Noisy Dual Principal Component Pursuit

Tianyu Ding [*1] **Zhihui Zhu** [*2] **Tianjiao Ding** [3] **Yunchen Yang** [3] **René Vidal** [2] **Manolis C. Tsakiris** [3] **Daniel P. Robinson** [1]

Abstract

Dual Principal Component Pursuit (DPCP) is a recently proposed non-convex optimization based method for learning subspaces of high relative dimension from *noiseless* datasets contaminated by as many outliers as the square of the number of inliers. Experimentally, DPCP has proved to be robust to noise and outperform the popular RANSAC on 3D vision tasks such as road plane detection and relative pose estimation from three views. This paper extends the global optimality and convergence theory of DPCP to the case of data corrupted by noise, and further demonstrates its robustness using synthetic and real data.

1. Introduction

Dual Principal Component Pursuit (DPCP) is a recently proposed method for learning a linear subspace $\mathcal{S} \subset \mathbb{R}^D$ from a dataset $\widetilde{\boldsymbol{\mathcal{X}}} \in \mathbb{R}^{D\times L}$ contaminated by outliers (Tsakiris & Vidal, 2015; 2017; 2018a; Zhu et al., 2018a;b). Specifically, DPCP minimizes an ℓ_1 co-sparse objective on the sphere:

$$\min_{\boldsymbol{b}} \ \left\|\widetilde{\boldsymbol{\mathcal{X}}}^\top \boldsymbol{b}\right\|_1 \ \text{s.t.} \ \left\|\boldsymbol{b}\right\|_2 = 1. \tag{1}$$

The aim is to estimate a basis for the orthogonal complement of the subspace, hence the attribute *dual*. As such, DPCP is ideally suited for subspaces of *high relative dimension*, i.e., those subspaces with dimension d such that d/D is close to one. A typical example is the case of hyperplanes ($d = D - 1$), which very commonly appear in 3D computer vision applications such as detecting planar structures in 3D point clouds (Geiger et al., 2013; Silberman et al., 2012) or estimating relative poses in multiple-view geometry (Hartley & Zisserman, 2000).

[*]Equal contribution [1]Department of Applied Mathematics & Statistics, Johns Hopkins University, USA [2]Mathematical Institute for Data Science, Johns Hopkins University, USA [3]School of Information Science and Technology, ShanghaiTech University, China. Correspondence to: Tianyu Ding <tding1@jhu.edu>, Zhihui Zhu <zzhu29@jhu.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

In the high relative dimension regime, state-of-the-art convex optimization based methods relying on sparse and low-rank representations (Xu et al., 2010; Soltanolkotabi & Candès, 2012; Rahmani & Atia, 2017; You et al., 2017) typically exhibit a significant decrease in performance. On the other hand, since its inception almost 40 years ago, the Random Sampling And Consensus (RANSAC) (Fischler & Bolles, 1981) algorithm has been one of the most popular methods in computer vision for the high relative dimension setting. RANSAC alternates between fitting a subspace to a randomly sampled minimal number of points ($D - 1$ in the case of a hyperplane) and then using the number of data-points close to the subspace as a measure of the quality of the subspace. The interplay between four factors governs when RANSAC is successful: the ambient dimension, the outlier ratio, the thresholding parameter for determining when points are considered close to a subspace, and the allocated time budget. In particular, RANSAC can be extremely effective when the probability of sampling outlier-free samples inside the allocated time budget is large.

Recently (Tsakiris & Vidal, 2018a), it has been shown that an *Iteratively-Reweighted-Least-Squares* algorithm (DPCP-IRLS) for solving the non-convex DPCP problem (1) can successfully handle 30%–50% outliers in the three-view geometry problem, while state-of-the-art RANSAC variations fail when given the running time of DPCP-IRLS as a time budget. Even more recently (Zhu et al., 2018a), it has been demonstrated that a certain projected subgradient method (DPCP-PSGM) solves (1) to global optimality using only matrix-vector multiplications, and correctly performs road plane detection from a 3D cloud of approximately $O(10^5)$ points with 50% outliers in just a few hundred milliseconds, a time window in which RANSAC can only perform a few iterations and thus fails. These results highlight the significance of DPCP as a potential alternative to RANSAC.

In the terminology of the review paper of Lerman & Maunu 2018, DPCP is in effect a *least absolute deviations* subspace learning method. Such methods compute the subspace by aiming to minimize the sum of the distances between all points in the dataset and the subspace; this is precisely the formulation (1) when the subspace is a hyperplane. For example, *REAPER* (Lerman et al., 2015) applies a convex relaxation that is solved via an IRLS scheme (Zhang & Lerman, 2014; Zhang, 2016). Although REAPER is known

to perform competitively, the regime in which theoretical guarantees are ensured excludes the high relative dimensional setting, which we conjecture is a consequence of using a convex relaxation. The work closest to DPCP is that of Maunu et al. 2019, which studies a *Geodesic Gradient Descent (GGD)* method for solving the least absolute deviations problem without any relaxation. GGD is shown to converge to the global optimum at a sublinear rate, and to be able to handle $M = O(N)$ outliers with N inliers; the latter property is common to many robust PCA methods (Lerman & Maunu, 2018). In contrast, Zhu et al. 2018a showed that under a noiseless spherical statistical model, any global minimizer to (1) is a normal vector to the subspace as long as $M = O(N^2)$. Moreover, the DPCP-PSGM algorithm mentioned above provably converges to the global optimum of (1) in a piece-wise linear rate providing its step-size is tuned in a piece-wise geometrically diminishing fashion.

Although the theoretical and algorithmic features of DPCP are appealing, they have only been established for the idealized case when inliers perfectly lie in the subspace. Yet, DPCP has proved to be competitive on noisy real datasets, so that it is reasonable to ask whether similar theoretical guarantees hold when there is noise in the data. This work bridges that gap by making the following contributions.

- We provide a geometric analysis of global optimality for DPCP that reveals that global minimizers of (1) are perturbed away from the orthogonal complement $\mathcal{S}^\perp$ of the inlier subspace by an amount proportional to the noise level, while still tolerating $M = O(N^2)$ outliers.
- We prove that the DPCP-PSGM method, even in the presence of noise, converges to a neighborhood of $\mathcal{S}^\perp$ at a piece-wise linear rate, if tuned properly.
- Connections are drawn to the literature of absolute least deviations in subspace learning, where in particular we establish the equivalence between DPCP-PSGM and the GGD method of Maunu et al. 2019.
- An experiment on road plane detection with real 3D data further strengthens the view that DPCP is superior to RANSAC in the high relative dimension setting.

2. Global Optimality for Noisy DPCP

2.1. Background and motivation

Consider a unit ℓ_2-norm dataset $\widetilde{\boldsymbol{\mathcal{X}}}_{\boldsymbol{\mathcal{E}}} = \begin{bmatrix}\boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{E}} & \boldsymbol{\mathcal{O}}\end{bmatrix}\boldsymbol{\Gamma}$, where $\boldsymbol{\mathcal{X}} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N] \in \mathbb{R}^{D\times N}$ are inlier points spanning a single d-dimensional underlying subspace $\mathcal{S}$ of $\mathbb{R}^D$, $\boldsymbol{\mathcal{E}} = [\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_N] \in \mathbb{R}^{D\times N}$ consists of additive noise on inlier points, $\boldsymbol{\mathcal{O}} = [\boldsymbol{o}_1, \cdots, \boldsymbol{o}_M] \in \mathbb{R}^{D\times M}$ are outlier points and $\boldsymbol{\Gamma}$ is an unknown permutation. Our goal is to estimate the underlying subspace $\mathcal{S}$ from $\widetilde{\boldsymbol{\mathcal{X}}}_{\boldsymbol{\mathcal{E}}}$. When there is no noise (i.e., $\boldsymbol{\mathcal{E}} = \boldsymbol{0}$) and the points are in general position, the vectors $\boldsymbol{b}$ that make $\widetilde{\boldsymbol{\mathcal{X}}}_{\boldsymbol{\mathcal{E}}}^\top \boldsymbol{b}$ as sparse as possible are precisely those satisfying $\boldsymbol{b} \perp \mathcal{S}$; this is the motivation for (1). Therefore, in the noisy case we expect $\widetilde{\boldsymbol{\mathcal{X}}}_{\boldsymbol{\mathcal{E}}}^\top \boldsymbol{b}$ to be close to a sparse vector $\boldsymbol{y}$ in the Euclidean sense, whenever $\boldsymbol{b}$ is close to a normal vector of $\mathcal{S}$. This motivates the following optimization problem (Tsakiris & Vidal, 2018a)[1]:

$$\min_{\boldsymbol{b}\in\mathbb{S}^{D-1}, \boldsymbol{y}\in\mathbb{R}^{N+M}} \tau\|\boldsymbol{y}\|_1 + \tfrac{1}{2}\|\boldsymbol{y} - \widetilde{\boldsymbol{\mathcal{X}}}_{\boldsymbol{\mathcal{E}}}^\top \boldsymbol{b}\|_2^2, \tag{2}$$

for some $\tau > 0$, where $\mathbb{S}^{D-1} := \{\boldsymbol{b} \in \mathbb{R}^D : \|\boldsymbol{b}\|_2 = 1\}$. As expected, the performance of (2) depends crucially on the parameter τ. An alternative way is to directly adopt (1):

$$\min_{\boldsymbol{b}\in\mathbb{S}^{D-1}} \|\widetilde{\boldsymbol{\mathcal{X}}}_{\boldsymbol{\mathcal{E}}}^\top \boldsymbol{b}\|_1. \tag{3}$$

