al., 2005), but is not robust to outliers and is computationally expensive when the ambient dimension is high. Third, K-subspaces (KSS) (Agarwal and Mustafa, 2004; Bradley and Mangasarian, 2000) is another attractive method that alternates between assigning data points to clus-

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

Table 1: Comparison of the theory and algorithms for learning a hyperplane under a UoH model.

	Theory			Algorithms		
	Which hyperplane does it recover?	Handle outliers	Analytical approach		Convergence guarantee	Scale well?
Lerman and Zhang (2014)	most significant plane (see (5))	✓	probabilistic	–	–	–
Tsakiris and Vidal (2017c)	dominant plane (see (4))	✗	geometric	LPs IRLS	✓ ✗	✗ ✓
This paper	geometrically dominant plane (see Definition 1)	✓	probabilistic + geometric	RSGM	✓	✓

ters and estimating a subspace for each cluster using PCA. KSS is scalable in practice, but it can easily get stuck near a local minimum due to its non-convex nature and it is not robust to outliers. The suboptimality issue can be addressed by leveraging ensembles of KSS (Lane et al., 2019; Lipor et al., 2018), while the lack of robustness stems from the fact that the squared ℓ_2 loss used in PCA is incapable of handling outliers.

In this paper, we analyze *Dual Principal Component Pursuit (DPCP)* (Tsakiris and Vidal, 2018) for learning a hyperplane from data under a UoH model, and to show the superiority of embedding DPCP into popular schemes (e.g., KSS) for clustering hyperplanes. It is known that DPCP can robustly learn a *single* hyperplane and tolerate outliers on the order of the square of the number of inliers (Ding et al., 2019; Zhu et al., 2018a) by computing a basis for the orthogonal complement of the subspace, which itself is computed by solving a non-convex ℓ_1 optimization problem on the sphere. It is not known, however, whether DPCP can learn a normal to one of the hyperplanes in the presence of both *structured* and *regular* outliers.[1] In fact, several related questions remain unanswered. Under what conditions is a global optimum of the DPCP problem a normal to one of the hyperplanes? When the global optimum is a normal, which hyperplane is it a normal to? Can the convergence of some optimization algorithm to a global solution to the non-convex DPCP problem under the UoH data model be guaranteed?

This paper addresses all of the above challenges associated with DPCP. Specifically, the main contributions of this paper can be summarized as follows.

- We introduce a new notion of geometric dominance for determining the hyperplane that is learned by DPCP under a UoH model, which then leads to an intuitive deterministic analysis that explicitly captures the data distribution *and* the geometric relationships among the hyperplanes.
- We derive conditions under which the global minimizer of DPCP for a UoH is guaranteed to be a normal vector of the geometrically dominant hyperplane. Our conditions replace the geometric quantities in Tsakiris and Vidal (2017c) with tighter ones that are amenable to outliers and easier to bound in probability. This approach leads to a new probabilistic guarantee for recovering the geometrically dominant hyperplane when it has sufficiently many points relative to the other planes with a mild requirement on the total number of points (e.g., $\Omega(D^3)$ with D the dimension of the ambient space), thus *significantly* improving upon Lerman and Zhang (2014), which requires $\Omega(D^{18}\log D)$ points.
- We prove that the objective problem of DPCP under a UoH data model satisfies a Riemannian Regularity Condition (RRC) (Zhu et al., 2019), and then use the RRC to show that a Riemannian subgradient method (RSGM, Algorithm 1) converges linearly to a normal vector of the geometrically dominant hyperplane if properly initialized. In particular, RSGM only involves matrix-vector multiplications, which makes it more scalable than the LP or SVD-based IRLS method proposed in Tsakiris and Vidal (2017c).
- We integrate DPCP into KSS (DPCP-KSS) by using DPCP to estimate the geometrically dominant hyperplane for each cluster, and leverage an ensemble of DPCP-KSS via the EKSS (Lipor et al., 2018) and CoRe (Lane et al., 2019) frameworks. Experiments demonstrate the superiority of using DPCP-KSS (implemented with RSGM) within various schemes for clustering hyperplanes.

Related work. Tsakiris and Vidal (2017c) have partially addressed the previous challenges of DPCP for a UoH without outliers while Lerman and Zhang (2014) analyzed ℓ_p recovery of a single subspace from a union of subspaces with UoH as a special case. Three key differences should be emphasized (see Table 1 for a summary). First, in the analysis of which hyperplane is recovered, Tsakiris and Vidal (2017c) and Lerman and Zhang (2014) introduce different notions of a "significant" or "dominant" hyperplane, which depend only on the (expected) number of points in each group. We argue that the global optimum depends not only on

[1] In learning a single hyperplane from data under a UoH model, the *structured* outliers are the data points that come from the remaining hyperplanes; *regular* outliers are uniformly distributed in the ambient space. Throughout this paper, unless stated otherwise, outliers refer to the regular kind, and a UoH model contains regular outliers.

the number of data points in each group, but also on geometric quantities related to their distribution. Currently there is no notion of *geometric dominance* that captures these aspects. Second, Tsakiris and Vidal (2017c) provide geometric conditions under which the global minimum is a normal to the "dominant" hyperplane, and Lerman and Zhang (2014) provide probabilistic conditions. However, neither have *both* types of analyses, nor do the analyses make connections to geometric dominance. Third, the provably convergent algorithm in Tsakiris and Vidal (2017c), which is based on a recursion of linear programs (LPs), is not scalable, while the recommended *Iteratively Reweighted Least Squares* (IRLS) (Lerman and Maunu, 2018a; Lerman et al., 2015) approach does not have a guarantee for the DPCP problem. In other words, there does not exist a scalable algorithm that ensures global convergence for learning a single hyperplane under a UoH model.

Other improvements on KSS. The theory of DPCP for a UoH is ideally matched to the subspace estimation step of KSS, where most of the points in the estimated cluster are expected to belong to a single hyperplane with the remaining points belonging to the other hyperplanes. This suggests using DPCP instead of PCA in KSS for its robustness in fitting a hyperplane. Although GGD (Maunu et al., 2019) and REAPER (Lerman et al., 2015) share similar objectives with DPCP, both are primarily designed for low-dimensional subspace recovery. For example, REAPER requires $d < (D-1)/2$ in theory, where d and D are the dimensions of the subspace and ambient space, respectively. In other related work, Median K-Flats (MKF) (Zhang et al., 2009) replaces the squared ℓ_2 objective in KSS with an unsquared one, but it lacks competitive performance as observed by Gitlin et al. (2018). Alternatively, Gitlin et al. (2018) substituted PCA in KSS by Coherence Pursuit (CoP) (Rahmani and Atia, 2017a), but the theory requires $d < \sqrt{D}$, thus making it unsuitable for hyperplanes.

2 BACKGROUND

We first describe the data model used in this paper. Consider the ℓ_2 column-normalized dataset $\widetilde{\boldsymbol{\mathcal{X}}} = [\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{O}}]\boldsymbol{\Gamma} \in \mathbb{R}^{D\times(N+M)}$, where $\boldsymbol{\mathcal{X}} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N] \in \mathbb{R}^{D\times N}$ are N inlier points that lie in the union of K hyperplanes $\mathcal{H}_1, \cdots, \mathcal{H}_K$ of $\mathbb{R}^D$ with unit normal vectors $\boldsymbol{b}_1, \cdots, \boldsymbol{b}_K$, respectively, $\boldsymbol{\mathcal{O}} = [\boldsymbol{o}_1, \cdots, \boldsymbol{o}_M] \in \mathbb{R}^{D\times M}$ are M outliers that lie on the unit sphere $\mathbb{S}^{D-1}$ in $\mathbb{R}^D$, and $\boldsymbol{\Gamma}$ is an unknown permutation. We assume that for every $k \in [K] := \{1, \cdots, K\}$, there are N_k inlier points, denoted by $\boldsymbol{\mathcal{X}}_k \subset \boldsymbol{\mathcal{X}}$, that belong to $\mathcal{H}_k$. Given this model, our goal is to estimate the underlying hyperplanes $\{\mathcal{H}_k\}$ from $\widetilde{\boldsymbol{\mathcal{X}}}$, as well as cluster the data points according to their nearest hyperplane.

