IMPLICIT BIAS OF PROJECTED SUBGRADIENT METHOD GIVES PROVABLE ROBUST RECOVERY OF SUBSPACES OF UNKNOWN CODIMENSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust subspace recovery (RSR) is a fundamental problem in robust representation learning. Although RSR has received considerable attention in the literature, there are still several aspects of RSR that remain highly unexplored. In particular, when the dimension of the underlying subspace is unknown and the data contains significant numbers of outlying entries many methods can struggle to identify the correct subspace. Here we focus on a recently proposed RSR method termed Dual Principal Component Pursuit (DPCP) approach, which aims to recover a basis of the orthogonal complement of the subspace. While prior work has shown that DPCP can provably recover the correct subspace in the presence of outliers, this relies on knowing the true dimension of the subspace, which is typically not possible in practice, and DPCP often fails when this dimension is unknown. Instead, we propose a very simple algorithm based on running multiple instances of a projected sub-gradient descent method (PSGM), with each problem instance seeking to find one vector in the null space of the subspace. Here we show that under mild conditions that this approach will succeed with high probability. In particular, we show that 1) all of the problem instances will converge to a vector in the null space of the subspace and 2) the ensemble of problem instance solutions will be sufficiently diverse to fully span the null space of the subspace (and thus reveal the true codimension of the subspace) even when the true subspace dimension is unknown. We provide empirical results that corroborate our theoretical results and showcase the remarkable implicit rank regularization behavior of PSGM algorithm that allows us to perform RSR without being aware of the subspace dimension.

1 INTRODUCTION

Robust subspace recovery (RSR) refers to methods designed to identify an underlying linear subspace (with dimension less than the ambient data dimension) in a dataset which is potentially corrupted with outliers (i.e., points that do not lie in the linear subspace). Many methods for RSR have been proposed in the literature over the past several years Xu et al. (2012); You et al. (2017a); Lerman & Maunu (2018). Formulations based on convex relaxations and decompositions of the data matrices into a low-rank matrices plus sparse corruptions – either uniformly at random as in Robust PCA (RPCA) Candès et al. (2011) or column-sparse corruptions as in Xu et al. (2012); McCoy & Tropp (2011)[1] – can, in certain situations, be shown to be provably recover the true subspace when the dimension is unknown. However, these theoretical guarantees often require the dimension of the subspace, d, to be significantly less than the ambient dimension of the data, D, and these methods are not suitable to the more challenging high relative subspace dimension regime (i.e., when $\frac{d}{D} \approx 1$).

The Dual Principal Component Pursuit approach for RSR. Recently, progress has been made towards solving the RSR problem in the high relative dimension regime by a formulation termed Dual Principal Component Pursuit (DPCP), which is provably robust in recovering subspaces of high relative dimension, Tsakiris & Vidal (2018). As implied by its name, DPCP follows a *dual* perspective of RSR aiming to recover a basis for the orthogonal complement of the inliers' subspace.

[1]Note that RPCA is closely related but distinct from RSR. In RPCA the corruptions are assumed to be entry-wise in a data matrix, while in RSR the corruptions apply to an entire datapoint (i.e., column-wise).

However, one of the key limitations of DPCP is that it requires *a priori* knowledge of the true subspace dimension.

Specifically, let $\tilde{\boldsymbol{X}} \in \mathbb{R}^{D\times(N+M)}$ denote the data matrix defined as $\tilde{\boldsymbol{X}} = [\boldsymbol{X} \quad \boldsymbol{O}]\boldsymbol{\Gamma}$ where $\boldsymbol{X} \in \mathbb{R}^{D\times N}$ is a matrix containing N inliers as its columns, $\boldsymbol{O} \in \mathbb{R}^{D\times M}$ is a matrix containing M outliers, and $\boldsymbol{\Gamma}$ is an unknown permutation matrix. DPCP was first formulated in Tsakiris & Vidal (2018) for handling subspaces of co-dimension $c = D - d$ equal to 1 (i.e., the subspace is a hyperplane with dimension $d = D - 1$) formulated as the following optimization problem:

$$\min_{\boldsymbol{b}\in\mathbf{R}^D} \quad \|\tilde{\boldsymbol{X}}^\top \boldsymbol{b}\|_1 \;\; \text{s.t.} \;\; \|\boldsymbol{b}\|_2 = 1. \tag{1}$$

Note that problem (1) is nonconvex due to the spherical constraint imposed on vector $\boldsymbol{b} \in \mathbb{S}^{D-1}$ which is a normal vector of the $D - 1$ dimensional hyperplane. In Tsakiris & Vidal (2018) the authors showed that the global minimizer of (1) will give a normal vector of the underlying true hyperplane as long as both inliers and outliers are well-distributed *or* the ratio between them is less than a deterministically derived upper bound as well as proposing an iteratively re-weighted least-squares (IRLS) algorithm. Following a probabilistic point of view, the authors in Zhu et al. (2018) presented an improved theoretical analysis of DPCP giving further insights on the remarkable tolerance of DPCP to recover the true underlying subspace even in datasets heavily corrupted by outliers. Moreover, the authors introduced a projected subgradient method which converges to a normal vector of the true subspace at a linear rate.

In addition, the authors of Zhu et al. (2018) also proposed an extension to DPCP which allows for subspaces with codimension $c > 1$ via a projected subgradient algorithm which attempted to learn c normal vectors to the subspace in a recursive manner. Specifically, after convergence to a normal vector, the projected subgradient algorithm is initialized with a vector *orthogonal* to the previously estimated normal vector. However, for that approach to be successful the knowledge of the true subspace codimension c becomes critical. Specifically, if an underestimate of the true codimension c is assumed the recovered basis for the null space, $\hat{\boldsymbol{B}}$, will fail to span the whole null space, $\mathcal{S}_\perp$. On the other hand, an overestimate of c will lead to columns of $\hat{\boldsymbol{B}}$ corresponding to vectors that lie in $\mathcal{S}$.

In Zhu et al. (2019), an alternative to (1) was proposed which attempts to solve for c normal vectors to the subspace at once via the formulation

$$\min_{\boldsymbol{B}\in\mathbb{R}^{D\times c}} \|\tilde{\boldsymbol{X}}^\top \boldsymbol{B}\|_{1,2} \quad \text{s.t.} \quad \boldsymbol{B}^\top\boldsymbol{B} = \mathbf{I}. \tag{2}$$

The authors also propose an optimization algorithm based on the projected Riemannian subgradient method (RSGM), which builds on similar ideas as the projected subgradient method of Zhu et al. (2018) and enjoys a linear converge rate when the step size is selected based on a geometrically diminishing rule. In Ding et al. (2021) a theoretical analysis is provided on the geometric properties of 2 showing the merits of the dual subspace recovery approach in handling datasets a) highly contaminated by outliers (in the order of $M = \mathcal{O}(N^2)$) and b) subspaces of high relative dimension. However again, it is critical to note that a key shortcoming of this approach is the fact that because all minimizers of problem (2) will be orthogonal matrices, a prerequisite for recovering the correct orthogonal complement of the inliers subspace is the *a priori* knowledge of the true co-dimension c (see Fig. 1).