We use synthetic experiments to compare (2) and (3), with data generated according to the following random model:

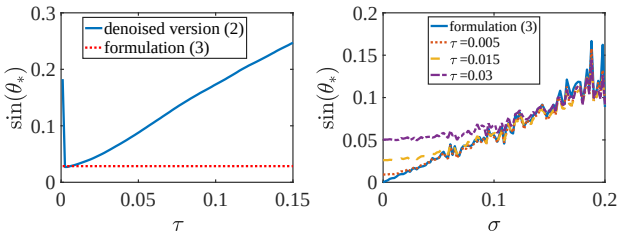
Definition 1 (Random spherical model)**.** *Consider a random spherical model where the columns of $\boldsymbol{\mathcal{O}}$ are drawn uniformly from the sphere $\mathbb{S}^{D-1}$, the columns of noisy inliers $\boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{E}}$ are drawn from the sphere $\mathbb{S}^{D-1}$ by normalizing i.i.d. $\mathcal{N}(\boldsymbol{0}, \frac{1}{d}\boldsymbol{\mathcal{P}}_{\mathcal{S}} + \frac{\sigma^2}{D}\boldsymbol{I}_D)$-distributed points to have unit ℓ_2-norm, where $d = \dim\mathcal{S}$, $\boldsymbol{\mathcal{P}}_{\mathcal{S}}$ is the ortho-projector onto $\mathcal{S}$, and σ is the standard deviation of the noise; under this model the SNR is $\mathbb{E}[\|\boldsymbol{\mathcal{X}}\|_F]/\mathbb{E}[\|\boldsymbol{\mathcal{E}}\|_F] = 1/\sigma$.*

Fig. 1 shows the numerical comparison between (2) and (3). We solve (2) by alternating minimization, which empirically converges fast even though it has no known convergence theory. We use DPCP-PSGM (Algorithm 1), whose convergence analysis will be discussed in Section 3, for solving (3). Fig. 1a shows that even though τ is chosen to be optimal, (2) and (3) perform competitively. Fig. 1b implies that the formulation (3), which does not depend on any hyperparameter, is robust to noise, whereas the solution of (2) is sensitive to τ. We have observed similar phenomena for different D, d, M, N and σ. Based on this evidence, in this paper we focus on (3).

Algorithm 1 DPCP-PSGM for (3)

Input: data $\widetilde{\boldsymbol{\mathcal{X}}}_{\boldsymbol{\mathcal{E}}} \in \mathbb{R}^{D\times(N+M)}$ and initial step size μ_0.
Initialization: Set $\widehat{\boldsymbol{b}}_0 \leftarrow \arg\min_{\boldsymbol{b}\in\mathbb{S}^{D-1}} \|\widetilde{\boldsymbol{\mathcal{X}}}_{\boldsymbol{\mathcal{E}}}^\top \boldsymbol{b}\|_2$.
for $k = 1, 2, \ldots$ **do**
  Compute sub-gradient: $\boldsymbol{g}_{k-1} \leftarrow \widetilde{\boldsymbol{\mathcal{X}}}_{\boldsymbol{\mathcal{E}}}\operatorname{sign}(\widetilde{\boldsymbol{\mathcal{X}}}_{\boldsymbol{\mathcal{E}}}^\top \widehat{\boldsymbol{b}}_{k-1})$.
  Update the step size μ_k according to a certain rule.
  Compute next iteration: $\boldsymbol{b}_k \leftarrow \widehat{\boldsymbol{b}}_{k-1} - \mu_k \boldsymbol{g}_{k-1}$.
  Project $\boldsymbol{b}_k$ onto the unit sphere: $\widehat{\boldsymbol{b}}_k \leftarrow \boldsymbol{b}_k/\|\boldsymbol{b}_k\|_2$.
end for

[1] Problem (2) has also appeared in the context of dictionary learning, see Qu et al. 2014.



(a) $\sigma = 0.05$ (b) Varying σ

Figure 1. Comparison between computed solutions of (2) and (3) in terms of their principal angle θ^* from $\mathcal{S}^\perp$. Here, we fix $D = 30, d = 29, N = 500$, and outlier ratio $M/(M+N) = 0.7$.

2.2. Geometric quantities and their concentrations

We aim to provide a global optimality analysis for (3). Since the objective in (3) is not continuously differentiable, we need to deal with its subdifferential. Denote the sign function by $\operatorname{sign}(a) = a/|a|$ when $a \neq 0$, and $\operatorname{sign}(0) = 0$, and denote the subdifferential of the absolute value function $|a|$ by $\operatorname{Sgn}(a) = \operatorname{sign}(a)$ when $a \neq 0$, and $\operatorname{Sgn}(0) = [-1, 1]$. We also apply sign and Sgn element-wise to vectors.

To analyze (3), first note that any global solution $\boldsymbol{b}^*$ to (3) must be a critical point, i.e., there exists $\boldsymbol{v}^* \in \partial\|\widetilde{\boldsymbol{\mathcal{X}}}_{\boldsymbol{\mathcal{E}}}^\top \boldsymbol{b}^*\|_1$ such that $(\boldsymbol{I} - \boldsymbol{b}^*\boldsymbol{b}^{*\top})\boldsymbol{v}^* = \boldsymbol{0}$, where

$$\partial\|\widetilde{\boldsymbol{\mathcal{X}}}_{\boldsymbol{\mathcal{E}}}^\top \boldsymbol{b}\|_1 = (\boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{E}})\operatorname{Sgn}((\boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{E}})^\top \boldsymbol{b}) + \boldsymbol{\mathcal{O}}\operatorname{Sgn}(\boldsymbol{\mathcal{O}}^\top \boldsymbol{b}). \quad (4)$$

When noise is not present (i.e., $\boldsymbol{\mathcal{E}} = \boldsymbol{0}$), the term $\operatorname{Sgn}\left((\boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{E}})^\top \boldsymbol{b}\right) = \operatorname{Sgn}(\boldsymbol{\mathcal{X}}^\top \boldsymbol{b})$ is simple since it only relates to inliers. In the noisy case, however, it is much more complicated to deal with this term. For example, since the function sign is discontinuous, $\operatorname{Sgn}\left((\boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{E}})^\top \boldsymbol{b}\right)$ cannot easily be separated into two parts with one part only involving the inliers and the other part only involving the noise. As a consequence, a significantly more technical analysis is required to analyze the effect of noise.