Note that if $\boldsymbol{b}$ is a normal vector to a hyperplane, it is orthogonal to all the data points within this hyperplane. Thus, we attempt to find a normal vector to one specific hyperplane by solving

$$\min_{\boldsymbol{b}\in\mathbb{S}^{D-1}} \; f(\boldsymbol{b}) := \|\widetilde{\boldsymbol{\mathcal{X}}}^\top \boldsymbol{b}\|_1 = \sum_{k=1}^{K} \|\boldsymbol{\mathcal{X}}_k^\top \boldsymbol{b}\|_1 + \|\boldsymbol{\mathcal{O}}^\top \boldsymbol{b}\|_1 \quad (1)$$

which is called *Dual Principal Component Pursuit (DPCP)*. For learning a *single* hyperplane $\mathcal{H}$, when the inliers are uniformly distributed in $\mathcal{H} \cap \mathbb{S}^{D-1}$ and the outliers are uniformly distributed in $\mathbb{S}^{D-1}$, the DPCP problem (1) can provably recover the true normal vector to $\mathcal{H}$ provided that the number of outliers is big-O of the square of the number of inliers (Ding et al., 2019; Zhu et al., 2018a). The problem is more challenging when $\boldsymbol{\mathcal{X}}$ consists of inliers from a union of K hyperplanes. The analysis of a single hyperplane *cannot* be applied here by treating the data points from one hyperplane as inliers and the rest as outliers since the data distribution in other planes is far from uniform and thus violates the prior analysis.

We now introduce several geometric quantities from Zhu et al. (2018a) that characterize how well the inliers and outliers are distributed. First, to characterize the distribution of outliers, we use the maximum norm Riemannian subgradient of the function $\frac{1}{M}\|\boldsymbol{\mathcal{O}}^\top \boldsymbol{b}\|_1$, which we denote by

$$\eta_{\boldsymbol{\mathcal{O}}} := \max_{\boldsymbol{b}\in\mathbb{S}^{D-1}} \|(\mathbf{I} - \boldsymbol{b}\boldsymbol{b}^\top)\boldsymbol{\mathcal{O}} \operatorname{sign}(\boldsymbol{\mathcal{O}}^\top \boldsymbol{b})\|_2/M, \quad (2)$$

where $\operatorname{sign}(a) = a/|a|$ if $a \neq 0$ else 0, and $\operatorname{sign}(\boldsymbol{a})$ denotes the application of the sign function elementwise to the vector $\boldsymbol{a}$. More uniformly distributed outliers lead to smaller values of $\eta_{\boldsymbol{\mathcal{O}}}$. This follows since if $M \to \infty$ and $\boldsymbol{\mathcal{O}}$ is well distributed, then $\boldsymbol{\mathcal{O}} \operatorname{sign}(\boldsymbol{\mathcal{O}}^\top \boldsymbol{b})/M$ approaches the direction of $\boldsymbol{b}$, which leads to $\eta_{\boldsymbol{\mathcal{O}}} \to 0$ (Zhu et al., 2018a). Second, for the inlier subset $\boldsymbol{\mathcal{X}}_k$ in hyperplane $\mathcal{H}_k$, we define

$$\begin{aligned} c_{\boldsymbol{\mathcal{X}}_k,\min} &:= \min_{\boldsymbol{b}\in\mathcal{H}_k\cap\mathbb{S}^{D-1}} \|\boldsymbol{\mathcal{X}}_k^\top \boldsymbol{b}\|_1/N_k, \\ c_{\boldsymbol{\mathcal{X}}_k,\max} &:= \max_{\boldsymbol{b}\in\mathcal{H}_k\cap\mathbb{S}^{D-1}} \|\boldsymbol{\mathcal{X}}_k^\top \boldsymbol{b}\|_1/N_k. \end{aligned} \quad (3)$$

Note that $c_{\boldsymbol{\mathcal{X}}_k,\min}$ is exactly the *permeance statistic* defined in Lerman et al. (2015). A well-distributed $\boldsymbol{\mathcal{X}}_k$ leads to a large value of $c_{\boldsymbol{\mathcal{X}}_k,\min}$ and small value of $c_{\boldsymbol{\mathcal{X}}_k,\max}$ since it is difficult to find a single direction $\boldsymbol{b}$ that is orthogonal to (or in line with) many points in $\boldsymbol{\mathcal{X}}_k$. Parallel to (2) and (3), we also define the following quantities that further characterize the distribution of inliers and outliers, respectively:

$$\begin{aligned} \eta_{\boldsymbol{\mathcal{X}}_k} &:= \max_{\boldsymbol{b}\in\mathcal{H}_k\cap\mathbb{S}^{D-1}} \|(\boldsymbol{P}_{\mathcal{H}_k} - \boldsymbol{b}\boldsymbol{b}^\top)\boldsymbol{\mathcal{X}}_k \operatorname{sign}(\boldsymbol{\mathcal{X}}_k^\top \boldsymbol{b})\|_2/N_k, \\ c_{\boldsymbol{\mathcal{O}},\min} &:= \min_{\boldsymbol{b}\in\mathbb{S}^{D-1}} \|\boldsymbol{\mathcal{O}}^\top \boldsymbol{b}\|_1/M, \\ c_{\boldsymbol{\mathcal{O}},\max} &:= \max_{\boldsymbol{b}\in\mathbb{S}^{D-1}} \|\boldsymbol{\mathcal{O}}^\top \boldsymbol{b}\|_1/M, \end{aligned}$$

where $\boldsymbol{P}_{\mathcal{H}_k}$ is the orthonormal projection onto $\mathcal{H}_k$. We will see shortly that the global optimality theory based on these geometric quantities is easier to interpret and facilitates a probabilistic analysis.

3 ANALYSIS OF DPCP FOR A UNION OF HYPERPLANES

3.1 Geometrically Dominant Hyperplane

We first review the definitions of a dominant hyperplane in Lerman and Zhang (2014); Tsakiris and Vidal (2017c). The hyperplane (say $\mathcal{H}_1$) with the most number of points is defined as the *dominant hyperplane* in Tsakiris and Vidal (2017c), i.e.,

$$N_1 > \max_{k\geq 2} N_k. \tag{4}$$

It is proved in Tsakiris and Vidal (2017c) that a global solution of (1) is a normal vector of $\mathcal{H}_1$ under certain conditions, which implicitly make use of the distribution of the data, but are deterministic in nature and difficult to interpret. On the other hand, the work of Lerman and Zhang (2014) considers a random model where inliers are sampled from $(\cup_{k=1}^{K}\mathcal{H}_k)\cap\mathbb{S}^{D-1}$ with weights $\{\alpha_k\}_{k=1}^{K}$ (α_k is the weight of sampling inliers in $\mathcal{H}_k$) and outliers are sampled from $\mathbb{S}^{D-1}$ with weight α_0, and $\sum_{k=0}^{K}\alpha_k = 1$. Then $\mathcal{H}_1$ is defined as the *most significant hyperplane* if

$$\alpha_1 > \sum_{k=2}^{K}\alpha_k. \tag{5}$$

The number of sampled points, in expectation, is equivalent to $N_1 > \sum_{k\geq 2} N_k$. In contrast to (4) and (5), the hyperplane that we target depends on the point weights as well as the distribution and geometric relationships among the planes. We call such a plane a *geometrically dominant* hyperplane.