Contributions. In this work, we address this key limitation by proposing a framework that allows us to perform robust subspace recovery in high relative subspace dimension regime **without** requiring *a priori* knowledge of the true subspace dimension. In particular, our proposed approach will be based on the simple approach of solving multiple, parallel instances of the DPCP formulation for solving for a single normal vector to the subspace,

$$\min_{\boldsymbol{B}\in\mathbb{R}^{D\times c'}} \sum_{i=1}^{c'} \|\tilde{\boldsymbol{X}}\boldsymbol{b}_i\|_1 \quad \text{s.t.} \quad \|\boldsymbol{b}_i\|_2 = 1, \quad i = 1, 2, \ldots, c' \tag{3}$$

where c' is assumed to be a safe overestimate of c i.e., $c' \geq c$. Contrary to (2), the objective function in (3) decouples over the columns $\boldsymbol{b}_i$ of matrix $\boldsymbol{B} = [\boldsymbol{b}_1 \; \boldsymbol{b}_2 \; \cdots \boldsymbol{b}_{c'}]$ and thus can be can be solved in a parallel manner by independently applying a projected subgradient algorithm (referred to as PSGM) from c' different *random initialization*. We then show that this simple strategy permits us to robustly recover the true subspace even without knowledge of the true codimension c since any overestimate

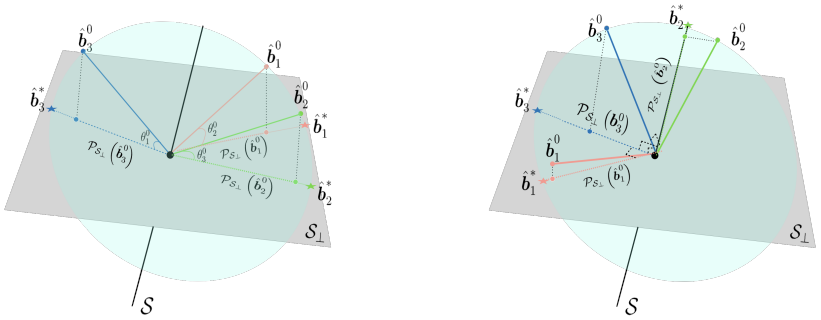

Figure 1: Graphical illustration of the recovered normal vectors of $\mathcal{S}$ by a) the proposed DPCP-PSGM approach (left) and b) methods imposing orthogonality constraints (right). Clearly, the rank of matrix $\hat{\boldsymbol{B}}^* = [\boldsymbol{b}_1^*, \boldsymbol{b}_2^*, \boldsymbol{b}_3^*]$ in a) equals to the true codimension $c = 2$ of $\mathcal{S}$ and $\text{span}(\boldsymbol{B}^*) \equiv \mathcal{S}_\perp$ as opposed to a) which gives rise to a $\boldsymbol{B}^*$ such that $\text{rank}(\boldsymbol{B}^*) = 3$ since $\boldsymbol{b}_2^* \in \mathcal{S}$.

will lead to redundant normal vectors of $\mathcal{S}$ and our recovered matrix $\hat{\boldsymbol{B}}$ that will span $\mathcal{S}_\perp$. As is detailed in Sections 3 and 4, this remarkable behavior of PSGM originates from the *implicit bias* that is induced in the optimization problem due to a) the relaxation of *orthogonality constraints* in (3) and b) the *random initialization* scheme that is adopted. Our specific contributions are as follows:

1. First, we focus on a continuous version of (3) i.e., in the case that inliers and outliers are distributed under continuous measures. We prove that DPCP problem formulated as in (3) can be solved via a projected subgradient algorithm that implicitly biases solutions towards low-rank matrices $\hat{\boldsymbol{B}} \in \mathbb{R}^{D \times c'}$ whose columns are the projections of the randomly initialized columns of $\boldsymbol{B}^0$ onto $\mathcal{S}_\perp$. As a result, $\hat{\boldsymbol{B}}$ spans $\mathcal{S}_\perp$ as long as it is *randomly initialized* with $c' \geq c$.
2. Second, we analyze the discrete version which is of more practical interest showing that iterates of DPCP-PSGM converge to a *scaled* and *perturbed* version of the initial matrix $\boldsymbol{B}^0$. This compelling feature of DPCP-PSGM allows to derive a sufficient condition and a probabilistic bound guaranteeing when the matrix $\hat{\boldsymbol{B}} \in \mathbb{R}^{D \times c'}$ spans $\mathcal{S}_\perp$.
3. We provide empirical results both on simulated and a real datasets, corroborating our theory and showing the robustness of our approach even without knowledge of the true subspace codimension.

2 Related work

Subspace Recovery. Learning underlying low-dimensional subspace representations of data has been a central topic of interest in machine learning research. Principal Component Analysis (PCA) has been the most celebrated method of this kind and is based on the minimization of the perpendicular distances of the data points from the estimates linear subspace, Jolliffe & Cadima (2016). Albeit, it is originally formulated as nonconvex optimization problem, PCA can be easily solved in closed form using a singular value decomposition (SVD) operation, Vidal et al. (2016). Despite its great success, PCA is prone to failure when handling datasets that contain *outliers* i.e., data points whose deviation from the inliers' subspace is "large" in the ℓ_2 norm sense.

Robust Subspace Recovery (RSR). To remedy this weakness of PCA *robust subspace recovery (RSR)* methods have come to the scene, Lerman & Maunu (2018); Maunu et al. (2019). The overall objective of RSR algorithms can be summarized as follows: "defy the outliers and correctly derive the low-dimensional subspace where clean data lie". Several strategies for "defying" outliers have been proposed giving rise to a wealth of RSR algorithms. Amongst the most popular ones is RANSAC Fischler & Bolles (1981), which is given a time budget and randomly chooses per iteration d points that span a d-dimensional subspace. RANSAC outputs the subspace that agrees with the largest number of points. The need to tackle inherent shortcomings of RANSAC pertaining to computational complexity issues inspired alternative convex formulations of RSR, Xu et al. (2012); You et al. (2017b); Rahmani & Atia (2017); Zhang & Lerman (2014). In Xu et al. (2012) the authors decompose the data matrix as a sum of a low-rank and a column-sparse component. However, theoretical guarantees obtained

for convex formulations hold for subspaces of relatively low-dimensional subspaces i.e., for $d \gg D$ where d and D denote the subspace and the ambient dimension, respectively. To the best of our knowledge, existing RSR algorithms rely heavily on a key assumption: *the subspace dimension is a priori known*. Undoubtedly, such a hypothesis is rather strong in real-world applications. Moreover, heuristic strategies for selecting the dimension of the subspace are hard to be applied in the RSR setting since they incur computationally prohibitive procedures, Lerman & Maunu (2018).

Relation to Orthogonal Dictionary Learning (ODL). Note that objective functions in the form of (3) show up beyond RSR problems i.e., in orthogonal dictionary learning (ODL), sparse blind deconvolution, etc., Qu et al. (2020). Specifically, based on a similar formulation the authors in Bai et al. (2019) proved that $c' = \mathcal{O}(clogc)$ independent random initial vectors suffice in order to recover with high probability a dictionary of size $D \times c$ with high accuracy.

3 Dual Principal Component Pursuit and the Projected Subgradient Method

In Zhu et al. (2018), the authors proposed a projected subgradient descent algorithm for addressing (1) that consists of a) a subgradient step of the form

$$\boldsymbol{b}^{k+1} = \hat{\boldsymbol{b}}^k - \mu^k \left(\boldsymbol{O}\mathrm{Sgn}(\boldsymbol{O}^\top \boldsymbol{b}^k) + \boldsymbol{X}\mathrm{Sgn}(\boldsymbol{X}^\top \boldsymbol{b}^k)\right) \tag{4}$$

where μ^k is the step size that is adaptively updated per iteration and $\hat{\boldsymbol{b}}^k$ is the unit ℓ_2 norm vector corresponding to the kth iteration, followed by b) a projection onto the sphere $\mathbb{S}^{D-1}$ i.e., $\hat{\boldsymbol{b}}^{k+1} = \mathcal{P}_{\mathbb{S}^{D-1}}(\boldsymbol{b}^{k+1})$.