We now introduce several helpful geometric quantities. We first characterize the maximum norm of a Riemannian subgradient of $\frac{1}{M}\|\boldsymbol{\mathcal{O}}^\top \boldsymbol{b}\|_1$:

$$\eta_{\boldsymbol{\mathcal{O}}} := \frac{1}{M} \max_{\boldsymbol{b} \in \mathbb{S}^{D-1}} \left\|(\boldsymbol{I} - \boldsymbol{b}\boldsymbol{b}^\top)\boldsymbol{\mathcal{O}} \operatorname{sign}(\boldsymbol{\mathcal{O}}^\top \boldsymbol{b})\right\|_2. \quad (5)$$

As it turns out, $\eta_{\boldsymbol{\mathcal{O}}}$ characterizes how well the outliers are distributed in the ambient space: more uniformly distributed outliers will lead to smaller value for $\eta_{\boldsymbol{\mathcal{O}}}$ (Zhu et al., 2018a). To facilitate an analysis, we decompose the noise as $\boldsymbol{\mathcal{E}} = \boldsymbol{\mathcal{E}}_{\boldsymbol{s}} + \boldsymbol{\mathcal{E}}_{\boldsymbol{n}}$, where $\boldsymbol{\mathcal{E}}_{\boldsymbol{s}}$ is the projection of the noise onto $\mathcal{S}$ and $\boldsymbol{\mathcal{E}}_{\boldsymbol{n}}$ is the projection of the noise onto $\mathcal{S}^\perp$. Denote $\widehat{\boldsymbol{\mathcal{X}}} := \boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{E}}_{\boldsymbol{s}}$ and $\widehat{\boldsymbol{\mathcal{E}}} := \boldsymbol{\mathcal{E}}_{\boldsymbol{n}}$ such that the columns of $\widehat{\boldsymbol{\mathcal{X}}}$ lie in $\mathcal{S}$ and the columns of $\widehat{\boldsymbol{\mathcal{E}}}$ lie in $\mathcal{S}^\perp$. $\widehat{\boldsymbol{\mathcal{X}}}$ can be viewed as effective inliers since they lie in $\mathcal{S}$, whereas $\widehat{\boldsymbol{\mathcal{E}}}$ can be interpreted as effective noise because it perturbs $\widehat{\boldsymbol{\mathcal{X}}}$ away from $\mathcal{S}$. Define the *permeance statistic* (Lerman et al., 2015)

$$c_{\widehat{\boldsymbol{\mathcal{X}}},\min} := \frac{1}{N} \min_{\boldsymbol{b} \in \mathcal{S} \cap \mathbb{S}^{D-1}} \left\|\widehat{\boldsymbol{\mathcal{X}}}^\top \boldsymbol{b}\right\|_1, \quad (6)$$

which attains larger values for better distributed inliers. We capture the effective noise $\widehat{\boldsymbol{\mathcal{E}}}$ via the quantity

$$c_{\widehat{\boldsymbol{\mathcal{E}}},\max} := \frac{1}{N} \max_{\boldsymbol{b} \in \mathcal{S}^\perp \cap \mathbb{S}^{D-1}} \left\|\widehat{\boldsymbol{\mathcal{E}}}^\top \boldsymbol{b}\right\|_1. \quad (7)$$

This is closely related to the *total inlier residual* $\mathscr{R}(\mathcal{S}) := \frac{1}{N}\sum_{j=1}^N \|\widehat{\boldsymbol{\epsilon}}_j\|_2$ used by Lerman et al. 2015 to measure the level of the effective noise. By the Cauchy-Schwartz inequality $|\widehat{\boldsymbol{\epsilon}}_j^\top \boldsymbol{b}| \leq \|\widehat{\boldsymbol{\epsilon}}_j\|_2 \|\boldsymbol{b}\|_2$, it is clear that $c_{\widehat{\boldsymbol{\mathcal{E}}},\max}$ is a lower bound of $\mathscr{R}(\mathcal{S})$ since $\|\boldsymbol{b}\|_2 = 1$. Indeed, $\mathscr{R}(\mathcal{S})$ only depends on the energy of $\widehat{\boldsymbol{\mathcal{E}}}$, whereas $c_{\widehat{\boldsymbol{\mathcal{E}}},\max}$ also depends on the distribution of $\widehat{\boldsymbol{\mathcal{E}}}$: the more uniformly distributed $\widehat{\boldsymbol{\mathcal{E}}}$ is in $\mathcal{S}^\perp$, the smaller $c_{\widehat{\boldsymbol{\mathcal{E}}},\max}$ is. Thus, $c_{\widehat{\boldsymbol{\mathcal{E}}},\max}$ leads to a tighter result in our analysis than if one used $\mathscr{R}(\mathcal{S})$.

Note that $c_{\widehat{\boldsymbol{\mathcal{X}}},\min}$ involves a mixture of inliers and components of noise projected onto $\mathcal{S}$. This particular integration of inliers and noise leads to tighter deterministic bounds in the deterministic phase of our analysis. In turn, this will be advantageous in the subsequent probabilistic analysis (Lerman et al., 2015; Tsakiris & Vidal, 2018b).

We recall the probabilistic result for $\eta_{\boldsymbol{\mathcal{O}}}$ from Zhu et al. 2018a;b and provide new bounds for $c_{\widehat{\boldsymbol{\mathcal{X}}},\min}$, $c_{\widehat{\boldsymbol{\mathcal{E}}},\max}$. Define $\delta : [0,1) \to \mathbb{R}$ and $\rho : [0,1) \to \mathbb{R}$ as

$$\delta(\sigma) := \sqrt{\sigma} + \sqrt{(1-\sigma)F_{d,D-d}(\sigma)}, \quad (8)$$

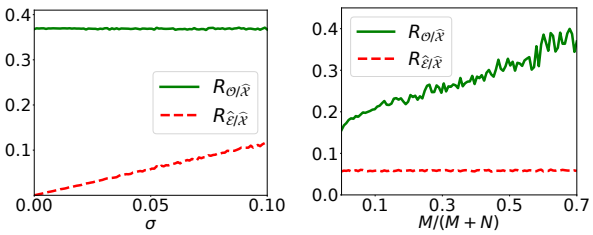
$$\rho(\sigma) := (1-\sigma)F_{D-d,d}(1/\sigma), \quad (9)$$

where $F_{d_1,d_2}(\cdot)$ is the cumulative density function (CDF) of the F-distribution with $F_{d_1,d_2}(0) = 0$ and $F_{d_1,d_2}(\infty) = 1$. Expanding the CDFs, we have $\delta(\sigma) = O(\sigma^{d/4} + \sqrt{\sigma})$ and $\rho(\sigma) = 1 - O(\sigma + \sigma^{d/2})$. We now state our first result.

Lemma 1. *Consider the random spherical model in Definition 1 and let $\sigma < 1$. For any $t > 0$, there exists a universal constant C independent of M, N, D, d, t, and σ such that*

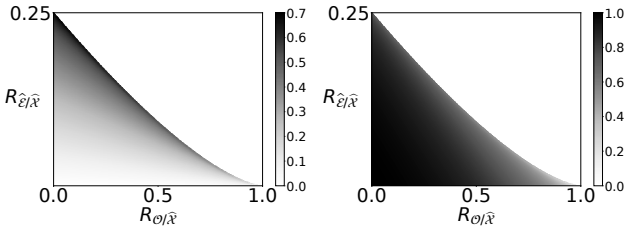
$$\begin{aligned}
&\mathbb{P}\left[\eta_{\boldsymbol{\mathcal{O}}} \leq C(\sqrt{D}\log D + t)/\sqrt{M})\right] \geq 1 - 2e^{-\frac{t^2}{2}}, \\
&\mathbb{P}\left[c_{\widehat{\boldsymbol{\mathcal{X}}},min} \leq \sqrt{2/\pi d}\rho(\sigma) - (2 + t/2)/\sqrt{N}\right] \leq 2e^{-\frac{t^2}{2}}, \\
&\mathbb{P}\left[c_{\widehat{\boldsymbol{\mathcal{E}}},max} \geq (1 + 2/\sqrt{N})\delta(\sigma) + t/\sqrt{N}\right] \leq 2e^{-\frac{t^2}{2}}. \quad (10)
\end{aligned}$$

Although our result for $c_{\widehat{\boldsymbol{\mathcal{X}}},\min}$ reduces to the one for $c_{\boldsymbol{\mathcal{X}},\min}$ in Zhu et al. 2018a when $\boldsymbol{\mathcal{E}} = \boldsymbol{0}$, the concentration derivation for $c_{\widehat{\boldsymbol{\mathcal{X}}},\min}$ is more involved since in the noisy case and under the above spherical statistical model, the columns of $\widehat{\boldsymbol{\mathcal{X}}}$ now lie *inside* the unit sphere as opposed to *on* the sphere.



(a) Varying σ (b) Varying outlier ratio

Figure 2. Plots of $R_{\mathcal{O}/\widehat{\mathcal{X}}}$ and $R_{\widehat{\mathcal{E}}/\widehat{\mathcal{X}}}$ as a function of (a) σ and (b) outlier ratio. Here we fix $D = 30, d = 29, N = 1500$, and $M/(M+N) = 0.7$ in (a), and $\sigma = 0.05$ in (b).



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

Figure 3. Plot of (a) t_1 and (b) t_2 when varying $R_{\mathcal{O}/\widehat{\mathcal{X}}}$ and $R_{\widehat{\mathcal{E}}/\widehat{\mathcal{X}}}$. In each plot, condition (12) holds only in the area below the curve, which corresponds to valid pairs of $(R_{\mathcal{O}/\widehat{\mathcal{X}}}, R_{\widehat{\mathcal{E}}/\widehat{\mathcal{X}}})$.