Geometrically dominant hyperplane. Recall our goal is to minimize the objective in (1). Intuitively, the outlier term $\|\boldsymbol{\mathcal{O}}^\top\boldsymbol{b}\|_1$ should be nearly constant for well distributed outliers, so that the minimizer of (1) is determined by the relative importance of the inlier terms $\|\boldsymbol{\mathcal{X}}_k^\top\boldsymbol{b}\|_1$.We also expect the relative orientation of the underlying hyperplanes to play an important role in determining the solution to (1). For example, in the case that data are uniformly sampled and each plane has the same point weights, the solution of (1) has a bias towards the normals of the planes that are close to each other. Noting that the geometric relationships between $\mathcal{H}_k$'s are determined by the principal angles between the $\boldsymbol{b}_k$'s, we define $\theta_{k\ell}\in[0,\pi/2]$ to be the principal angle between $\boldsymbol{b}_k$ and $\boldsymbol{b}_\ell$. By analyzing the first-order necessary condition for problem (1), we define ζ_k that measures the relative dominance for $\boldsymbol{\mathcal{X}}_k$ and considers the integrated information of point weights, data distribution, and relative orientation of the hyperplanes:

$$\zeta_k := \frac{N_k c_{\boldsymbol{\mathcal{X}}_k,\min}}{\sqrt{\mathbf{1}^\top\boldsymbol{W}^{\max}_{(k,k)}\mathbf{1}} + \sum_{\ell\neq k} N_\ell\eta_{\boldsymbol{\mathcal{X}}_\ell} + M\eta_{\boldsymbol{\mathcal{O}}} + D}, \tag{6}$$

where $\boldsymbol{W}^{\max}\in\mathbb{R}^{K\times K}$ whose (k,ℓ)th entry is $N_k c_{\boldsymbol{\mathcal{X}}_k,\max} N_\ell c_{\boldsymbol{\mathcal{X}}_\ell,\max}\cos(\theta_{k\ell})$ and represents the joint importance of $\boldsymbol{\mathcal{X}}_k$ and $\boldsymbol{\mathcal{X}}_\ell$ weighted by $\cos(\theta_{k\ell})$, $\boldsymbol{W}^{\max}_{(k,k)}$ is the principal submatrix obtained by deleting the kth row and kth column of $\boldsymbol{W}^{\max}$, and $\mathbf{1}$ is the vector of all 1's. Noting that: (i) the numerator $N_k c_{\boldsymbol{\mathcal{X}}_k,\min}$ of (6) represents the contribution from $\boldsymbol{\mathcal{X}}_k$; (ii) the term $\mathbf{1}^\top\boldsymbol{W}^{\max}_{(k,k)}\mathbf{1}$ in the denominator counts the sum of the entries in $\boldsymbol{W}^{\max}_{(k,k)}$, capturing the total contributions from $\{\boldsymbol{\mathcal{X}}_\ell\}_{\ell\neq k}$; and (iii) the last term $\sum_{\ell\neq k} N_\ell\eta_{\boldsymbol{\mathcal{X}}_\ell} + M\eta_{\boldsymbol{\mathcal{O}}} + D$ is typically small[2] compared with the former two terms. Thus, overall ζ_k measures the relative dominance of $\boldsymbol{\mathcal{X}}_k$ over $\{\boldsymbol{\mathcal{X}}_\ell\}_{\ell\neq k}$. We see that larger relative dominance of $\boldsymbol{\mathcal{X}}_k$ (i.e. larger ζ_k) results from better distributed data points, larger N_k relative to M and N_ℓ for $\ell\neq k$, and better separation of the other hyperplanes (large $\theta_{ij}, i,j\neq k, i\neq j$).

Definition 1. *With ζ_k in (6), we say that $\mathcal{H}_k$ is a* geometrically dominant hyperplane *if $\zeta_k \geq \zeta_\ell, \forall\ell$.*

The notion of geometric dominance makes the deterministic analysis (Sec. 3.2) tighter, and allows a probabilistic analysis (Sec. 3.3) that is easier to be satisfied with only mild number of sampled points

Proposition 1. *There is at most one $k\in[K]$ such that $\zeta_k > 1$, and then $\zeta_\ell < 1$ for all $\ell\in[K]\backslash k$.*

It follows from Proposition 1 that if $\zeta_k > 1$ then $\mathcal{H}_k$ is the *unique* geometrically dominant hyperplane. For the rest of the analysis, we assume that there always exists $k\in[K]$ such that $\zeta_k > 1$; the scenario that such a geometrically dominant hyperplane does not exist is left for future work. We note that this assumption ensures a simple landscape of the non-convex DPCP problem (1) that allows us to show that under certain conditions the global minimizers of (1) are guaranteed to be normal vectors of the geometrically dominant hyperplane (Theorem 1). The assumption may be stronger than needed in theory[3] since it excludes the possibility that normals of the other hyperplanes are global solutions to (1), which are also of our interest. Related works make similar assumptions—Tsakiris and Vidal (2017c)

[2]Assuming points in $\boldsymbol{\mathcal{X}}_k$ and $\boldsymbol{\mathcal{O}}$ are uniformly sampled from $\mathbb{S}^{D-1}\cap\mathcal{H}_k$ and $\mathbb{S}^{D-1}$, respectively, both $N_k c_{\boldsymbol{\mathcal{X}}_k,\max}$ and $N_k c_{\boldsymbol{\mathcal{X}}_k,\min}$ scale as $O(N_k)$, while $N_k\eta_{\boldsymbol{\mathcal{X}}_k}$ scales as $O(\sqrt{N_k})$ and $M\eta_{\boldsymbol{\mathcal{O}}}$ scales as $O(\sqrt{M})$ (Zhu et al., 2018a).

[3]In fact, Tsakiris and Vidal (2017c, Proposition 5) shows that for three equi-angular hyperplanes, global minimizers of (1) can be normal vectors of any of the planes when they are well-separated and the data points are well-distributed.

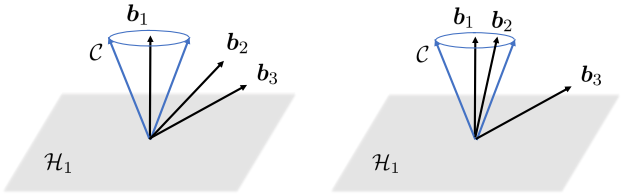

Figure 1: (Left) Since $\boldsymbol{b}_2, \boldsymbol{b}_3 \notin \mathcal{C}$, they could be critical points; (Right) Since $\boldsymbol{b}_2 \in \mathcal{C}$ it cannot be a critical point, but $\boldsymbol{b}_3$ could be because $\boldsymbol{b}_3 \notin \mathcal{C}$.

requires (4) and Lerman and Zhang (2014) requires (5). We will see that, when data is sampled from a specific random spherical model (Theorem 2), the geometric dominance not only implies both (4) and (5), but also has the advantage that it characterizes the data distribution. As mentioned above, this assumption is likely to be satisfied in the subspace estimation step of KSS where most of the points in the estimated cluster are expected to be sampled from one dominant hyperplane, which works well in practice as we will see in Section 4.

3.2 Deterministic Analysis of DPCP for a UoH

Without loss of generality, we assume $\zeta_1 > 1$, i.e., that $\mathcal{H}_1$ is the geometrically dominant hyperplane. We first characterize critical points of (1) with respect to the geometrically dominant hyperplane $\mathcal{H}_1$.

Lemma 1. *Any critical point $\boldsymbol{b}^*$ of (1) must belong to $\{\pm\boldsymbol{b}_1\}$ or have a principal angle θ from $\boldsymbol{b}_1$ satisfying $\theta \geq \arcsin(\sqrt{1-(1/\zeta_1)^2})$.*

Intuitively, Lemma 1 suggests that any critical point of (1) is either a normal vector of $\mathcal{H}_1$, or very close to $\mathcal{H}_1$ (i.e., within a region defined by the geometric dominance level of $\boldsymbol{\mathcal{X}}_1$). As the relative dominance of $\boldsymbol{\mathcal{X}}_1$ increases (larger ζ_1), the location of $\boldsymbol{b}^*$ becomes more restricted. In particular, Lemma 1 allows us to conclude that $\boldsymbol{b}_1$ is the single (up to direction) critical point inside of the cone $\mathcal{C} := \{\boldsymbol{y} \in \mathbb{R}^D : |\boldsymbol{y}^\top \boldsymbol{b}_1| > 1/\zeta_1, \|\boldsymbol{y}\|_2 = 1\}$ centered around $\pm\boldsymbol{b}_1$.