The convergence properties of the projected subgradient algorithm described above depend on specific quantities that reflect the geometry of the problem and are defined next,

$$c_{\boldsymbol{X},\min} = \frac{1}{N} \min_{\boldsymbol{b} \in \mathbb{S}^{d-1} \cap \mathcal{S}} \|\boldsymbol{X}^\top \boldsymbol{b}\|_1 \text{ and } c_{\boldsymbol{X},\max} = \frac{1}{N} \max_{\boldsymbol{b} \in \mathbb{S}^{D-1} \cap \mathcal{S}} \|\boldsymbol{X}^\top \boldsymbol{b}\|_1 \tag{5}$$

Note that the more well distributed the inliers are in the subspace $\mathcal{S}$ the higher the value of the quantity $c_{\boldsymbol{X},\min}$ (called as permeance statistic and first appeared in Lerman et al. (2015)) it becomes since it gets harder to find a vector $\boldsymbol{b}$ that is orthogonal to the inliers. Moreover, $c_{\boldsymbol{X},\min}$ and $c_{\boldsymbol{X},\max}$ converge to the same value c_d as $N \to \infty$ and inliers are uniformly distributed in the subspace i.e., $c_{\boldsymbol{X},\min} \to c_d, c_{\boldsymbol{X},\max} \to c_d$ where c_d is the average height of the unit hemisphere on $\mathbb{R}^d$ as,

$$c_d := \frac{(d-2)!!}{(d-1)!!} \begin{cases} \frac{2}{\pi}, \text{ if } d \text{ is even,} \\ 1, \text{ if } d \text{ is odd} \end{cases} \text{ where } k!! = \begin{cases} k(k-2)(k-4)\cdots 4 \cdot 2, \ k \text{ is even,} \\ k(k-2)(k-4)\cdots 3 \cdot 1, \ k \text{ is odd} \end{cases} \tag{6}$$

Similarly to $c_{\boldsymbol{X},\min}, c_{\boldsymbol{X},\max}$ the quantities $c_{\boldsymbol{O},\min}, c_{\boldsymbol{O},\max}$ indicate how well-distributed the outliers are in the ambient space and are defined as

$$c_{\boldsymbol{O},\min} = \min_{\boldsymbol{b} \in \mathbb{S}^{D-1}} \frac{1}{M} \|\boldsymbol{X}^\top \boldsymbol{b}\|_1 \text{ and } c_{\boldsymbol{O},\max} = \max_{\boldsymbol{b} \in \mathbb{S}^{D-1}} \frac{1}{M} \|\boldsymbol{X}^\top \boldsymbol{b}\|_1 \tag{7}$$

$c_{\boldsymbol{O},\max}$ can be viewed as the *dual* permeance statistic and is bounded away from small values while its difference from $c_{\boldsymbol{O},\min}$ tends to zero as $M \to \infty$. Specifically, $c_{\boldsymbol{O},\max} \to c_D$ and $c_{\boldsymbol{O},\min} \to c_D$ where c_D is defined as in (6), (Zhu et al., 2018).

We also define the following quantities,

$$\eta_{\boldsymbol{O}} = \frac{1}{M} \max_{\boldsymbol{b} \in \mathbb{S}^{D-1}} \|(\boldsymbol{I} - \boldsymbol{b}\boldsymbol{b}^\top)\boldsymbol{O}\mathrm{Sgn}(\boldsymbol{O}^\top \boldsymbol{b})\|_2 \ , \ \eta_{\boldsymbol{X}} = \frac{1}{M} \max_{\boldsymbol{b} \in \mathbb{S}^{D-1}} \|(\mathcal{P}_{\mathcal{S}} - \boldsymbol{b}\boldsymbol{b}^\top)\boldsymbol{X}\mathrm{Sgn}(\boldsymbol{X}^\top \boldsymbol{b})\|_2 \tag{8}$$

As $M \to \infty$ and assuming outliers in $\boldsymbol{O}$ are well-distributed we get $\boldsymbol{O}\mathrm{Sgn}(\boldsymbol{O}^\top \boldsymbol{b}) \to c_D \boldsymbol{b}$ thus $\eta_{\boldsymbol{O}} \to 0$ (Tsakiris & Vidal, 2018). Similarly to $\eta_{\boldsymbol{O}}$, $\eta_{\boldsymbol{X}} \to 0$ as $N \to \infty$ and given that inliers uniformly fill out the d-dimensional subspace they "live" in. The following theorem provides convergence guarantees of the projected subgradient method that was proposed in Zhu et al. (2018) for addressing problem 1.

Theorem 1 *(Informal - Theorem 3 of Zhu et al. (2018) Let $\{\hat{\boldsymbol{b}}_k\}$ the sequence generated by the projected subgradient algorithm in Zhu et al. (2018), with initialization $\hat{\boldsymbol{b}}_0$ such that*

$$\theta_0 < \arctan\left(\frac{Nc_{\boldsymbol{X},\min}}{N\eta_{\boldsymbol{X}} + M\eta_{\boldsymbol{O}}}\right) \text{ and } Nc_{\boldsymbol{X},\min} \geq N\eta_{\boldsymbol{X}} + M\eta_{\boldsymbol{O}} \tag{9}$$

where θ_0 denotes the principal angle of $\boldsymbol{b}^0$ from $\mathcal{S}_\perp$. If the step size μ^k is updated according to a piecewise geometrically diminishing rule given as

$$\mu^k = \begin{cases} \mu^0, & k < K_0 \\ \mu^0 \beta^{\lfloor (k-K_0)/K_* \rfloor + 1}, & k \geq K_0 \end{cases} \tag{10}$$

where $\beta < 1$, $\lfloor \cdot \rfloor$ is the floor function, then the iterates $\boldsymbol{b}^k$ converge to a normal vector of $\mathcal{S}$.

The full version of Theorem 1 is given in Appendix. Theorem 1 alludes to a relatively large initial step size μ^0 that leads to a small value of K_0. After iteration $k > K_0$ and with a piecewise geometrically diminishing step size rule the projected subgradient algorithm will converge to a normal of $\mathcal{S}$.

4 Dual Principal Component Pursuit in subspaces of unknown codimension

Knowing c is a rather strong requirement and is far from being true in real word applications. In this light, we next aim at alleviating that need and thus propose to lift the orthogonality constraint on $\boldsymbol{B}$.

4.1 PSGM's iterates convergence in the continuous version of DPCP

By assuming that inliers and outliers are uniformly distributed over the sphere *under a continuous measure* we can get the continuous version of problem defined in (3). Note that the continuous version of DPCP can be viewed as a special case of the discrete one. That being said, it inherits all the theoretical properties that have been derived in Tsakiris & Vidal (2018), Zhu et al. (2018). As such, it was first analyzed in Tsakiris & Vidal (2018) for the case of codimension $c = 1$ with the aim to provide an insight into to the landscape properties of the DPCP problem. Let us now define for any $\boldsymbol{b} \in \mathbb{S}^{D-1}$ the function $f_{\boldsymbol{b}} : \mathbb{S}^{D-1} \to \mathbb{R}$, $f_{\boldsymbol{b}}(\mathbf{z}) = |\mathbf{z}^\top \boldsymbol{b}|$.The following lemma provides the continuous version of the discrete objective function given in (3).

Lemma 2 *In the continuous case, the discrete DPCP objective given in (3) can be written in the following form,*

$$\sum_{i=1}^{c'} \left(M\mathbb{E}_{\boldsymbol{\mu}_{\mathbb{S}^{D-1}}}[f_{\mathbf{b}_i}] + N\mathbb{E}_{\boldsymbol{\mu}_{\mathbb{S}^{D-1}\cap\boldsymbol{S}}}[f_{\boldsymbol{b}_i}] \right) = \sum_{i=1}^{c'} \|\boldsymbol{b}_i\|_2 \left(Mc_D + Nc_d \cos(\phi_i) \right) \tag{11}$$

where ϕ_i is the principal angle of $\mathbf{b}_i$ from the inliers subspace $\mathcal{S}$.

where $\boldsymbol{\mu}_{\mathbb{S}^{D-1}}, \boldsymbol{\mu}_{\mathbb{S}^{D-1}\cap\boldsymbol{S}}$ denote the continuous measures associated with the outliers and inliers, respectively. Evidently, (11) attains its minimum value for vectors $\boldsymbol{b}_i$s whose principal angles ϕ_is from the inliers' subspace are equal to $\frac{\pi}{2}$.