Since $\delta(\sigma) = O(\sigma^{d/4} + \sqrt{\sigma})$, (10) essentially implies that $c_{\widehat{\mathcal{E}},\max} = O(\sigma^{d/4} + \sqrt{\sigma})$ with high probability [2].

Two more definitions are needed for our analysis:

$$R_{\mathcal{O}/\widehat{\mathcal{X}}} := \frac{M}{N}\frac{\overline{\eta}_{\mathcal{O}}}{c_{\widehat{\mathcal{X}},\min}}, \quad R_{\widehat{\mathcal{E}}/\widehat{\mathcal{X}}} := \frac{c_{\widehat{\mathcal{E}},\max}}{c_{\widehat{\mathcal{X}},\min}}, \tag{11}$$

where $\overline{\eta}_{\mathcal{O}} := \eta_{\mathcal{O}} + D/M$. $R_{\mathcal{O}/\widehat{\mathcal{X}}}$ and $R_{\widehat{\mathcal{E}}/\widehat{\mathcal{X}}}$ can be simply viewed as outlier-to-inlier and noise-to-inlier type of ratios, respectively. When we fix inliers and outliers, $R_{\widehat{\mathcal{E}}/\widehat{\mathcal{X}}}$ is proportional to the noise level (see Fig. 2a). Similarly, when we fix inliers and noise level, $R_{\mathcal{O}/\widehat{\mathcal{X}}}$ is proportional to the number of outliers (see Fig. 2b).

2.3. Geometry of the critical points

For the rest of the analysis, let $\theta \in [0, \pi/2]$ be the principal angle of a vector $\boldsymbol{b} \in \mathbb{S}^{D-1}$ from the orthogonal complement subspace $\mathcal{S}^{\perp}$. Thus, $\boldsymbol{b}$ is normal to $\mathcal{S}$ if and only if $\theta = 0$. Using $R_{\mathcal{O}/\widehat{\mathcal{X}}}$ and $R_{\widehat{\mathcal{E}}/\widehat{\mathcal{X}}}$ defined in (11), we can now characterize the geometry of the critical points of (3).

Lemma 2. *Assume* $R_{\mathcal{O}/\widehat{\mathcal{X}}} < 1$ *and*

$$\frac{32 R_{\widehat{\mathcal{E}}/\widehat{\mathcal{X}}}}{\left(\sqrt{R^2_{\mathcal{O}/\widehat{\mathcal{X}}}+8} - 3R_{\mathcal{O}/\widehat{\mathcal{X}}}\right)^{\frac{3}{2}}\left(\sqrt{R^2_{\mathcal{O}/\widehat{\mathcal{X}}}+8} + R_{\mathcal{O}/\widehat{\mathcal{X}}}\right)^{\frac{1}{2}}} < 1, \tag{12}$$

then any critical point $\boldsymbol{b}^*$ *of problem* (3) *must have its principal angle* θ^* *from* $\mathcal{S}^{\perp}$ *satisfy:*

$$\theta^* \le \sin^{-1}(t_1) \quad \text{or} \quad \theta^* \ge \sin^{-1}(t_2), \tag{13}$$

where $0 \le t_1 \le t_2 \le 1$ *are the two nonnegative roots of the following quartic equation:*

$$t^4 + (R^2_{\mathcal{O}/\widehat{\mathcal{X}}} - 1)t^2 + 4R_{\mathcal{O}/\widehat{\mathcal{X}}} R_{\widehat{\mathcal{E}}/\widehat{\mathcal{X}}} t + 4R^2_{\widehat{\mathcal{E}}/\widehat{\mathcal{X}}} = 0. \tag{14}$$

First note that $R_{\mathcal{O}/\widehat{\mathcal{X}}} < 1$ ensures that the denominator of the left hand side in (12) is well-defined. Since the function $a \mapsto f(a) = \left(\sqrt{a^2+8} - 3a\right)^{\frac{3}{2}}\left(\sqrt{a^2+8} + a\right)^{\frac{1}{2}}$ is decreasing between $[0, 1]$ with $f(0) = 8$ and $f(1) = 0$, (12) implies that larger noise levels lead to smaller numbers of outliers that DPCP can tolerate. With (12), it can be shown that (14) has two nonnegative roots $0 \le t_1 \le t_2 \le 1$, and (13) implies that none of the critical points have principal angle in $(\sin^{-1}(t_1), \sin^{-1}(t_2))$. Fig. 3 displays t_1 and t_2 while varying $R_{\mathcal{O}/\widehat{\mathcal{X}}}$ and $R_{\widehat{\mathcal{E}}/\widehat{\mathcal{X}}}$ under condition (12). One can observe that smaller percentages of outliers and noise levels lead to t_1 being closer to 0 and t_2 being closer to 1, which means that critical points of (3) either lie in a neighborhood of $\mathcal{S}^{\perp}$ or very close to $\mathcal{S}$. The following bound helps in further interpreting Lemma 2:

$$t_1 \le 25 R_{\widehat{\mathcal{E}}/\widehat{\mathcal{X}}}/(1 - R_{\mathcal{O}/\widehat{\mathcal{X}}})^2. \tag{15}$$

In particular, this means that t_1 is close to 0 when $R_{\widehat{\mathcal{E}}/\widehat{\mathcal{X}}}$ and $R_{\mathcal{O}/\widehat{\mathcal{X}}}$ are small. More generally, for fixed $\mathcal{O}$ and $\widehat{\mathcal{X}}$, (15) guarantees that t_1 is perturbed away from 0 by at most the effective noise level, which makes sense intuitively.

When there is no noise ($\mathcal{E} = \mathbf{0}$), Lemma 2 reduces to Lemma 1 in Zhu et al. 2018a: $R_{\widehat{\mathcal{E}}/\widehat{\mathcal{X}}} = 0$ and $R_{\mathcal{O}/\widehat{\mathcal{X}}} = \overline{\eta}_{\mathcal{O}}/c_{\mathcal{X},\min}$, so that (12) always holds and (14) becomes $t^4 + ((\overline{\eta}_{\mathcal{O}}/c_{\mathcal{X},\min})^2 - 1)t^2 = 0$, which implies $t_1 = 0$ and $t_2 = \sqrt{1 - (\overline{\eta}_{\mathcal{O}}/c_{\mathcal{X},\min})^2}$. Nevertheless, we stress that the proof for Lemma 2 is far more complicated than for the noiseless case, partly because of the need to deal with $\operatorname{Sgn}\left((\mathcal{X} + \mathcal{E})^\top \boldsymbol{b}\right)$ as per the discussion right after (4).

2.4. Global optimality

Before characterizing the global solutions of (3), we recall two outlier-based quantities

$$c_{\mathcal{O},\min} := \frac{1}{M}\min_{\boldsymbol{b}\in\mathbb{S}^{D-1}} \left\|\mathcal{O}^\top \boldsymbol{b}\right\|_1, c_{\mathcal{O},\max} := \frac{1}{M}\max_{\boldsymbol{b}\in\mathbb{S}^{D-1}} \left\|\mathcal{O}^\top \boldsymbol{b}\right\|_1$$

that are already used by Zhu et al. 2018a;b and scales as $O(\frac{1}{\sqrt{M}})$. We note that better distributed outliers lead to smaller values of $c_{\mathcal{O},\max} - c_{\mathcal{O},\min}$. The next result gives a condition under which global solutions to (3) lie in a neighborhood of $\mathcal{S}^{\perp}$.

[2] We note that $\frac{t}{2\sqrt{N}}$ in (10) may be improved to a quantity proportional to σ using a more sophisticated analysis.

Theorem 1. *If $R_{\boldsymbol{\mathcal{O}}/\widehat{\boldsymbol{\mathcal{X}}}} < 1$, (12) holds, and*

$$\frac{M}{N}\frac{c_{\boldsymbol{\mathcal{O}},max} - c_{\boldsymbol{\mathcal{O}},min}}{c_{\widehat{\boldsymbol{\mathcal{X}}},min}} < t_2 - 2R_{\widehat{\boldsymbol{\mathcal{E}}}/\widehat{\boldsymbol{\mathcal{X}}}}, \tag{16}$$

then any global minimizer $\boldsymbol{b}^*$ of (3) must have its principal angle θ^* from $\mathcal{S}^\perp$ satisfy

$$\theta^* \le \sin^{-1}(t_1), \tag{17}$$

where $0 \le t_1 \le t_2 \le 1$ are the nonnegative roots of (14).