The above observation ensures that every normal in the set $\{\pm\boldsymbol{b}_2, \cdots, \pm\boldsymbol{b}_K\}$ that lies inside of $\mathcal{C}$ is not a critical point (see Figure 1). We will later see how this facilitates the convergence of an algorithm to $\{\pm\boldsymbol{b}_1\}$ when it is initialized inside $\mathcal{C}$ because $\boldsymbol{b}_1$ (up to direction) is the only possible solution within the region.

Lemma 1 helps us understand global solutions of (1). To show that any global minimizer $\boldsymbol{b}^*$ satisfies $\boldsymbol{b}^* \in \{\pm\boldsymbol{b}_1\}$, we need to ensure that every critical point close to $\mathcal{H}_1$ is not a global solution. Inspired by the analysis in Tsakiris and Vidal (2017c), we define

$$\gamma_k := \frac{N_k c_{\boldsymbol{\mathcal{X}}_k,\min}}{\sum_{\ell\neq k} N_\ell c_{\boldsymbol{\mathcal{X}}_\ell,\max}\sin(\theta_{k\ell}) - \sqrt{\sum_{i=2}^{K-1}\lambda_i\big(\boldsymbol{W}^{\min}_{(k,k)}\big)} + \Delta},$$
$$\text{where } \Delta := M(c_{\boldsymbol{\mathcal{O}},\max} - c_{\boldsymbol{\mathcal{O}},\min}). \tag{7}$$

Here, $\boldsymbol{W}^{\min}$ is the same as $\boldsymbol{W}^{\max}$ (see (6)) by replacing $c_{\boldsymbol{\mathcal{X}}_k,\max}c_{\boldsymbol{\mathcal{X}}_\ell,\max}$ with $c_{\boldsymbol{\mathcal{X}}_k,\min}c_{\boldsymbol{\mathcal{X}}_\ell,\min}$, and $\lambda_1(\boldsymbol{A}) \geq \cdots \geq \lambda_n(\boldsymbol{A})$ are the eigenvalues of an n-by-n matrix $\boldsymbol{A}$. In fact, we can show every global solution of (1) is not far from $\{\pm\boldsymbol{b}_1\}$ in the sense that its principal angle θ from $\boldsymbol{b}_1$ satisfies $\theta \leq \arcsin(1/\gamma_1)$. Combining this fact with Lemma 1 establishes our main theoretical result.

Theorem 1. *Any global solution of (1) is a normal to the geometrically dominant hyperplane $\mathcal{H}_1$ if*

$$(1/\zeta_1)^2 + (1/\gamma_1)^2 < 1. \tag{8}$$

Additional interpretation of γ_k and ζ_k is useful. Note that γ_k is similar to ζ_k, which characterizes the relative dominance of $\boldsymbol{\mathcal{X}}_k$ from a different perspective. First, the Δ term $M(c_{\boldsymbol{\mathcal{O}},\max} - c_{\boldsymbol{\mathcal{O}},\min})$ in the denominator of (7) represents the impact of outliers: uniformly distributed outliers with $M \to \infty$ cause the difference $c_{\boldsymbol{\mathcal{O}},\max} - c_{\boldsymbol{\mathcal{O}},\min}$ to vanish, making the term small. Next, to better analyze the square root part in (7), for simplicity we consider the equi-angular case for $\{\mathcal{H}_\ell\}_{\ell\neq k}$ such that $\theta_{ij} \equiv \theta'$ for all $i,j \neq k, i \neq j$, then one can obtain (see supplementary) $\sum_{i=2}^{K-1}\lambda_i(\boldsymbol{W}^{\min}_{(k,k)}) = (1-\cos(\theta'))\sum_{\ell\neq k,r} N_\ell^2 c^2_{\boldsymbol{\mathcal{X}}_\ell,\min}$, where $r = \arg\max_{\ell\neq k} N_\ell c_{\boldsymbol{\mathcal{X}}_\ell,\min}$. For a global solution to be a normal of $\mathcal{H}_k$, one may expect: (i) a large relative disparity in significance between $\boldsymbol{\mathcal{X}}_k$ and $\boldsymbol{\mathcal{X}}_\ell$ for all $\ell \neq k$ so that $\frac{N_k c_{\boldsymbol{\mathcal{X}}_k,\min}}{N_\ell c_{\boldsymbol{\mathcal{X}}_\ell,\max}}$ is large; (ii) $\mathcal{H}_k$ to be relatively close to the other planes so that the energy concentrated around $\mathcal{H}_k$ is relatively large, i.e., $\theta_{k\ell}$ is relatively small; and (iii) the other planes $\{\mathcal{H}_\ell\}_{\ell\neq k}$ are relatively separated so that the energy concentrated around any of them is relatively small, i.e., θ' is relatively large. All these make γ_k large.

An interpretation of Theorem 1 follows from the above discussion about ζ_k and γ_k: for a fixed number of inliers $\{N_k\}$ and outliers M, if data points are well-distributed (large $c_{\boldsymbol{\mathcal{X}}_k,\min}$, small $c_{\boldsymbol{\mathcal{X}}_k,\max}$, small $\eta_{\boldsymbol{\mathcal{X}}_k}$, small $\eta_{\boldsymbol{\mathcal{O}}}$, small $c_{\boldsymbol{\mathcal{O}},\max} - c_{\boldsymbol{\mathcal{O}},\min}$), $\mathcal{H}_1$ is closer to the other planes (relatively small $\theta_{1\ell}, \ell \neq 1$) than the other planes are to each other (relatively large $\theta_{ij}, i,j \neq 1, i \neq j$), then both ζ_1 and γ_1 tend to be large, (8) is more likely to be satisfied, and any global minimizer is a normal vector of $\mathcal{H}_1$. In contrast to the discrete result in Tsakiris and Vidal (2017c), which is based on a continuous variant of (1) without outliers and uses quantities such as the *spherical cap discrepancy* or *circumradii of polytopes* that are difficult to interpret, the global geometric analysis here focuses on the discrete problem (1) and leverages geometric quantities to explicitly characterize the underlying distribution of both inliers and outliers.

When the dataset is further contaminated with noise, one may expect that the error between the global minimizer and the true normal vector to $\mathcal{H}_1$ is proportional to the noise level, as analyzed for a single subspace case in Ding et al. (2019). We leave it as future work.

3.3 Probabilistic Analysis of DPCP for a UoH

Since the geometric quantities have corresponding concentrations in probability (Zhu et al., 2018a), the new approach leads to the following probabilistic guarantee.