Based on (3) and due to Lemma 2, we can now minimize the objective function of the "continuous version" of DPCP by employing a projected subgradient methods (PSGM) that performs the following steps per iteration [2]

$$\begin{aligned} \boldsymbol{b}_i^{k+1} &= \hat{\boldsymbol{b}}_i^k - \mu_i^k (Mc_D \hat{\boldsymbol{b}}_i^k + Nc_d \hat{\mathbf{s}}_i^k) \quad , i = 1, 2, \ldots, c' \\ \hat{\boldsymbol{b}}_i^{k+1} &= \mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{b}_i^{k+1}) \quad , i = 1, 2, \ldots, c' \end{aligned} \tag{12}$$

Corollary 3 *As $k \to \infty$, a projected subgradient algorithm consisting of the steps described in (12) and for specific step-size updates (piecewise geometrically diminishing step size rule) will almost surely converge to a matrix $\hat{\boldsymbol{B}}^* \in \mathbb{R}^{D\times c'}$ whose columns $\hat{\boldsymbol{b}}_i^*$, $i = 1, 2, \ldots, c'$ will be normal vectors of the inliers' subspace when randomly initialized with independent vectors $\boldsymbol{b}_i^0 \in \mathbb{S}^{D-1}$, $i = 1, 2, \ldots, c'$ uniformly distributed over the sphere $\mathbb{S}^{D-1}$.*

[2]Note that $\partial\|\boldsymbol{b}\|_2 = \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|_2}$ for $\boldsymbol{b} \neq \mathbf{0}$ and $\|\boldsymbol{b}_i\|_2 \cos(\phi_i) = \boldsymbol{b}_i^\top \hat{\mathbf{s}}_i$ where $\hat{\mathbf{s}}_i = \frac{\mathcal{P}_{\mathcal{S}}(\boldsymbol{b}_i)}{\|\mathcal{P}_{\mathcal{S}}(\boldsymbol{b}_i)\|_2}$.

Corollary 3 allows us to claim that we can always recover $c' > c$ normal vectors to the inliers' subspace using a PSGM algorithm consisting of steps given in (12). However, this does not tell the whole story yet, since our ultimate objective is to recover a matrix $\hat{\boldsymbol{B}}$ that spans $\mathcal{S}_\perp$. It is thus imperative to ensure that the rank of $\hat{\boldsymbol{B}}$ is equal to the true and *unknown* codimension of the inliers' subspace c. Next we show that by initializing with a $\hat{\boldsymbol{B}}_0$ such that $\operatorname{rank}(\hat{\boldsymbol{B}}_0) = c'$, we can guarantee that we can solve the continuous version of DPCP using PSGM and converge to a $\hat{\boldsymbol{B}}$ such that $\operatorname{rank}(\hat{\boldsymbol{B}}) = c$ thus getting $\operatorname{span}(\hat{\boldsymbol{B}}) \equiv \mathcal{S}_\perp$. First observe that by projecting the PSGM iterates given in (12) onto $\mathcal{S}_\perp$ we have,

$$\mathcal{P}_{\mathcal{S}_\perp}(\hat{\boldsymbol{b}}_i^{k+1}) = \mathcal{P}_{\mathcal{S}_\perp}(\hat{\boldsymbol{b}}_i^{k}) - \mu_i^k M c_D \mathcal{P}_{\mathcal{S}_\perp}(\hat{\boldsymbol{b}}_i^{k}) \text{ and } \mathcal{P}_{\mathcal{S}_\perp}(\hat{\boldsymbol{b}}_i^{k+1}) = \mathcal{P}_{\mathcal{S}_\perp}(\mathcal{P}_{\mathbb{S}^{D-1}}(\boldsymbol{b}_i^{k+1})) \tag{13}$$

We can thus deduce that the projections of successive iterates of PSGM for updating $\boldsymbol{b}_i$ are scaled versions of the corresponding projections of the previous iterates. We can now state Lemma 4.

Lemma 4 *The PSGM iterates $\hat{\boldsymbol{b}}_i^k,\ i = 1, 2, \ldots, c',\ k = 1, 2, \ldots$ given in (12), when randomly initialized with $\hat{\boldsymbol{b}}_i^0$s, $i = 1, 2, \ldots, c'$ that are independently drawn from a spherical distribution with unit ℓ_2 norm converge almost surely to c' normal vectors of the inliers subspace $\mathcal{S}$ denoted as $\hat{\boldsymbol{b}}_i^*,\ i = 1, 2, \ldots, c'$ that are given by $\hat{\boldsymbol{b}}_i^* = \frac{\mathcal{P}_{\mathcal{S}_\perp}(\hat{\boldsymbol{b}}_i^0)}{\|\mathcal{P}_{\mathcal{S}_\perp}(\hat{\boldsymbol{b}}_i^0)\|_2}, \quad i = 1, 2, \ldots, c'$.*

Lemma 4 shows that the initialization of PSGM plays a pivotal role since it determines the direction of the recovered normal vectors $\hat{\boldsymbol{b}}_i^*$s. Corollary 3 and Lemma 4 pave the way for Theorem 5.

Theorem 5 *Let $\hat{\boldsymbol{B}}^0 \in \mathbb{R}^{D \times c'}$ where $c' \geq c$ with c denoting the true codimension of the inliers subspace $\mathcal{S}$, consisting of unit ℓ_2 norm column vectors $\hat{\boldsymbol{b}}_i^0 \in \mathbb{S}^{D-1}, i = 1, 2, \ldots, c'$ that are independently drawn from uniform distribution over the sphere $\mathbb{S}^{D-1}$. A PSGM algorithm initialized with $\hat{\boldsymbol{B}}^0$ will almost surely converge to a matrix $\hat{\boldsymbol{B}}^*$ such that $\operatorname{span}(\hat{\boldsymbol{B}}^*) \equiv \mathcal{S}_\perp$.*

From Theorem 5 we observe that in the benign scenario where inliers and outliers are uniformly distributed under continuous measures, DPCP problem can be formulated in a way that allow us to recover the correct orthogonal complement of the inlier subspace even when we are oblivious to its true codimension. Remarkably, this is achieved without any explicit regularization being enforced but rather by exploiting the *implicit bias* induced by multiple random initializations of the PSGM algorithm for solving the DPCP formulation without orthogonality constraints equation 3.

4.2 PSGM'S ITERATES CONVERGENCE IN THE DISCRETE VERSION OF DPCP

From this analysis of the continuous version of DPCP we now extend to the the discrete version, which is of more practical relevance for finite data, yet also presents more challenges. To begin, we reformulate the DPCP objective as follows

$$\sum_{i=1}^{c'} \|\tilde{\boldsymbol{X}}^\top \boldsymbol{b}_i\|_1 = \sum_{i=1}^{c'} \|\boldsymbol{X}^\top \boldsymbol{b}_i\|_1 + \|\boldsymbol{O}^\top \boldsymbol{b}_i\|_1 = M \sum_{i=1}^{c'} \mathbf{b}_i^\top \boldsymbol{o}_{\boldsymbol{b}_i} + N \sum_{i=1}^{c'} \boldsymbol{b}_i^\top \boldsymbol{x}_{\boldsymbol{b}_i} \tag{14}$$

where $\boldsymbol{x}_{\boldsymbol{b}_i}$ and $o_{\boldsymbol{b}_i}$ are called as *average inliers* and *average outliers* terms, defined as $\boldsymbol{x}_{\boldsymbol{b}_i} = \frac{1}{N}\sum_{j=1}^{N} Sgn(\boldsymbol{b}_i^\top \boldsymbol{x}_j)\boldsymbol{x}_j$ and $\boldsymbol{o}_{\boldsymbol{b}_i} = \frac{1}{M}\sum_{j=1}^{M} Sgn(\boldsymbol{b}_i^\top \boldsymbol{o}_j)\boldsymbol{o}_j$.