Theorem 1 builds upon Lemma 2, with the intuition that critical points that are close to the subspace $\mathcal{S}$ (i.e., for which $\theta^* \ge \sin^{-1}(t_2)$) cannot be global minimizers as they result in large objective values. As long as data points are well-distributed (small $c_{\boldsymbol{\mathcal{O}},\max} - c_{\boldsymbol{\mathcal{O}},\min}$, large $c_{\widehat{\boldsymbol{\mathcal{X}}},\min}$, large t_2) and effective noise is mild (small $c_{\widehat{\boldsymbol{\mathcal{E}}},\max}$), (16) will be satisfied and global minimizers must be close to $\mathcal{S}^\perp$. When $\boldsymbol{\mathcal{E}} = \boldsymbol{0}$, we have already remarked that $t_1 = 0$ and $t_2 = \sqrt{1-(\overline{\eta}_{\boldsymbol{\mathcal{O}}}/c_{\boldsymbol{\mathcal{X}},\min})^2}$, which together with (16) and (17) imply that global minimizers are orthogonal to $\mathcal{S}$ when

$$\frac{M}{N}\frac{c_{\boldsymbol{\mathcal{O}},\max} - c_{\boldsymbol{\mathcal{O}},\min}}{c_{\boldsymbol{\mathcal{X}},\min}} < \sqrt{1-(\overline{\eta}_{\boldsymbol{\mathcal{O}}}/c_{\boldsymbol{\mathcal{X}},\min})^2},$$

which is precisely Theorem 1 of Zhu et al. 2018a.

We further interpret the above global optimality result, via the following probabilistic characterization.

Theorem 2. *Consider the random spherical model of Definition 1 and let $\sigma < 1$. If $0 < t < 2\left(\sqrt{\frac{2N}{\pi d}}\rho(\sigma) - 2\right)$, then with probability at least $1 - 10e^{-t^2/2}$, any global solution to (3) must have its principal angle θ^* from $\mathcal{S}^\perp$ satisfy*

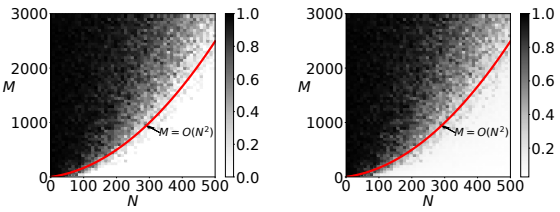
$$\sin(\theta^*) \le \frac{C_1\delta(\sigma) + \frac{t}{2\sqrt{N}}}{\sqrt{\frac{2}{\pi d}}\rho(\sigma) - C_2\frac{t\sqrt{M}+\sqrt{DM}\log D}{N} - \frac{4+t}{\sqrt{N}}} \tag{18}$$

as long as

$$M\Big((4\sqrt{2}+\sqrt{2}t)^2 + C_3(\sqrt{D}\log D + t)^2\Big) \le N^2\Big(\frac{1}{\sqrt{\pi d}}\rho(\sigma) - C_4\delta(\sigma) - \frac{4+3t}{2\sqrt{2N}}\Big)^2, \tag{19}$$

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

The effect of the noise in perturbing the global solution away from $\mathcal{S}^\perp$ is captured by (18), where the right hand side (RHS) approaches 0 when $\sigma \to 0$, except for the small term $\frac{t}{2\sqrt{N}}$, which as we stated after Lemma 1 can be improved to a quantity proportional to σ. Moreover, (18) together with $\delta(\sigma) = O(\sigma^{d/4} + \sqrt{\sigma})$ and $\rho(\sigma) = 1 - O(\sigma + \sigma^{d/2})$ imply that $\sin(\theta^*) = O((\sqrt{\sigma} + \sigma^{d/4}))$ when σ is small. The inequality (19) suggests that, unlike existing state-of-the-art $O(N)$ outlier bounds (Lerman et al., 2015; Maunu et al., 2019), DPCP can tolerate $O(N^2)$ outliers even for noisy data. Fig. 4 verifies this point by plotting $\sin(\theta^*)$ ($\boldsymbol{b}^*$ is computed via Algorithm 1).



(a) $\sigma = 0$ (b) $\sigma = 0.1$

Figure 4. Plot of $\sin(\theta^*)$ where θ^* is the principal angle between the computed solution $\boldsymbol{b}^*$ to the DPCP problem (3) and $\mathcal{S}^\perp$ when varying N and M for noise level (a) $\sigma = 0$ and (b) $\sigma = 0.1$. Here $D = 30$ and $d = 29$.

3. Convergence Analysis of Noisy DPCP-PSGM

The convergence of DPCP-PSGM (Algorithm 1) has been analyzed by Zhu et al. 2018a;b in the absence of noise. Their main finding is that selecting the step size in a piecewise geometrically diminishing fashion guarantees piecewise linear convergence to a vector orthogonal to $\mathcal{S}$. In the noisy case, one can only expect Algorithm 1 to converge to a neighborhood of $\mathcal{S}^\perp$. A significantly more involved analysis yields the following convergence result.

Theorem 3 (Piecewise linear convergence)**.** *Suppose that*

$$Nc_{\widehat{\boldsymbol{\mathcal{X}}},min} - N\eta_{\widehat{\boldsymbol{\mathcal{X}}},\widehat{\boldsymbol{\mathcal{E}}}} - \tfrac{5}{2}Nc_{\widehat{\boldsymbol{\mathcal{E}}},max} - \tfrac{7}{2}\sqrt{N}\|\widehat{\boldsymbol{\mathcal{E}}}\|_2 > M\eta_{\boldsymbol{\mathcal{O}}}, \tag{20}$$

where $\eta_{\widehat{\boldsymbol{\mathcal{X}}},\widehat{\boldsymbol{\mathcal{E}}}} = \frac{1}{N}\sup_{(c,\boldsymbol{b},\boldsymbol{g})\in\mathbb{F}} \big\|(\boldsymbol{I}-\boldsymbol{b}\boldsymbol{b}^\top)\widehat{\boldsymbol{\mathcal{X}}}\,\mathrm{sign}(\widehat{\boldsymbol{\mathcal{X}}}^\top\boldsymbol{b} + c\cdot\widehat{\boldsymbol{\mathcal{E}}}^\top\boldsymbol{g})\big\|_2$ with $\mathbb{F} := \{(c,\boldsymbol{b},\boldsymbol{g}) : c \in [0,\infty), \boldsymbol{g} \in \mathcal{S}^\perp \cap \mathbb{S}^{D-1}, \boldsymbol{b} \in \mathcal{S} \cap \mathbb{S}^{D-1}\}$.[3] Let $\{\widehat{\boldsymbol{b}}_k\}$ be the sequence generated by Algorithm 1 with noisy data and initialization $\widehat{\boldsymbol{b}}_0$ whose principal angle from the orthogonal subspace $\mathcal{S}^\perp$ satisfies

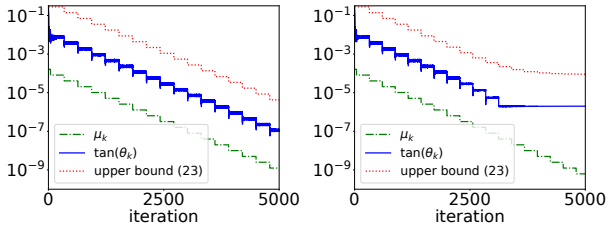
$$\theta_0 < \tan^{-1}\big(\kappa_{\widetilde{\boldsymbol{\mathcal{X}}}\boldsymbol{\mathcal{E}}}/\nu_{\widetilde{\boldsymbol{\mathcal{X}}}\boldsymbol{\mathcal{E}}}\big), \tag{21}$$

where $\kappa_{\widetilde{\boldsymbol{\mathcal{X}}}\boldsymbol{\mathcal{E}}} := Nc_{\widehat{\boldsymbol{\mathcal{X}}},min} - Nc_{\widehat{\boldsymbol{\mathcal{E}}},max} - \sqrt{N}\|\widehat{\boldsymbol{\mathcal{E}}}\|_2$ and $\nu_{\widetilde{\boldsymbol{\mathcal{X}}}\boldsymbol{\mathcal{E}}} := N\eta_{\widehat{\boldsymbol{\mathcal{X}}},\widehat{\boldsymbol{\mathcal{E}}}} + M\eta_{\boldsymbol{\mathcal{O}}} + \sqrt{N}\|\widehat{\boldsymbol{\mathcal{E}}}\|_2$. Consider the following piecewise geometrically diminishing step size

$$\mu_k = \begin{cases} \mu_0, & k < K_0, \\ \mu_0\beta^{\lfloor (k-K_0)/K\rfloor + 1}, & k \ge K_0, \end{cases} \tag{22}$$

where $\mu_0 \le \mu' := \frac{1/16}{\max\{Nc_{\widehat{\boldsymbol{\mathcal{X}}},max}, Mc_{\boldsymbol{\mathcal{O}},max}, \sqrt{N}\|\widehat{\boldsymbol{\mathcal{X}}}\|_2\}}$, $\beta < 1$, $\lfloor\cdot\rfloor$ is the floor function, $K_0, K \in \mathbb{N}$ are chosen such that

[3] $\eta_{\widehat{\boldsymbol{\mathcal{X}}},\widehat{\boldsymbol{\mathcal{E}}}} = O(\frac{1}{\sqrt{N}})$ for the random spherical model.