Theorem 2. *Consider a random spherical model where the M columns of $\boldsymbol{\mathcal{O}}$ are drawn uniformly from the sphere $\mathbb{S}^{D-1}$, and the N_k columns of $\boldsymbol{\mathcal{X}}_k$ are drawn uniformly from $\mathbb{S}^{D-1} \cap \mathcal{H}_k$ for $k \in [K]$, where $\mathcal{H}_k$ is a hyperplane in $\mathbb{R}^D$. Then the probability that any global solution of (1) is a normal vector of $\mathcal{H}_1$ is at least $1 - 2(K+1)e^{-t^2/2}$, where $t > 0$ satisfies*

$$C_0 \sum_{k\neq 1} N_k + \Big(C_1\sqrt{D}\log(D) + \frac{3t}{2}\Big)\sum_{k\neq 1}\sqrt{N_k} \qquad (9)$$
$$+ \big(C_2\sqrt{D}\log D + t\big)\sqrt{M} < C_0 N_1 - \big(\sqrt{2} + \frac{t}{2\sqrt{2}}\big)\sqrt{N_1},$$

C_1 and C_2 are universal constants that are independent of K, $\{N_k\}$, M, D and t, and

$$C_0 := \frac{(D-3)!!}{(D-2)!!} \cdot \begin{cases} \frac{2}{\pi} & \text{if } D \text{ is even,} \\ 1 & \text{if } D \text{ is odd.} \end{cases} \qquad (10)$$

Note that $C_0 \in \big[\sqrt{\frac{2}{\pi(D-1)}}, \sqrt{\frac{1}{D-1}}\big]$(Zhu et al., 2018b, footnote 9) is a constant for fixed D. As the number of inliers from the hyperplanes goes to infinity and the other parameters are fixed, (9) roughly requires $\sum_{k\neq 1} N_k < N_1$, which coincides with (5) of Lerman and Zhang (2014) (in expectation). Also, as the number of inliers goes to infinity, (9) implies that the DPCP approach can tolerate $M = O((N_1 - \sum_{k\neq 1} N_k)/D)^2)$ outliers, which generalizes the result in Zhu et al. (2018a) for a single subspace.

A similar probabilistic result is provided in Lerman and Zhang (2014, Theorem 1.1) but for a different generative model where the number of points sampled in each hyperplane is not fixed in advance, as opposed to M and $\{N_k\}$ here, but is controlled by the sampling weights $\{\alpha_k\}_{k=0}^K$ (see Sec. 3.1). With this difference in mind, we now compare Lerman and Zhang (2014, Theorem 1.1) with (9). Towards that goal, dividing both sides of (9) by the total number of data points $N+M$, and viewing $\frac{M}{N+M}$ as α_0 and $\frac{N_k}{N+M}$ as α_k, gives

$$\alpha_1 > \sum_{k=2}^K \alpha_k + \frac{3\sqrt{D}\cdot t + \rho(D)}{\sqrt{N+M}} \sum_{k=0}^K \sqrt{\alpha_k}, \qquad (11)$$

where $\rho(D) := \sqrt{2}D\log D \max(C_1, C_2)$. Our result and Lerman and Zhang (2014, Theorem 1.1) require a similar condition on α_k to guarantee that any global solution of (1) is a normal vector of $\mathcal{H}_1$ with certain probability. On one hand, (11) requires α_1 to be larger than $\sum_{k=2}^K \alpha_k$ by a positive amount (which goes to 0 if the total number of points goes to infinity), which is slightly stronger than (5) in Lerman and Zhang (2014). On the other hand, Lerman and Zhang (2014, Theorem 1.1) only ensures a probability of $1 - C_3\exp(-\frac{N+M}{C_4})$, where $C_3 = O(D^{D(D-1)/2} + D^{8(D-1)})$ and $C_4 = O(D^{16})$ (assuming the other parameters such as K are fixed), thus needing to sample $\Omega(D^{18}\log D)$ points to make the probability overwhelming (e.g., probability of $1 - O(\exp(-D))$ if $N + M = \Omega(D^{19}\log D)$). For comparison, by taking $t = \sqrt{\frac{N+M}{D^3}}$, Theorem 2 now requires α_1 to be larger than $\sum_{k=2}^K \alpha_k$ by a small amount of $\big(\frac{3}{D} + \frac{\rho(D)}{\sqrt{N+M}}\big)\sum\sqrt{\alpha_k}$ and guarantees with probability $1 - 2(K+1)\exp(-\frac{N+M}{2D^3})$, which only requires a total sampling of $\Omega(D^3)$ points to make the probability overwhelming (e.g., probability of $1 - O(\exp(-D))$ if $N + M = \Omega(D^4)$), which is much smaller than the $\Omega(D^{18}\log D)$ needed in Lerman and Zhang (2014).

Algorithm 1 Riemannian Subgradient Method

1: **Initialization:** $\widehat{\boldsymbol{b}}_0 \in \mathbb{S}^{D-1}$, μ_0, and $\beta \in (0,1)$.
2: **for** $t = 0, 1, 2, \cdots$ **do**
3: Update the step size: $\mu_t \leftarrow \mu_0\beta^t$.
4: Compute a Riemannian subgradient:
$$\mathcal{G}(\widehat{\boldsymbol{b}}_t) \leftarrow (\mathbf{I} - \widehat{\boldsymbol{b}}_t\widehat{\boldsymbol{b}}_t^\top)\widetilde{\boldsymbol{\mathcal{X}}}\operatorname{sign}(\widetilde{\boldsymbol{\mathcal{X}}}^\top\widehat{\boldsymbol{b}}_t).$$
5: Update the iterate as:
$$\widetilde{\boldsymbol{b}}_{t+1} \leftarrow \widehat{\boldsymbol{b}}_t - \mu_t\mathcal{G}(\widehat{\boldsymbol{b}}_t),$$
$$\widehat{\boldsymbol{b}}_{t+1} \leftarrow \widetilde{\boldsymbol{b}}_{t+1}/\|\widetilde{\boldsymbol{b}}_{t+1}\|_2.$$
6: **end for**

3.4 Analysis of Projected Riemannian Subgradient Descent for a UoH

We have shown that the non-convex DPCP problem (1) is effective in robustly recovering a specific hyperplane for a UoH. The work of Tsakiris and Vidal (2017c) proposed to solve (1) by either an LP-based algorithm that involves a sequence of convex optimization problems thus is computationally expensive, or an IRLS algorithm that requires doing an SVD in each iteration and lacks a convergence guarantee. Here, we will utilize the efficient Riemannian subgradient method (RSGM) stated as Algorithm 1, and focus on its convergence to the geometrically dominant hyperplane that solves (1).

Each iterate of the RSGM computes a Riemannian subgradient $(\mathbf{I} - \boldsymbol{b}\boldsymbol{b}^\top)\widetilde{\boldsymbol{\mathcal{X}}}\operatorname{sign}(\widetilde{\boldsymbol{\mathcal{X}}}^\top\boldsymbol{b})$, which is computationally efficient compared with solving an LP. Moreover, RSGM has been proved to converge to a global solution at a linear rate with appropriate initialization in the single subspace case (Li et al., 2019; Zhu et al., 2019). Here, we extend this analysis to the UoH model and prove a linear convergence rate. Towards that goal, we measure the distance between any vector $\boldsymbol{b} \in \mathbb{S}^{D-1}$ and our target solution set $\{\pm\boldsymbol{b}_1\}$ by

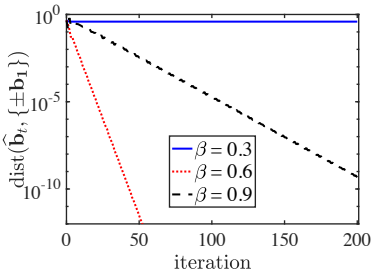

Figure 2: Linear convergence to $\mathcal{H}_1$ for different β in Algorithm 1. Here $D = 9$, $K = 3$, $N = 1200$ ($N_3 = 0.8N_2 = 0.8^2 N_1$), and outlier ratio $\frac{M}{M+N} = 0.3$.

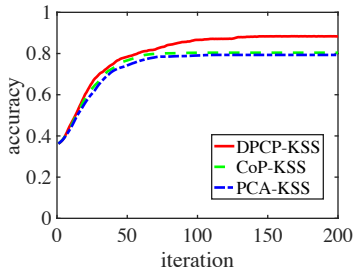

Figure 3: Mean clustering accuracy evolution over 100 independent experiments. Here $D = 9, K = 3, N_1 = N_2 = N_3 = 400$, and outlier ratio $\frac{M}{M+N} = 0.3$.