Likewise, in Algorithm 1, we give the projected sub-gradient method applied on the DPCP problem in (3) (named DPCP-PSGM) in terms of these quantities.

Note that the average outliers and inliers terms of (**??**) are discrete versions of the corresponding continuous average terms $c_D \boldsymbol{b}_i$ and $c_d \hat{\mathbf{s}}_i$ where $\hat{\mathbf{s}}_i = \mathcal{P}_{\mathcal{S}}(\boldsymbol{b}_i)$, respectively, Tsakiris & Vidal (2018). We now express the sub-gradient step of Algorithm 1 as

$$\boldsymbol{b}_i^{k+1} = \hat{\boldsymbol{b}}_i^k - \mu_i^k \left(M(c_D \hat{\boldsymbol{b}}_i^k + \boldsymbol{e}_{\boldsymbol{O}}^{i,k}) + N(c_D \hat{\mathbf{s}}_i^k + \boldsymbol{e}_{\boldsymbol{X}}^{i,k}) \right) \tag{15}$$

where the quantities $\boldsymbol{e}_{\boldsymbol{O}}^{i,k} = \boldsymbol{o}_{\hat{\boldsymbol{b}}_i^k} - c_D \hat{\boldsymbol{b}}_i^k -$ and $\boldsymbol{e}_{\boldsymbol{X}}^{i,k} = \boldsymbol{x}_{\hat{\boldsymbol{b}}_i^k} - c_D \hat{\mathbf{s}}_i^k$ account for the error between the continuous and discrete versions of the average outliers and the average inliers terms, respectively.

Algorithm 1: DPCP-PSGM algorithm for solving (3)

Result: $\hat{\boldsymbol{B}} = [\hat{\boldsymbol{b}}_1^k, \hat{\boldsymbol{b}}_2^k, \ldots, \hat{\boldsymbol{b}}_{c'}^k]$
Initialize: $\hat{\boldsymbol{B}}_0 = [\hat{\boldsymbol{b}}_0^1\ \hat{\boldsymbol{b}}_0^2\ \ldots \hat{\boldsymbol{b}}_0^{c'}]$;
for $k = 1, 2, \ldots$ **do**
  for $i = 1, 2, \ldots, c'$ **do**
    Update the step-size according to a specific rule;
    $\boldsymbol{b}_i^{k+1} = \hat{\boldsymbol{b}}_i^k - \mu_i^k (M \boldsymbol{o}_{\hat{\boldsymbol{b}}_i}^k + N \boldsymbol{x}_{\hat{\boldsymbol{b}}_i}^k)$;
    $\hat{\boldsymbol{b}}_i^{k+1} = \mathcal{P}_{\mathbb{S}^{D-1}}(\boldsymbol{b}_i^{k+1})$;
  end
end

Following a similar path as in the continuous case we next project the iterates of (15) onto $\mathcal{S}_\perp$,

$$\begin{aligned}\mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{b}_i^{k+1}) &= \mathcal{P}_{\mathcal{S}_\perp}(\hat{\boldsymbol{b}}_i^k) - \mu_i^k \left(M \mathcal{P}_{\mathcal{S}_\perp}(c_D \hat{\mathbf{b}}_i^k + \boldsymbol{e}_{\boldsymbol{O}}^{i,k}) + N \cancelto{0}{\mathcal{P}_{\mathcal{S}_\perp}(c_D \hat{\mathbf{s}}_i^k + \boldsymbol{e}_{\boldsymbol{X}}^{i,k})} \right) \\ &= (1 - \mu_i^k M c_D) \mathcal{P}_{\mathcal{S}_\perp}(\hat{\boldsymbol{b}}_i^k) - \mu_i^k M \mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{e}_{\boldsymbol{O}}^{i,k})\end{aligned} \tag{16}$$

Remark. *Eq. (16) reveals that DPCP-PSGM, applied on the discrete problem, gives rise to updates whose projections to $\mathcal{S}_\perp$ are* **scaled** *and* **perturbed** *versions of the previous estimates, and the magnitude of perturbation depends on the discrepancy between the continuous and the discrete problem.*

Recall that $\mathcal{P}_{\mathcal{S}_\perp}(\hat{\boldsymbol{b}}_i^k) = \frac{\mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{b}_i^k)}{\|\boldsymbol{b}_i^k\|_2}$, and we can rewrite the update of the 2nd iteration of DPCP-PSGM,

$$\begin{aligned}\mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{b}_i^2) &= \frac{(1 - \mu_i^1 M c_D)}{\|\boldsymbol{b}_i^1\|_2} \left((1 - \mu_i^0 M c_D) \mathcal{P}_{\mathcal{S}_\perp}(\hat{\boldsymbol{b}}_i^0) - \mu_i^0 M \mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{e}_{\boldsymbol{O}}^{i,0}) \right) - \mu_i^1 M \mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{e}_{\boldsymbol{O}}^{i,1}) \\ &= \frac{(1 - \mu_i^1 M c_D)(1 - \mu_i^0 M c_D)}{\|\boldsymbol{b}_i^1\|_2 \|\boldsymbol{b}_i^0\|_2} \mathcal{P}_{\mathcal{S}_\perp}(\hat{\mathbf{b}}_i^0) - \frac{(1 - \mu_i^1 M c_D)}{\|\boldsymbol{b}_i^1\|_2} \mu_i^0 M \mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{e}_{\boldsymbol{O}}^{i,0}) - \mu_i^1 M \mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{e}_{\boldsymbol{O}}^{i,1})\end{aligned} \tag{17}$$

where we have assumed that $\|\boldsymbol{b}_i^0\|_2 = 1$. By repeatedly applying the same steps, we can reach to the following recursive expression for $\mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{b}_i^K)$,

$$\mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{b}_i^K) = \left(\prod_{k=0}^{K-1} \frac{(1 - \mu_i^k M c_D)}{\|\boldsymbol{b}_i^k\|_2} \right) \mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{b}_i^0) - \sum_{k=0}^{K-1} \left(\prod_{j=k+1}^{K-1} \frac{(1 - \mu_i^j M c_D)}{\|\boldsymbol{b}_i^j\|_2} \right) \mu_i^k M \mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{e}_{\boldsymbol{O}}^{i,k}) \tag{18}$$

where for $j > K - 1$ we set $\prod_{j=k+1}^{K-1} \frac{(1 - \mu_i^j M c_D)}{\|\boldsymbol{b}_i^j\|_2} = 1$.