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

Figure 5. Convergence of DPCP-PSGM in both noiseless and noisy case. θ_k is the principal angle of iterate $\widehat{\boldsymbol{b}}_k$ from $\mathcal{S}^\perp$. The red dotted line represents the upper bound on $\tan(\theta_k)$ given by (23), while the green dashed line indicates the step size (22).

$$K \geq \frac{2}{\beta\mu'(Nc_{\widehat{\boldsymbol{X}},min} - \nu_{\widetilde{\boldsymbol{X}}_{\boldsymbol{\mathcal{E}}}})} \text{ and } K_0 \geq \frac{\tan(\theta_0)}{\mu_0 \cdot \Delta} \text{ with}$$

$$\Delta := \min\left\{\kappa_{\widetilde{\boldsymbol{X}}_{\boldsymbol{\mathcal{E}}}} - \tan(\theta_0)\nu_{\widetilde{\boldsymbol{X}}_{\boldsymbol{\mathcal{E}}}},\ \tfrac{1}{6}(Nc_{\widehat{\boldsymbol{X}},min} - \nu_{\widetilde{\boldsymbol{X}}_{\boldsymbol{\mathcal{E}}}})\right\} > 0.$$

Then the principal angle θ_k of $\widehat{\boldsymbol{b}}_k$ from $\mathcal{S}^\perp$ satisfies

$$\tan(\theta_k) \leq \frac{\mu_0}{\mu'}\beta^{\lfloor (k-K_0)/K \rfloor} + \tan(\theta') \text{ for } k \geq K_0, \quad (23)$$

$$\tan(\theta_k) \leq \max\{\tan(\theta_0),\ \tfrac{\mu_0}{\mu'} + \tan(\theta')\} \text{ for } k < K_0 \text{ with}$$

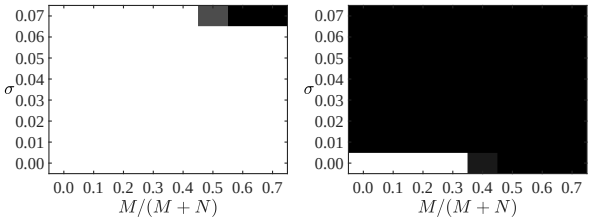
$$\theta' := \tan^{-1}\left(\frac{2(Nc_{\widehat{\boldsymbol{\mathcal{E}}},max} + \sqrt{N}\|\widehat{\boldsymbol{\mathcal{E}}}\|_2)}{Nc_{\widehat{\boldsymbol{X}},min} - \nu_{\widetilde{\boldsymbol{X}}_{\boldsymbol{\mathcal{E}}}}}\right). \quad (24)$$

Under the random spherical model and for small noise levels, the main hypothesis of Theorem 3, (20) is satisfied as long as there are at most $M = O(N^2)$ outliers. In that regime, Theorem 3 guarantees that DPCP-PSGM converges to a neighborhood of $\mathcal{S}^\perp$ providing data points are well-distributed (small $\eta_{\widehat{\boldsymbol{X}},\widehat{\boldsymbol{\mathcal{E}}}}$, small $\eta_{\boldsymbol{\mathcal{O}}}$, large $c_{\widehat{\boldsymbol{X}},\text{min}}$) and the effective noise is mild (small $c_{\widehat{\boldsymbol{\mathcal{E}}},\text{max}}$ and $\|\widehat{\boldsymbol{\mathcal{E}}}\|$): (23) implies that the principal angle θ_k decays in a piecewise linear rate until θ', which reflects the noise effect; see Fig. 5b. Also, larger noise levels lead to larger $c_{\widehat{\boldsymbol{\mathcal{E}}},\text{max}}$ and $\|\widehat{\boldsymbol{\mathcal{E}}}\|_2$, and thus to larger θ'. If no noise is present, $\theta' = 0$ and Theorem 3 is consistent with Zhu et al. 2018a; see Fig. 5a.

We note that for the sake of interpretability the final angle θ' in (24) has intentionally been made looser than the analytical bound θ^* in (17). This causes no harm since it can be shown that both angles scale as $\delta(\sigma)$. Finally, the spectral initialization in Algorithm 1 can be shown to satisfy (21) subject to some further mild conditions on the data.

4. Comparison with state-of-the-art

As noted in §1, DPCP is very closely related to least-absolute-deviations subspace learning methods. Two important representatives of that class are REAPER (Lerman et al., 2015) and the Geodesic-Gradient-Descent (GGD) method of Maunu et al. 2019. For the case of a hyperplane, GGD attempts to solve the same optimization problem as DPCP, while REAPER a convex relaxation of it. In this section we compare the results of this paper to those known for REAPER and GGD. We show that the global optimality conditions for DPCP given in the present paper are much looser compared to those required for REAPER (§4.1). In fact, they are an improvement even over conditions enabling a local stability characterization of the function landscape given by Maunu et al. 2019 (§4.2). Finally, we show that for a suitable tuning of the step-size GGD is equivalent to DPCP-PSGM (§4.3).



(a) This paper (b) Lerman et al. 2015

Figure 6. Check whether (a) (16) and (b) (26) are satisfied (white) or not (black) when varying the outlier ratio $M/(M+N)$ and σ. Here we fix $D = 30$, $d = 29$, and $N = 1000$.

4.1. Comparison with REAPER (Lerman et al., 2015)

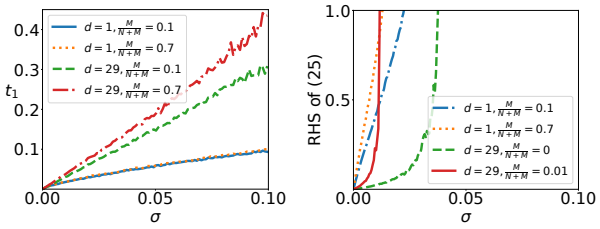
Theorem 2.1 of Lerman et al. 2015 asserts that any global minimizer of the REAPER problem must satisfy

$$\sin(\theta_*) \leq \frac{2N\mathscr{R}(\mathcal{S})}{\left[\frac{N}{4\sqrt{d}}c_{\widehat{\boldsymbol{X}},\text{min}} - M\mathscr{A}(\mathcal{S}) - N\mathscr{R}(\mathcal{S})\right]_+}, \quad (25)$$

where $\mathscr{R}(\mathcal{S}) := \frac{1}{N}\sum_{j=1}^N \|\widehat{\boldsymbol{\epsilon}}_j\|_2 \geq c_{\widehat{\boldsymbol{\mathcal{E}}},\text{max}}$ is the total inlier residual, $\mathscr{A}(\mathcal{S}) := \frac{1}{M}\|\boldsymbol{\mathcal{O}}\|_2\|\boldsymbol{\mathcal{P}}_{\mathcal{S}^\perp}\boldsymbol{\mathcal{O}}\|_2 \geq 0$ is an *alignment statistic* that measures the amount of linear structure in the outliers, and $[\alpha]_+ = \alpha$ if $\alpha > 0$ and 0 otherwise. Here $\boldsymbol{\mathcal{P}}_{\mathcal{S}^\perp}$ is the orthoprojection onto $\mathcal{S}^\perp$ and the overline spherization operator normalizes the columns of a matrix. Note that (25) is meaningful only when

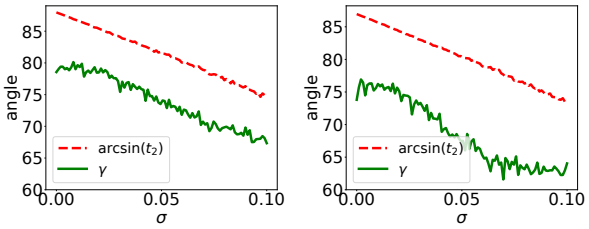
$$\frac{M\mathscr{A}(\mathcal{S})}{Nc_{\widehat{\boldsymbol{X}},\text{min}}} < \frac{1}{4\sqrt{d}} - R_{\widehat{\boldsymbol{\mathcal{E}}}/\widehat{\boldsymbol{X}}}. \quad (26)$$

We compare the necessary condition (26) for REAPER to (12) and (16) for DPCP. When there are no outliers (26) requires $R_{\widehat{\boldsymbol{\mathcal{E}}}/\widehat{\boldsymbol{X}}} < \frac{1}{4\sqrt{d}}$. By contrast, (12) only requires $R_{\widehat{\boldsymbol{\mathcal{E}}}/\widehat{\boldsymbol{X}}} < \frac{1}{4}$ (see Fig. 3). More generally, in the presence of outliers, $M\mathscr{A}(\mathcal{S})$ (the numerator in LHS of (26)) scales as $O(M)$ in a random model (Lerman et al., 2015), whereas the quantity $M(c_{\boldsymbol{\mathcal{O}},\text{max}} - c_{\boldsymbol{\mathcal{O}},\text{min}})$ (the numerator in LHS of (16))



(a) This paper (b) Lerman et al. 2015

Figure 7. Evaluation of (a) t_1 and (b) (25), with $D = 30$ and $N = 1500$. In (b) for $d = 29$, we only plot (25) for $\frac{M}{M+N} \in \{0, 0.01\}$ since (26) does not hold for a mild size of the outlier ratio.