$\mathrm{dist}(\boldsymbol{b}, \{\pm \boldsymbol{b}_1\}) := \min(\|\boldsymbol{b} - \boldsymbol{b}_1\|_2, \|\boldsymbol{b} + \boldsymbol{b}_1\|_2)$. The next result establishes the Riemannian regularity condition (RRC) (Zhu et al., 2019) for (1), which we use to obtain a linear convergence rate.

Lemma 2 (Riemannian regularity condition (RRC))**.** *For any* $\epsilon \in (0, \sqrt{2(1 - 1/\zeta_1)})$ *and* $\tau = \frac{\sqrt{2}}{2} N_1 c_{\boldsymbol{\mathcal{X}}_1, min} \big((1 - \epsilon^2/2) - 1/\zeta_1\big)$ *with* ζ_1 *defined in* (6)*, the DPCP problem* (1) *satisfies the following* $(\tau, \epsilon, \boldsymbol{b}_1)$*-RRC: for every* $\boldsymbol{b} \in \mathbb{S}^{D-1}$ *satisfying* $\mathrm{dist}(\boldsymbol{b}, \{\pm \boldsymbol{b}_1\}) \le \epsilon$*, we have*

$$\begin{aligned} &\langle \mathrm{sign}(\boldsymbol{b}^\top \boldsymbol{b}_1)\boldsymbol{b}_1 - \boldsymbol{b}, -(\mathbf{I} - \boldsymbol{b}\boldsymbol{b}^\top)\widetilde{\boldsymbol{\mathcal{X}}}\,\mathrm{sign}(\widetilde{\boldsymbol{\mathcal{X}}}^\top \boldsymbol{b})\rangle \\ &\ge \tau\, \mathrm{dist}(\boldsymbol{b}, \{\pm \boldsymbol{b}_1\}). \end{aligned} \tag{12}$$

In words, (12) guarantees that when $\boldsymbol{b}$ is close to a target solution $\pm \boldsymbol{b}_1$ (a normal vector of the geometrically dominant hyperplane $\mathcal{H}_1$), the negative Riemannian subgradient points towards the target solution. The choice of ϵ and τ in Lemma 2 depends on the geometric dominance level of $\boldsymbol{\mathcal{X}}_1$. A larger relative dominance of $\boldsymbol{\mathcal{X}}_1$ (larger ζ_1) leads to larger ϵ (i.e., a larger initialization region) and larger τ (i.e., the negative Riemannain subgradient points closer to $\pm \boldsymbol{b}_1$). Using the RRC, we now apply Zhu et al. (2019, Theorem 1) to obtain a convergence guarantee for RSGM.

Theorem 3. *Let* $\{\widehat{\boldsymbol{b}}_t\}$ *be the sequence generated by Algorithm 1 for solving problem* (1) *with initialization* $\widehat{\boldsymbol{b}}_0$ *satisfying* $\widehat{\theta}_0 = \arccos(|\boldsymbol{b}_1^\top \widehat{\boldsymbol{b}}_0|) < \arccos(1/\zeta_1)$ *and step size* $\mu_t = \mu_0 \beta^t$ *such that*

$$0 < \mu_0 \le \frac{\tau \epsilon}{2\xi^2} \quad \text{and} \quad 1 > \beta \ge \sqrt{1 - 2\frac{\tau \mu_0}{\epsilon} + \frac{\mu_0^2 \xi^2}{\epsilon^2}}, \tag{13}$$

where $\epsilon = \sqrt{2(1 - \cos(\widehat{\theta}_0))}$,

$$\begin{aligned} &\tau = (\sqrt{2}/2) N_1 c_{\boldsymbol{\mathcal{X}}_1, min}\big(\cos(\widehat{\theta}_0) - 1/\zeta_1\big), \text{ and} \\ &\xi = \sqrt{\mathbf{1}^\top \boldsymbol{W}^{\max} \mathbf{1}} + \sum_{k=1}^{K} N_k \eta_{\boldsymbol{\mathcal{X}}_k} + M \eta_{\boldsymbol{\mathcal{O}}} + D. \end{aligned} \tag{14}$$

Then the principal angle $\widehat{\theta}_t$ *between* $\widehat{\boldsymbol{b}}_t$ *and* $\boldsymbol{b}_1$ *decays at a linear rate:* $\sin(\widehat{\theta}_t) \le \epsilon \cdot \beta^t$ *for all* $t \ge 0$.

Theorem 3 ensures that a properly initialized Algorithm 1 converges linearly to a normal vector of the geometrically dominant hyperplane $\mathcal{H}_1$, i.e., $\pm \boldsymbol{b}_1$, provided a certain diminishing step size is used. Note that Theorem 1 implies that $\pm \boldsymbol{b}_1$ are global solutions to (1) when condition (8) is satisfied. The initialization requirement coincides with Lemma 1, which states that any critical point inside the cone $\mathcal{C} = \{\boldsymbol{y} \in \mathbb{R}^D : |\boldsymbol{y}^\top \boldsymbol{b}_1| > 1/\zeta_1, \|\boldsymbol{y}\|_2 = 1\}$ must be normal vectors of $\mathcal{H}_1$ (see Figure 1). Note that β is crucial to the convergence properties of Algorithm 1: convergence may fail if β is too small, and convergence may be slow when β is too large. This is illustrated in Figure 2 for data sampled from the random model of Theorem 2, the initial step size is $\mu_0 = 0.01$, and a spectral initialization is used (the bottom eigenvectors of $\widetilde{\boldsymbol{\mathcal{X}}}\widetilde{\boldsymbol{\mathcal{X}}}^\top$, which were shown to be appropriate in practice (Zhu et al., 2019)).

Computational complexity. Let T be the number of iterations, and $L := N + M$ be the total number of points. The computational complexity of RSGM is $O(TLD)$, which is preferable over the SVD-based IRLS solver whose complexity is $O(TLD^2)$, especially when the ambient dimension D is large.

4 HYPERPLANE CLUSTERING WITH DPCP

Recall that KSS alternates between assigning data points to clusters and fitting a hyperplane to each cluster. The previous discussion concentrated on the theory and algorithms for solving the DPCP problem (1) for a UoH, showing it recovers the geometrically dominant hyperplane. Inspired by the fact that condition (9) in Theorem 2 is likely to hold in the subspace estimation step of KSS (since we expect most of the points in the estimated cluster to belong to a *single* hyperplane), we use a family of KSS variants for hyperplane clustering. The better performance of the KSS approach over the sequential use of RANSAC was observed in Tsakiris and Vidal (2017c) where the DPCP problem was solved by IRLS. Aside from the standard KSS, we also consider the following two improved variants.

Table 2: Mean hyperplane clustering accuracy for different methods over 50 independent experiments.

	$D = 4$				$D = 9$			
	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
MKF	0.7937	0.6263	0.5548	0.4643	0.5840	0.3973	0.2949	0.2470
SCC	0.9445	0.9209	0.9093	0.8784	0.9126	0.5940	0.3138	0.2519
EnSC	0.7011	0.4912	0.3913	0.3254	0.6223	0.3996	0.3125	0.2540
SSC-ADMM	0.6801	0.4810	0.3795	0.3175	0.6683	0.4010	0.2999	0.2548
SSC-OMP	0.5707	0.4134	0.3291	0.2747	0.5267	0.3573	0.2732	0.2232
DPCP-KSS	**0.9834**	**0.9463**	0.8985	0.8103	0.9927	**0.9807**	0.8051	0.5004
CoP-KSS	0.9614	0.8747	0.8300	0.7630	0.9706	0.9358	0.8380	0.5110
PCA-KSS	0.9601	0.8623	0.8142	0.7461	0.9619	0.9243	0.8074	0.5130
DPCP-EKSS	**0.9889**	0.8807	**0.9778**	**0.9489**	**0.9938**	0.9517	0.4908	0.2920
CoP-EKSS	0.8278	0.8393	0.8772	0.7938	0.8271	0.7900	0.3706	0.2867
PCA-EKSS	0.8278	0.8274	0.8517	0.7542	0.8221	0.7539	0.3660	0.2868
DPCP-CoRe-KSS	0.9832	**0.9715**	**0.9561**	**0.9599**	**0.9928**	**0.9857**	**0.9784**	**0.9628**
CoP-CoRe-KSS	0.9612	0.8992	0.9065	0.8907	0.9706	0.9415	0.9258	0.9089
PCA-CoRe-KSS	0.9603	0.8981	0.8769	0.8586	0.9619	0.9370	**0.9278**	**0.9083**

Ensemble KSS (EKSS). The performance of KSS is sensitive to its initialization because the problem is non-convex. A practical approach is to repeat the process for multiple random initializations and then pick the best one, or combine the results together in a certain way. The Ensemble KSS (EKSS) (Lipor et al., 2018) constructs an affinity matrix whose (i, j)th entry is the number of times the ith and jth points are clustered together, and then applies spectral clustering to obtain clustering results.