By dividing (18) with $\prod_{k=0}^{K-1} \frac{(1 - \mu_i^k M c_D)}{\|\boldsymbol{b}_i^k\|_2}$ and by projecting onto the sphere $\mathbb{S}^{D-1}$ we get

$$\mathcal{P}_{\mathbb{S}^{D-1}}(\mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{b}_i^K)) = \mathcal{P}_{\mathbb{S}^{D-1}}(\mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{b}_i^0) - \mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{\delta}_i^K)) \tag{19}$$

where $\boldsymbol{\delta}_i$ is defined as follows,

$$\boldsymbol{\delta}_i^K = \sum_{k=0}^{K-1} \left(\prod_{j=0}^{k} \frac{\|\boldsymbol{b}_i^j\|_2}{(1 - \mu_i^j M c_D)} \right) \mu_i^k M \boldsymbol{e}_{\boldsymbol{O}}^{i,k} \tag{20}$$

Assumption 1. *We assume that the principal angles θ_0^i for all $\boldsymbol{b}_i^0$s satisfy the inequality $\theta_0^i < \arctan\left(\frac{N c_{\boldsymbol{X},\min}}{N \eta_{\boldsymbol{X}} + M \eta_{\boldsymbol{O}}}\right)\ \ \forall i, i = 1, 2, \ldots, c'$.*

Assumption 1 essentially assumes that the sufficient condition given in eq. (9) required by PSGM algorithm for converging to a normal vector is satisfied which is the same condition for success in

Zhu et al. (2018). Under Assumption 1 we can invoke the convergence properties of PSGM given in Theorem 1 and get as $K \to \infty, \mathcal{P}_{\mathbb{S}^{D-1}}(\mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{b}^K)) \to \hat{\boldsymbol{b}}^* \in \mathcal{S}_\perp \cap \mathbb{S}^{D-1}$. That being said, we denote $\hat{\boldsymbol{b}}_i^* = \mathcal{P}_{\mathbb{S}^{D-1}}\left(\mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{b}_i^0) - \mathcal{P}_{\mathcal{S}_\perp}(\hat{\boldsymbol{\delta}}_i)\right)$, where $\hat{\boldsymbol{\delta}}_i = \lim_{K\to\infty} \boldsymbol{\delta}_i^K$[3]. Following the same steps as in Section 4.1 and by defining matrices $\boldsymbol{B}^* = [\boldsymbol{b}_1^*, \boldsymbol{b}_2^*, \ldots, \boldsymbol{b}_{c'}^*], \boldsymbol{B}^0 = [\boldsymbol{b}_1^0, \boldsymbol{b}_2^0, \ldots, \boldsymbol{b}_{c'}^0], \hat{\boldsymbol{\Delta}} = [\hat{\boldsymbol{\delta}}_1, \hat{\boldsymbol{\delta}}_2, \ldots, \hat{\boldsymbol{\delta}}_{c'}]$ we can express the matrix $\boldsymbol{B}^*$ as $\hat{\boldsymbol{B}}^* = \mathcal{P}_{\mathbb{S}^{D-1}}\left(\mathcal{P}_{\mathcal{S}_\perp}\left(\boldsymbol{B}^0 - \hat{\boldsymbol{\Delta}}\right)\right)$ where $\boldsymbol{B}^*$ will now consist of normal vectors of the inliers' subspace. In order to guarantee that $\operatorname{span}(\boldsymbol{B}^*) \equiv \mathcal{S}_\perp$ it thus suffices to ensure that $\operatorname{rank}(\boldsymbol{B}^*) = c$. Here we show that a sufficient condition for this to hold is that the matrix $\boldsymbol{A} = \boldsymbol{B}^0 - \hat{\boldsymbol{\Delta}}$ is full-rank.

Lemma 6 *If $\sigma_{c'}(\boldsymbol{B}^0) \geq \|\hat{\boldsymbol{\Delta}}\|_2$ then matrix $\boldsymbol{A} = \boldsymbol{B}^0 - \hat{\boldsymbol{\Delta}}$ is full-rank.*

From Lemma 6 we can see that the success of DPCP-PSGM hinges on how well-conditioned the matrix $\boldsymbol{B}^0$ is. Specifically, it says that if a lower-bound on the smallest singular is satisfied then DPCP-PSGM is guaranteed to converge to the correct complement of the inlier without knowledge of the correct codimension c. From this, we can prove the following Theorem.

Theorem 7 *Let $\mathbf{B}^0 \in \mathbb{R}^{D\times c'}$ with columns randomly sampled from a unit ℓ_2 norm spherical distribution where $c' \geq c$ with c denoting the true codimension of the inliers subspace $\mathcal{S}$ that satisfies Assumption 1. If*

$$1 - C_1\sqrt{\frac{c'}{D}} - \frac{\epsilon}{\sqrt{D}} > \sqrt{c'}\kappa(\boldsymbol{\eta}_{\boldsymbol{\mathcal{O}}} + c_{\boldsymbol{\mathcal{O}},\max} - c_d) \tag{21}$$

where $\kappa = \max_i \frac{M\mu_i^0}{\beta^{K_0/K_*}(1-r_i)}$ and $r_i = \frac{\left(1+\mu_i^0(N(\boldsymbol{\eta}_{\boldsymbol{\mathcal{X}}}+c_{\boldsymbol{\mathcal{X}},\max})+M(\boldsymbol{\eta}_{\boldsymbol{\mathcal{O}}}+c_{\boldsymbol{\mathcal{O}},\max}))\right)}{1-\mu_i^0 M c_D}\beta^{1/K_*}$ then with probability at least $1 - 2\exp(-\epsilon^2 C_2)$ (where C_1, C_2 are absolute constants), Algorithm 1 with a geometrically diminishing step size rule will converge to a matrix $\hat{\boldsymbol{B}}^*$ such that $\operatorname{span}(\hat{\boldsymbol{B}}^*) \equiv \mathcal{S}_\perp$.

Theorem 7 shows that we can randomly initialize DPCP-PSGM, with a matrix $\hat{\boldsymbol{B}}^0$ whose number of columns c' is an overestimate of the true co-dimension c of the inliers' subspace and with columns sampled *independently* by a uniform distribution over the unite sphere, and recover a matrix that will span the orthogonal complement of $\mathcal{S}$. The probability of success depends on the geometry of the problem since condition (21) is trivially satisfied (RHS of (21) tends to 0) in the continuous case which incurs a benign geometry. Moreover, the a less benign geometry requires a smaller initial codimension c' that would increase the LHS of (21).

5 Numerical Simulations

In this section we manifest the effectiveness of the proposed DPCP formulation and the derived DPCP-PSGM algorithm in recovering orthogonal complements of subspace of unknown codimension. We compare the proposed algorithm with previously developed methods i.e., DPCP-IRLS Tsakiris & Vidal (2018) and the Riemannian Subgradient Method (RSGM) Zhu et al. (2019). Recall that both DPCP-IRLS and RSGM address DPCP problem enforcing orthogonality constraints thus both algorithms are quite sensitive if true co-dimension of $\mathcal{S}$ is not known a-priori, while they are both initialized using of spectral initialization i.e., $\hat{\boldsymbol{B}}^0 \in \mathbb{R}^{D\times c'}$ contains the first c' eigenvectors of matrix $\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}^\top$ as its columns, as proposed in Tsakiris & Vidal (2018); Zhu et al. (2019).

Robustness to outliers in the unknown co-dimension regime. In this experiment we set the dimension of the ambient space to $D = 200$. We randomly generate N inliers uniformly distributed with unit ℓ_2 norm in a $d = 195$ dimensional subspace (hence for its co-dimension we have $c = D - d = 5$). Following a similar process we generate M outliers that live in the ambient space and are sampled from a uniform distribution over the unit sphere. Fig. 2, illustrates the distances [4] of the recovered matrix $\hat{\boldsymbol{B}}$ as obtained by the proposed DPCP-PSGM algorithm initialized with

[3] For the sake of brevity we assume that the step size has been selected such that existence of the limit is guaranteed. We refer the reader to the proof Lemma 8 for further details.