(a) $M/(M+N) = 0.1$ (b) $M/(M+N) = 0.2$

Figure 8. Comparison between the quantity γ of Maunu et al. 2019 and $\arcsin(t_2)$ in the hyperplane case with outlier ratio (a) 0.1 and (b) 0.2. Here we fix $D = 30$, $d = 29$, and $N = 3000$.

scales as $O(\sqrt{M})$; see Zhu et al. 2018b. Numerically, this is captured in Fig. 6, in which we observe that (16) is satisfied for a much larger range of outlier ratio and noise levels. Finally, note that $\mathscr{R}(\mathcal{S})$ appears both in the numerator and denominator in the RHS of (25), which makes the entire upper bound blow up quickly when the noise level increases; see Fig. 7b. In contrast, according to Theorem 1 and (15), $\sin(\theta^*)$ is roughly proportional to the effective noise level (see Fig. 7a).

4.2. Comparison with the local optimality conditions of Maunu et al. 2019

Let $\boldsymbol{b}^*$ be a critical point of (3). Then, given $0 < \eta < \gamma < \pi/2$ such that a certain *stability condition* holds, Theorem 2 of Maunu et al. 2019 asserts that either

$$\theta^* < \eta \quad \text{or} \quad \theta^* > \gamma. \tag{27}$$

Note that a tighter analysis corresponds to a smaller η (closer to 0) and a larger γ (closer to $\pi/2$). To fairly compare (27) and (13) numerically, we manually set η equal to $\arcsin(t_1)$ and compare $\arcsin(t_2)$ and γ. Fig. 8 shows the comparison between γ and $\arcsin(t_2)$ under different percentages of outliers and noise levels. We can observe that $\arcsin(t_2)$ is always larger than γ by a significant amount, under the restriction that η is equal to $\arcsin(t_1)$, thus suggesting that (13) is a tighter result compared with (27). Moreover, (27) is sensitive to the variation of the outliers, while (13) is rather stable (compare Fig. 8a to Fig. 8b). Finally, we mention that the relationship between η and γ is not as clear as for our t_1 and t_2, with the latter being the two non-negative roots of the univariate quartic in (14). In conclusion, we believe that Lemma 2 represents a theoretical and computational improvement over the important characterization of the critical points of (3) given previously by Maunu et al. 2019.

4.3. Equivalence with GGD of Maunu et al. 2019

We now relate DPCP-PSGM to the GGD of Maunu et al. 2019. Towards that end, we consider the hyperplane case and interpret the GGD as finding a normal vector instead of a basis for the hyperplane. Rewriting the subgradient in Algorithm 1 as $\boldsymbol{g}_{k-1} = \widetilde{\boldsymbol{g}}_{k-1} + \overline{\boldsymbol{g}}_{k-1}$, where $\widetilde{\boldsymbol{g}}_{k-1} = (\boldsymbol{I} - \widehat{\boldsymbol{b}}_{k-1}\widehat{\boldsymbol{b}}_{k-1}^\top)\boldsymbol{g}_{k-1}$ is the Riemannian subgradient and $\overline{\boldsymbol{g}}_{k-1} = \widehat{\boldsymbol{b}}_{k-1}\widehat{\boldsymbol{b}}_{k-1}^\top\boldsymbol{g}_{k-1}$, the GGD is the same as Algorithm 1 except that the iterate is updated by $\widehat{\boldsymbol{b}}_k^\natural = \cos(\mu_k^\natural)(\widehat{\boldsymbol{b}}_{k-1} - \tan(\mu_k^\natural)\widetilde{\boldsymbol{g}}_{k-1}/\|\widetilde{\boldsymbol{g}}_{k-1}\|_2)$, where $\mu_k^\natural$ is the step size used in GGD. To relate $\widehat{\boldsymbol{b}}_k^\natural$ to $\widehat{\boldsymbol{b}}_k$ in Algorithm 1, we first rewrite $\boldsymbol{b}_k = (1 - \mu_k \boldsymbol{g}_{k-1}^\top \widehat{\boldsymbol{b}}_{k-1})(\widehat{\boldsymbol{b}}_{k-1} - \frac{\mu_k \widetilde{\boldsymbol{g}}_{k-1}}{1-\mu_k \boldsymbol{g}_{k-1}^\top \widehat{\boldsymbol{b}}_{k-1}})$. Noting that $\widehat{\boldsymbol{b}}_k$ is obtained by normalizing $\boldsymbol{b}_k$, we have $\widehat{\boldsymbol{b}}_k^\natural = \widehat{\boldsymbol{b}}_k$ if we set $\mu_k^\natural = \tan^{-1}(\frac{\mu_k \|\widetilde{\boldsymbol{g}}_{k-1}\|_2}{1-\mu_k \boldsymbol{g}_{k-1}^\top \widehat{\boldsymbol{b}}_{k-1}})$.

Proposition 1. *For the hyperplane case, with a suitable choice of step size the GGD (Maunu et al., 2019) is equivalent to DPCP-PSGM.*

Thus, the convergence guarantee in Theorem 3 can be directly applied for GGD under a suitable choice of step size. For the general case where the subspace has co-dimension larger than 1, we conjecture that the analysis in Theorem 3 will be helpful for the convergence analysis of GGD.

5. Road Plane Estimation Using Real 3D Data

We use the experimental setup of Zhu et al. 2018a to further compare DPCP, RANSAC, and alternative methods in the task of 3D road plane detection. In this problem one is given a 3D point cloud of a road scene and the goal is to learn an affine plane $\mathcal{A} = \mathcal{H} + \boldsymbol{t} \subset \mathbb{R}^3$ as a model for the road. This is important in autonomous driving applications. Here $\mathcal{H}$ is a plane through the origin with normal vector $\boldsymbol{b}$ and $\boldsymbol{t}$ is its translation with respect to the origin; this latter is the center of the laser sensor. Hence the task is to estimate $\boldsymbol{b}$ and $\boldsymbol{t}$, which are taken to be co-linear in order to resolve the inherent ambiguity in estimating $\boldsymbol{t}$. In turn, this can be converted to a linear subspace learning problem by working in homogeneous coordinates, i.e., by embedding $\mathcal{A}$ into the linear hyperplane $\bar{\mathcal{H}} \subset \mathbb{R}^4$ with normal vector $\bar{\boldsymbol{b}} = [\boldsymbol{b}^\top \ -\boldsymbol{t}^\top \boldsymbol{b}]^\top$, through the mapping $\boldsymbol{x} \mapsto [\boldsymbol{x}^\top \ 1]^\top$.

Methods/metric	ROC	$\hat{\bar{\theta}}$	$\hat{\theta}$	$\hat{t}$	iter.	time
SVD	0.76	4.40	1.73	14%	N/A	1
RANSAC×1	0.78	3.74	4.18	12%	3.8	31
RANSAC×10	0.91	1.58	2.85	5%	18.7	149
RANSAC×100	**0.93**	**1.47**	2.77	**4%**	64.1	515
$\ell_{2,1}$-RPCA	0.77	4.35	1.72	14%	2.8	30
REAPER	0.88	2.48	1.07	8%	4.1	27
DPCP-IRLS	0.81	3.67	1.48	12%	3.0	29
DPCP-d	0.92	1.51	0.82	5%	6.5	16
DPCP-PSGM	0.92	1.59	**0.76**	5%	37.3	24

Table 1. 3D road plane estimation using 125 annotated frames of the KITTI dataset. Running time is in msec.