Cooperative Re-initialization (CoRe) KSS. The Cooperative Re-initialization (CoRe) (Lane et al., 2019) framework optimizes a group of clustering results (replicas) by greedily swapping clusters between them to improve the overall quality. Both EKSS and CoRe expect the clustering in each replica to be partially correct, and that the same pattern of errors will not be made by all replicas. CoRe is capable of identifying bad clusters in a replica and swapping them with better alternatives by monitoring the change in the objective value, and hence it is observed to be more efficient than EKSS.

Since the above vaiants of KSS use PCA to fit a hyperplane to a cluster, we denote them as PCA-KSS, PCA-EKSS, and PCA-CoRe-KSS. To improve their performance, we replace PCA by our DPCP approach with RSGM (Algorithm 1) and denote these KSS variants by DPCP-KSS, DPCP-EKSS, and DPCP-CoRe-KSS. We also use the CoP (Rahmani and Atia, 2017a) to fit the hyperplane for each cluster, resulting in the three KSS variants CoP-KSS (Gitlin et al., 2018), CoP-EKSS (Lipor et al., 2018), and CoP-CoRe-KSS.

Synthetic Experiments. The data are generated based on the random model in Theorem 2. All results are obtained on a 64-bit machine with 2.6GHz Intel Core i7 CPU. We first test the effect of using PCA, DPCP, and CoP in KSS. The DPCP approach is implemented with RSGM (Algorithm 1), where the initial step size μ_0 is determined by using a backtracking line search during the first iteration and the diminishing factor β is fixed to be 0.9. Figure 3 shows the mean hyperplane clustering accuracy (over 100 independent experiments) versus iterations, with all methods using the same initialization. DPCP-KSS outperforms the others on the configuration, with average running times for DPCP-KSS, CoP-KSS, and PCA-KSS of 0.99s, 2.11s, and 0.20s, respectively.

Next, we compare the performance of the methods discussed above with other state-of-the-art subspace clustering algorithms that include MKF (Zhang et al., 2009), SCC (Chen and Lerman, 2009), SSC-ADMM (Elhamifar and Vidal, 2013), EnSC (You et al., 2016a), and SSC-OMP (You et al., 2016b). The test[4] uses $D = 4, 9$, $K = 2, 3, 4, 5$, $N = 50KD$ (each plane has the same number of points so that $N_k = 50D$), and $\frac{M}{M+N} = 0.3$. Since the KSS-style methods (without ensemble) are sensitive to initialization, we run them 10 times with random initializations until convergence (tolerance of 0.001) or 100 iterations is reached, and then select the best (i.e., the one with the lowest objective value). The CoRe methods operate directly on these 10 replicas to return an improved clustering result by aggregating the knowledge. For the EKSS-like methods, in each replica we run the KSS-style methods for only 10 iterations but build the affinity matrix based on 1000 such replicas, which is suggested in Lipor et al. (2018). Table 2 reports the mean clustering accuracy of the methods on 50 independent instances with the highest two scores in each column given in bold.

One can see that the SC methods EnSC, SSC-ADMM, and SSC-OMP, which are designed for the low-relative dimension setting, are among the least competitive for clustering *hyperplanes*. Also, MKF and SCC do not perform well. Among the other methods, we observe that within each scheme, algorithms that involve DPCP

[4]The ambient dimension D for the synthetic experiments follows Tsakiris and Vidal (2017c).

Table 3: Mean clustering error for KSS variants with different backbones on 89 annotated images of NYUdepthV2.

	KSS	CoRe-KSS	EKSS
DPCP	**10.2%**	**9.3%**	**8.0%**
PCA	12.4%	11.7%	10.8%
CoP	11.0%	10.8%	13.8%

(implemented by RSGM in Algorithm 1) almost always perform the best. As a result, in each column the best method is the one that uses DPCP as the internal solver for identifying the dominant hyperplane in a cluster. We made the conservative choice of fixing $\beta = 0.9$ in RSGM, which empirically works well but additional tuning for β would further improve performance. We find that with as little as 10 replicas, the methods built on the CoRe framework perform very well. We believe this is because CoRe is able to correct bad cluster estimates by swapping with other estimates.

Real Experiments. We further explore the performance of DPCP in hyperplane clustering using the real dataset NYUdepthV2 (Nathan Silberman and Fergus, 2012), which contains indoor RGB images of size $480 \times 640 \times 3$ together with depth information for each pixel. We use 89 annotated images from Tsakiris and Vidal (2017c), each of which can be transformed to 307,200 3D points and has dominant hyperplanes such as floors, walls and so on. For computational reasons, we perform superpixel representation where each image is segmented to about 1000 superpixels and the set of pixels corresponding to each superpixel is substituted by their median depth. Moreover, since the planes associated with an indoor scene are affine in $\mathbb{R}^3$, we use homogeneous coordinates by adding 1 at the fourth coordinate and normalize it to unit length in $\mathbb{R}^4$. Finally, since different superpixels represent different numbers of underlying pixels, we assign a weight to each superpixel according to its size.

We now compare the KSS variants with different backbones, namely PCA, CoP and DPCP, in clustering hyperplanes on 89 annotated images of NYUdepthV2. The parametric setting for each method is the same as for the synthetic experiments. Note that here we ignore the other general subspace clustering algorithms discussed in the synthetic experiments since they have been shown less competitive for the hyperplane clustering task (see Table 2). We first show in Table 3 the averaged clustering error for the KSS variants applied to the real data. One can see that a similar phenomenon appears as in the synthetic experiments, namely that the algorithm achieving the lowest mean clustering error is the one using DPCP as the internal solver for estimating the dominant hyperplane within each KSS framework. On the other hand, the KSS method is generally not comparable with CoRe-KSS or EKSS in this test. Finally, in Figure 4 we give visual comparisons of various approaches on clustering four hyperplanes from a 3D point cloud of NYUdepthV2.