[4] We define the distance between two subspaces spanned by matrices $\boldsymbol{B}$ and $\boldsymbol{A}$ as $dist(\boldsymbol{B}, \boldsymbol{A}) = \min_{\boldsymbol{Q}\in\mathbb{O}(D,c)} \|\boldsymbol{B} - \boldsymbol{A}\boldsymbol{Q}\|_F$ where $\mathbb{O}(D, c)$ denotes the Stiefel manifold of orthogonal matrices of rank c. Note that $dist(\boldsymbol{B}, \boldsymbol{A}) = 0 \iff \operatorname{span}(\boldsymbol{B}) \equiv \operatorname{span}(\boldsymbol{A})$ see Zhu et al. (2019))

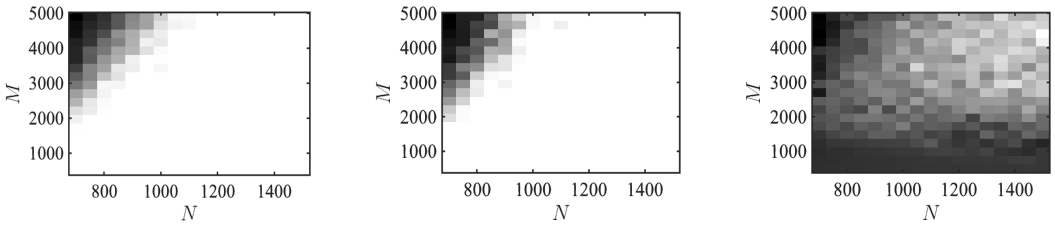

Figure 2: Distances of the recovered $\hat{\boldsymbol{B}}$ from the true orthogonal complements $\mathcal{S}_{\perp}$ as recovered by a) the proposed DPCP-PSGM algorithm provided an overestimated of the true c i.e.,$c' = 10$ (left) b) RSGM provided the true c (middle) and c) RSGM provided $c' = 10$ (right). Darker colors reflect higher values of distances while lighter colors indicate successful recoveries of $\boldsymbol{B}$.

an overestimate $c' = 10$ of the true codimension c codimension and two versions of RSGM i.e., a) RSGM when it is given as input true $c = 5$ and b) RSGM when being incognizant of c and hence it initialized with a $c' = 10$ of c. As is shown in Fig. 2(c), RSGM fails to recover the correct orthogonal complement of $\mathcal{S}$ when it is provided with an overestimate of the true c which is attributed to spectral initialization and the imposed orthogonality constraints. Contrary, DPCP-PSGM albeit incognizant to c displays a remarkably robust behavior (Fig. 2(b)) performing similarly to RSGM when though the latter knows beforehand the correct co-dimension.

Recovery of the true codimension. Here we test DPCP-PSGM on the recovery of the true co-dimension c of the inliers' subspace $\mathcal{S}$. Again, we set $D = 200$ and follow the same process described above for generating $N = 1500$ inliers. We vary the true co-dimension of $\mathcal{S}$ from $c = 10$ to 20 and consider two different outlier's ratios r, defined as $r = \frac{M}{M+N}$, namely $r = 0.6$ and $r = 0.7$. In both cases, DPCP-PSGM is initialized with the same overestimate of c i.e., $c' = 30$. In Fig. 3 we report the estimated codimensions obtained by DPCP-PSGM for 10 independent trials of the experiments. It can be observed that DPCP-PSGM achieves 100% for all different codimensions for $r = 0.6$. Moreover, it shows a remarkable performance in estimating the correct c's even in the more challenging case corresponding to outliers' ratios equal to 0.7. with the estimated codimensions being close to the true values even in the cases that it fails to exactly compute c. The results corroborate the theory showing that the DPCP-PSGM with random initialization biases the solutions of $\hat{\mathbf{B}}$ towards matrices with rank c.

A hyperspectral imaging experiment See Appendix B.1.

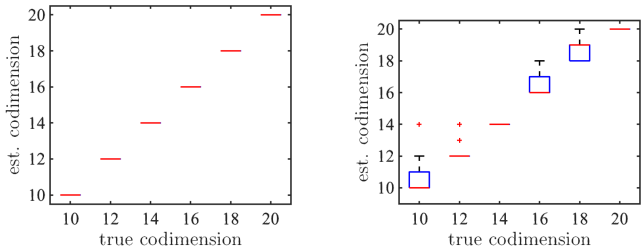

Figure 3: Estimated by DPCP-PSGM codimensions for two different outliers' ratios $r = \frac{M}{M+N}$ (a) $r = 0.6$ (left) and (b) $r = 0.6$ (right)

6 Conclusions

We proposed a simple framework which allows us to perform robust subspace recovery without requiring a priori knowledge of the subspace codimension. This is based on Dual Principal Component Pursuit (DPCP) and thus is amenable to handling subspaces of high relative dimensions. We observed that a projected subgradient methods (PSGM) induces implicit bias and converges to a matrix that spans a basis of the orthogonal complement of the inliers subspace even as long as a) we overestimate it codimension, b) lift orthogonality constraints enforced in previous DPCP formulations and c) use random initialization. We provide empirical results that corroborate the developed theory and showcase the merits of our approach.

Ethics Statement This work focuses on theoretical aspects of robust subspace recovery problem which is a well-established topic in machine learning research. The research conducted in the framework of this work raises no ethical issues or any violations vis-a-vis the ICLR Code of Ethics.

REFERENCES

Yu Bai, Qijia Jiang, and Ju Sun. Subgradient descent learns orthogonal dictionaries. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HklSf3CqKm.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

Tianyu Ding, Zhihui Zhu, Rene Vidal, and Daniel P Robinson. Dual principal component pursuit for robust subspace learning: Theory and algorithms for a holistic approach. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2739–2748. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ding21b.html.

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

Paris V. Giampouras, Athanasios A. Rontogiannis, and Konstantinos D. Koutroumbas. Alternating iteratively reweighted least squares minimization for low-rank matrix factorization. *IEEE Transactions on Signal Processing*, 67(2):490–503, 2019. doi: 10.1109/TSP.2018.2883921.

Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

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

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

Tyler Maunu, Teng Zhang, and Gilad Lerman. A well-tempered landscape for non-convex robust subspace recovery. *Journal of Machine Learning Research*, 20(37), 2019.

Michael McCoy and Joel A Tropp. Two proposals for robust PCA using semidefinite programming. *Electronic Journal of Statistics*, 5:1123–1160, 2011.

Qing Qu, Zhihui Zhu, Xiao Li, Manolis C Tsakiris, John Wright, and René Vidal. Finding the sparsest vectors in a subspace: Theory, algorithms, and applications. *arXiv preprint arXiv:2001.06970*, 2020.

Mostafa Rahmani and George Atia. Coherence pursuit: Fast, simple, and robust subspace recovery. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2864–2873. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/rahmani17a.html.

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

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

René Vidal, Yi Ma, and S Shankar Sastry. *Generalized principal component analysis*, volume 5. Springer, 2016.

Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust PCA via outlier pursuit. *IEEE Transactions on Information Theory*, 58(5):3047–3064, 2012.

C. You, D. Robinson, and R. Vidal. Provable self-representation based outlier detection in a union of subspaces. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4323–4332, 2017a.

Chong You, Daniel P Robinson, and René Vidal. Provable self-representation based outlier detection in a union of subspaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3395–3404, 2017b.

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

Z. Zhu, T. Ding, M. C. Tsakiris, D. P. Robinson, and R. Vidal. A linearly convergent method for non-smooth non-convex optimization on the grassmannian with applications to robust subspace and dictionary learning. In *Neural Information Processing Systems (NIPS)*, 2019.