We use the 3D point clouds from the KITTI dataset (Geiger et al., 2013). In addition to the 7 frames annotated in Zhu et al. 2018a, we further annotate 131 frames. Each point cloud contains around 10^5 points with approximately 50% outliers. The data are homogenized and normalized to unit ℓ_2-norm. We compare DPCP-PSGM (Algorithm 1), DPCP-IRLS and DPCP-d (Tsakiris & Vidal, 2018a) to RANSAC, REAPER and $\ell_{2,1}$-RPCA (Xu et al., 2010). We also include SVD, which calculates $\bar{\boldsymbol{b}}$ as the bottom singular vector of the data. Since DPCP-PSGM and DPCP-d are among the fastest methods with comparable running times, we let them run to convergence, and set the running time of the slowest as time budget for the rest methods. We update the step size in DPCP-PSGM via a modified backtracking line search, which is known to perform well in practice. For RANSAC, we also include a version with $10\times$ and $100\times$ that time budget. We tune the parameters of the algorithms on a randomly selected training set of 13 frames and use the rest of the frames for evaluation. Each method is tuned to achieve an optimal error and then re-tuned to be as fast as possible without exceeding 5% of that error. The λ of $\ell_{2,1}$-RPCA is set to $\frac{1.92}{\sqrt{M}}$, the τ of DPCP-d is set to $\frac{2.76}{\sqrt{N+M}}$, μ_{min} for DPCP-PSGD is set to 10^{-9}, and the relative convergence accuracy, wherever applicable, is set to 10^{-6}.

Table 1 reports geometric, clustering and algorithmic metrics for the various methods. Once a method has computed an estimated normal vector $\hat{\bar{\boldsymbol{b}}} \in \mathbb{R}^4$, we extract from it estimates $\hat{\boldsymbol{b}}, \hat{\boldsymbol{t}}$. We report the corresponding estimation errors, i.e., the angle $\hat{\bar{\theta}}$ between $\bar{\boldsymbol{b}}^*$ and $\hat{\bar{\boldsymbol{b}}}$, the angle $\hat{\theta}$ between $\boldsymbol{b}^*$ and $\hat{\boldsymbol{b}}$, and $100\left\|\boldsymbol{t}^* - \hat{\boldsymbol{t}}\right\|_2 / \left\|\boldsymbol{t}^*\right\|_2 \%$, where $\bar{\boldsymbol{b}}^*, \boldsymbol{b}^*, \boldsymbol{t}^*$ are the ground-truth values. By varying a threshold on the distances of all points to the estimated affine plane, the area under the ROC curve is obtained[4], with higher values indicating better performance. Finally, the number of iterations executed by each method and its running time in msec[5] are also reported. Notably, not only DPCP-PSGM and DPCP-d outperform RANSAC×1 and RANSAC×10, rather their performance is comparable with that of RANSAC×100, which they further surpass, e.g., in estimating the orientation of the normal vector $\boldsymbol{b}^*$: RANSAC×100 is off by 2.77° on average, while DPCP-PSGM and DPCP-d only by 0.76° and 0.82° respectively; see also Fig. 9. On the other hand, DPCP-IRLS and REAPER make heavy use of SVD, which makes them slow to run on $\mathcal{O}(10^5)$ points, and eventually inaccurate given the limited time budget.

[4]For RANSAC this is also its internal thresholding parameter.

[5]Experiments done on a laptop with Intel i7-6700HQ @ 2.6GHz CPU, 16GB 2133MHz DDR4 Memory.

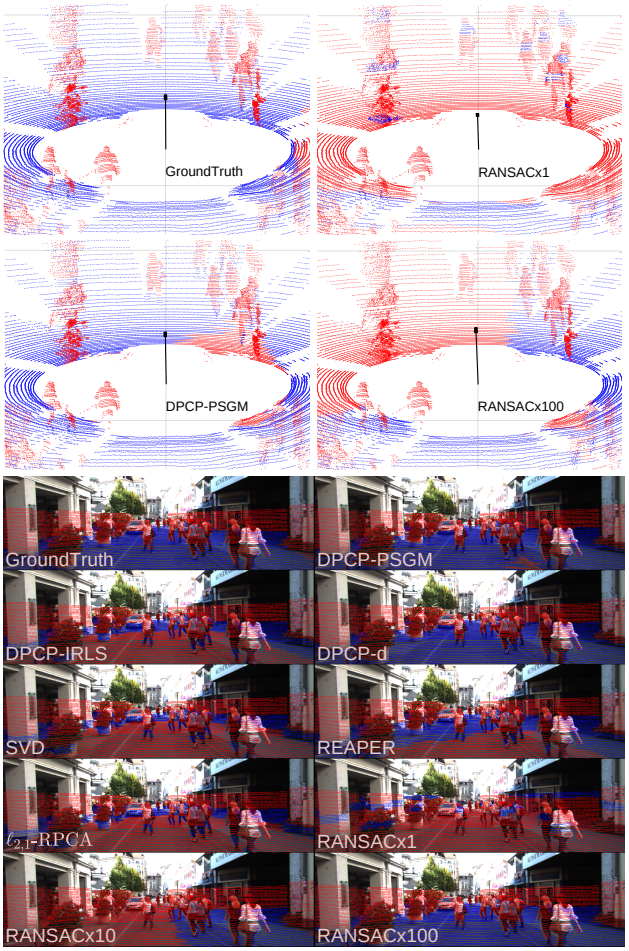


Figure 9. Frame 328 of KITTY-CITY-71, with inliers in blue and outliers in red. Top: 3D point clouds and estimated translations. Ground-truth thresholding parameter is used for RANSAC. Bottom: Projections of 3D point clouds onto the image.

Acknowledgment

The co-authors from JHU were supported by NSF grant 1704458. The co-authors from ShanghaiTech were supported by ShanghaiTech grant 2017F0203-000-16.

References

Chakraborty, R., Hauberg, S., and Vemuri, B. C. Intrinsic grassmann averages for online linear and robust subspace learning. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Chiang, K.-Y., Dhillon, I. S., and Hsieh, C.-J. Using side information to reliably learn low-rank matrices from missing and corrupted observations. *Journal of Machine Learning Research*, 19(76):1–35, 2018.

Dai, B., Wang, Y., Aston, J., Hua, G., and Wipf, D. Connections with robust pca and the role of emergent sparsity in variational autoencoder models. *Journal of Machine Learning Research*, 19(41):1–42, 2018.

Diakonikolas, I., Kamath, G., Kane, D., and Li, J. Being robust (in high dimensions) can be practical. *International Conference on Machine Learning*, 2017.

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

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

Hartley, R. and Zisserman, A. *Multiple View Geometry in Computer Vision*. Cambridge, 2000.

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

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.

Marinov, T. V., Mianjy, P., and Arora, R. Streaming principal component analysis in noisy setting. *International Conference on Machine Learning*, 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.

Qu, Q., Sun, J., and Wright, J. Finding a sparse vector in a subspace: Linear sparsity using alternating directions. In *Advances in Neural Information Processing Systems*, pp. 3401–3409, 2014.

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

Soltanolkotabi, M. and Candès, E. J. A geometric analysis of subspace clustering with outliers. *Annals of Statistics*, 40(4):2195–2238, 2012.

Tsakiris, M. and Vidal, R. Dual principal component pursuit. *ICCV Workshop on Robust Subspace Learning and Computer Vision*, pp. 10–18, 2015.

Tsakiris, M. and Vidal, R. Hyperplane clustering via dual principal component pursuit. *International Conference on Machine Learning*, 2017.

Tsakiris, M. and Vidal, R. Dual principal component pursuit. *Journal of Machine Learning Research*, 19(18): 1–50, 2018a.

Tsakiris, M. and Vidal, R. Theoretical analysis of sparse subspace clustering with missing entries. *International Conference on Machine Learning*, 2018b.

Xu, H., Caramanis, C., and Sanghavi, S. Robust pca via outlier pursuit. In *Neural Information Processing Systems*, 2010.

You, C., Robinson, D., and Vidal, R. Provable self-representation based outlier detection in a union of subspaces. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Zhang, T. Robust subspace recovery by tyler's m-estimator. *Information and Inference: A Journal of the IMA*, 5(1): 1–21, 2016.

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

Zhang, T. and Yang, Y. Robust pca by manifold optimization. *Journal of Machine Learning Research*, 19(80): 1–39, 2018.

Zhang, Y., Shi, D., Gao, J., and Cheng, D. Low-rank-sparse subspace representation for robust regression. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

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*, 2018a.

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