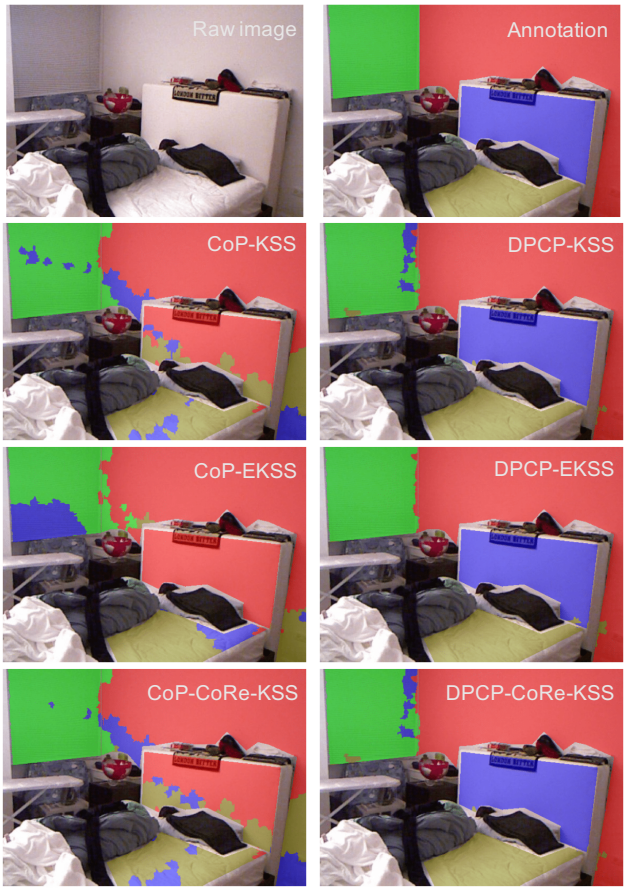

Figure 4: Visualization of various approaches in clustering four hyperplanes from a 3D point cloud of NYUdepthV2.

5 CONCLUSIONS

We considered the Dual Principal Component Pursuit (DPCP) for learning a union of hyperplanes (UoH). We provided a new geometric characterization as well as an interpretable probabilistic analysis on global minimizers of DPCP, which suggests that the solution is a normal vector to the geometrically dominant hyperplane. Moreover, we established the convergence guarantee for a scalable projected Riemannian subgradient method for solving DPCP for a UoH. By integrating DPCP into KSS (DPCP-KSS), and utilizing an ensemble of DPCP-KSS via EKSS or CoRe, we were able to achieve state-of-the-art performance in hyperplane clustering.

One could try to extend the analytical framework to a union of high dimensional *subspaces*, but the analysis would be significantly more complex since the geometry between the subspaces is no longer easily characterized by the principal angle between normal vectors. This topic will be the subject of future work.

Acknowledgements

This research is supported in part by NSF grant 1704458, 2008460 and ShanghaiTech grant 2017F0203-000-16. Tianyu Ding would like to thank Yunchen Yang and Tianjiao Ding for fruitful discussions and help with the real experiments.

References

Agarwal, P. K. and Mustafa, N. H. (2004). K-means projective clustering. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 155–165.

Aldroubi, A. and Tessera, R. (2011). On the existence of optimal unions of subspaces for data modeling and clustering. *Foundations of Computational Mathematics*, 11(3):363–379.

Bako, L. (2011). Identification of switched linear systems via sparse optimization. *Automatica*, 47(4):668–677.

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

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

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

Elhamifar, E. and Vidal, R. (2009). Sparse subspace clustering. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2790–2797. IEEE.

Elhamifar, E. and Vidal, R. (2013). Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2765–2781.

Fischler, M. A. and Bolles, R. C. (1981). Ransac random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 26:381–395.

Georgiev, P., Theis, F., and Cichocki, A. (2005). Sparse component analysis and blind source separation of underdetermined mixtures. *IEEE transactions on neural networks*, 16(4):992–996.

Gitlin, A., Tao, B., Balzano, L., and Lipor, J. (2018). Improving k-subspaces via coherence pursuit. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1575–1588.

He, Z. and Cichocki, A. (2007). An efficient k-hyperplane clustering algorithm and its application to sparse component analysis. In *International Symposium on Neural Networks*, pages 1032–1041. Springer.

Lane, C., Haeffele, B., and Vidal, R. (2019). Adaptive online k-subspaces with cooperative re-initialization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0.

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

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

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

Lerman, G. and Zhang, T. (2014). l_p-recovery of the most significant subspace among multiple subspaces with outliers. *Constructive Approximation*, 40(3):329–385.

Lerman, G., Zhang, T., et al. (2011). Robust recovery of multiple subspaces by geometric lp minimization. *The Annals of Statistics*, 39(5):2686–2715.

Li, X., Chen, S., Deng, Z., Qu, Q., Zhu, Z., and So, A. M. C. (2019). Weakly convex optimization over stiefel manifold using riemannian subgradient-type methods. *arXiv*, pages arXiv–1911.

Lipor, J., Hong, D., Zhang, D., and Balzano, L. (2018). Subspace clustering using ensembles of k-subspaces. *ArXiv*, abs/1709.04744v2.

Liu, G., Lin, Z., and Yu, Y. (2010). Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 663–670.

Lu, C.-Y., Min, H., Zhao, Z.-Q., Zhu, L., Huang, D.-S., and Yan, S. (2012). Robust and efficient subspace segmentation via least squares regression. In *European conference on computer vision*, pages 347–360. Springer.

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

Nathan Silberman, Derek Hoiem, P. K. and Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. In *ECCV*.

Rahmani, M. and Atia, G. K. (2017a). Coherence pursuit: Fast, simple, and robust principal component

analysis. *IEEE Transactions on Signal Processing*, 65(23):6260–6275.

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

Rahmani, M. and Li, P. (2019). Outlier detection and data clustering via innovation search. *arXiv preprint arXiv:1912.12988*.

Tron, R. and Vidal, R. (2007). A benchmark for the comparison of 3-d motion segmentation algorithms. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8.

Tsakiris, M. C. and Vidal, R. (2017a). Algebraic clustering of affine subspaces. *IEEE transactions on pattern analysis and machine intelligence*, 40(2):482–489.

Tsakiris, M. C. and Vidal, R. (2017b). Filtrated algebraic subspace clustering. *SIAM Journal on Imaging Sciences*, 10(1):372–415.

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

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

Vidal, R. (2011). Subspace clustering. *IEEE Signal Processing Magazine*, 28(2).

Vidal, R. and Favaro, P. (2014). Low rank subspace clustering (lrsc). *Pattern Recognition Letters*, 43:47–61.

Vidal, R., Ma, Y., and Sastry, S. (2005). Generalized principal component analysis (gpca). *IEEE transactions on pattern analysis and machine intelligence*, 27(12):1945–1959.

Vidal, R., Ma, Y., Soatto, S., and Sastry, S. (2006). Two-view multibody structure from motion. *International Journal of Computer Vision*, 68(1):7–25.

Vidal, R., Soatto, S., Ma, Y., and Sastry, S. (2003). An algebraic geometric approach to the identification of a class of linear hybrid systems. In *42nd IEEE International Conference on Decision and Control (IEEE Cat. No. 03CH37475)*, volume 1, pages 167–172. IEEE.

Vidal, R., Tron, R., and Hartley, R. (2008). Multiframe motion segmentation with missing data using power factorization and gpca. *International Journal of Computer Vision*, 79(1):85–105.

Xu, X., Zhong, M., and Guo, C. (2018). A hyperplane clustering algorithm for estimating the mixing matrix in sparse component analysis. *Neural Processing Letters*, 47(2):475–490.

Yang, H., Yang, X., Zhang, F., and Ye, Q. (2020). Robust plane clustering based on l1-norm minimization. *IEEE Access*, 8:29489–29500.

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

You, C., Robinson, D., and Vidal, R. (2016b). Scalable sparse subspace clustering by orthogonal matching pursuit. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3918–3927.

Zhang, T., Szlam, A., and Lerman, G. (2009). Median k-flats for hybrid linear modeling with many outliers. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 234–241. IEEE.

Zhu, Z., Ding, T., Robinson, D., Tsakiris, M., and Vidal, R. (2019). 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*, pages 9437–9447.

Zhu, Z., Wang, Y., Robinson, D., Naiman, D., Vidal, R., and Tsakiris, M. (2018a). Dual principal component pursuit: Improved analysis and efficient algorithms. In *Advances in Neural Information Processing Systems*, pages 2171–2181.

Zhu, Z., Wang, Y., Robinson, D. P., Naiman, D. Q., Vidal, R., and Tsakiris, M. C. (2018b). Dual principal component pursuit: probability analysis and efficient algorithms. *arXiv preprint arXiv:1812.09924*.