Zhihui Zhu, Yifan Wang, Daniel Robinson, Daniel Naiman, René Vidal, and Manolis Tsakiris. Dual principal component pursuit: Improved analysis and efficient algorithms. In *Advances in Neural Information Processing Systems 2018*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

A APPENDIX

Theorem 1 *(Theorem 3 of Zhu et al. (2018)) Let $\{\hat{\boldsymbol{b}}_k\}$ the sequence generated by the projected subgradient method in Zhu et al. (2018), with initialization $\hat{\boldsymbol{b}}_0$ such that*

$$\theta_0 < \arctan\left(\frac{Nc_{\boldsymbol{X},\min}}{N\eta_{\boldsymbol{X}} + M\eta_{\boldsymbol{O}}}\right) \tag{22}$$

where θ_0 denotes the principal angle of $\boldsymbol{b}^0$ from $\mathcal{S}_\perp$, and

$$Nc_{\boldsymbol{X},\min} \geq N\eta_{\boldsymbol{X}} + M\eta_{\boldsymbol{O}} \tag{23}$$

Let $\mu' := \frac{1}{4\max\{Nc_{\boldsymbol{X},\min}, Mc_{\boldsymbol{O},\max}\}}$. If $\mu^0 \leq \mu'$ and the step size μ^k is updated according to a piece-wise geometrically diminishing rule given as

$$\mu^k = \begin{cases} \mu^0, & k < K_0 \\ \mu^0\beta^{\lfloor (k-K_0)/K_* \rfloor + 1}, & k \geq K_0 \end{cases} \tag{24}$$

where $\beta < 1$, $\lfloor \cdot \rfloor$ is the floor function, and $K_0, K_* \in \mathbb{N}$ are chosen such that

$$\begin{aligned} &K_0 \geq K^\diamondsuit(\mu^0), \\ &K_* \geq \left(\sqrt{2}\beta\mu'\left(Nc_{\boldsymbol{X}} - (N\eta_{\boldsymbol{X}} + M\eta_{\boldsymbol{O}})\right)\right)^{-1} \end{aligned}$$

where,

$$K^\diamondsuit(\mu) := \frac{\tan(\theta_0)}{\mu\left(Nc_{\boldsymbol{X},\min} - \max\{1, \tan(\theta_0\}\left(N\eta_{\boldsymbol{X}} + M\eta_{\boldsymbol{O}}\right)\right)} \tag{25}$$

then for the angle θ_k between $\hat{\boldsymbol{b}}^k$ and $\mathcal{S}_\perp$ it holds

$$\tan(\theta_0) \leq \begin{cases} \max\{\tan(\theta_0), \frac{\mu_0}{\sqrt{2}\mu'}\}, & k < K_0 \\ \frac{\mu^0}{\sqrt{2}\mu'}\beta^{\lfloor (k-K_0)/K_* \rfloor}, & k \geq K_0 \end{cases} \tag{26}$$

A.1 PROOF OF LEMMA 2

Lemma 2 *In the continuous case, the discrete DPCP objective given in (3) can be written in the following form,*

$$\sum_{i=1}^{c'}\left(M\mathbb{E}_{\boldsymbol{\mu}_{\mathbb{S}^{D-1}}}[f_{\mathbf{b}_i}] + N\mathbb{E}_{\boldsymbol{\mu}_{\mathbb{S}^{D-1}\cap\boldsymbol{S}}}[f_{\boldsymbol{b}_i}]\right) = \sum_{i=1}^{c'}\|\boldsymbol{b}_i\|_2\left(Mc_D + Nc_d\cos(\phi_i)\right) \tag{27}$$

where ϕ_i is the principal angle of $\boldsymbol{b}_i$ from the inliers subspace $\mathcal{S}$.

Proof We define the discrete measures $\boldsymbol{\mu}_{\boldsymbol{\mathcal{X}}}, \boldsymbol{\mu}_{\boldsymbol{\mathcal{O}}}$ associated with the inliers and outliers, respectively as,

$$\boldsymbol{\mu}_{\boldsymbol{\mathcal{X}}}(\mathbf{z}) = \frac{1}{N}\sum_{j=1}^{N}\delta(\mathbf{z} - \boldsymbol{o}_j), \;\; \boldsymbol{\mu}_{\boldsymbol{\mathcal{O}}}(\mathbf{z}) = \frac{1}{M}\sum_{j}^{N}\delta(\mathbf{z} - \boldsymbol{x}_j) \tag{28}$$

where $\delta(\cdot)$ is the Dirac function. Recall that,

$$\int_{\mathbf{z}\in\mathbb{S}^{D-1}} g(\mathbf{z})\delta(\mathbf{z} - \mathbf{z}_0)d\boldsymbol{\mu}_{\mathbb{S}^{D-1}} = g(z_0) \tag{29}$$

where $g : \mathbb{S}^{D-1} \to \mathbb{R}$ and $\boldsymbol{\mu}_{\mathbb{S}^{D-1}}$ is the uniform measure on $\mathbb{S}^{D-1}$.

The DPCP objective for the discrete version of the problem can be written as,

$$\begin{aligned}
\sum_{i=1}^{c'} \|\tilde{\boldsymbol{\mathcal{X}}}^\top \mathbf{b}_i\|_1 &= \sum_{i=1}^{c'} \left(\|\boldsymbol{\mathcal{X}}^\top \mathbf{b}_i\|_1 + \|\boldsymbol{\mathcal{O}}^\top \mathbf{b}_i\|_1 \right) = \sum_{i}^{c'} \left(\sum_{j=1}^{N} |\boldsymbol{x}_j^\top \mathbf{b}_i| + \sum_{j=1}^{M} |\boldsymbol{o}_j^\top \mathbf{b}_i| \right) \\
&= \sum_{i=1}^{c'} \left(\sum_{j=1}^{N} \int_{\mathbf{z} \in \mathbb{S}^{D-1}} |\mathbf{z}^\top \mathbf{b}_i| \delta(\mathbf{z} - \boldsymbol{x}_j) d\boldsymbol{\mu}_{\mathbb{S}^{D-1}} + \sum_{j}^{M} \int_{\mathbf{z} \in \mathbb{S}^{D-1}} |\mathbf{z}^\top \mathbf{b}_i| \delta(\mathbf{z} - \boldsymbol{o}_j) d\boldsymbol{\mu}_{\mathbb{S}^{D-1}} \right) \\
&= . \sum_{i=1}^{c'} \left(\int_{\mathbf{z} \in \mathbb{S}^{D-1}} |\mathbf{z}^\top \mathbf{b}_i| \sum_{j=1}^{N} \delta(\mathbf{z} - \boldsymbol{x}_j) d\boldsymbol{\mu}_{\mathbb{S}^{D-1}} + \int_{\mathbf{z} \in \mathbb{S}^{D-1}} |\mathbf{z}^\top \mathbf{b}_i| \sum_{j}^{M} \delta(\mathbf{z} - \boldsymbol{o}_j) d\boldsymbol{\mu}_{\mathbb{S}^{D-1}} \right) \\
&= \sum_{i=1}^{c'} (M \mathbb{E}_{\boldsymbol{\mu}_{\boldsymbol{\mathcal{X}}}}[f_{\mathbf{b}_i}] + N \mathbb{E}_{\boldsymbol{\mu}_{\boldsymbol{\mathcal{O}}}}[f_{\mathbf{b}_i}])
\end{aligned}$$

Note that $\boldsymbol{\mu}_{\boldsymbol{\mathcal{X}}}, \boldsymbol{\mu}_{\boldsymbol{\mathcal{O}}}$ arise by discretizing the continuous uniform measures $\boldsymbol{\mu}_{\mathbb{S}^{D-1}}$ and $\boldsymbol{\mu}_{\mathbb{S}^{D-1} \cap \boldsymbol{\mathcal{S}}}$ respectively ($\boldsymbol{\mu}_{\mathbb{S}^{D-1} \cap \boldsymbol{\mathcal{S}}}$ denotes the uniform measure on $\mathbb{S}^{D-1} \cap \mathcal{S}$). That being said, the continuous version of DPCP can be simply stated by replacing $\boldsymbol{\mu}_{\boldsymbol{\mathcal{X}}}, \boldsymbol{\mu}_{\boldsymbol{\mathcal{O}}}$ with $\boldsymbol{\mu}_{\mathbb{S}^{D-1}}$ and $\boldsymbol{\mu}_{\mathbb{S}^{D-1} \cap \boldsymbol{\mathcal{S}}}$ in equation 30 as follows,

$$\min_{\mathbf{B}} \sum_{i=1}^{c'} \left(M \mathbb{E}_{\boldsymbol{\mu}_{\mathbb{S}^{D-1}}}[f_{\mathbf{b}_i}] + N \mathbb{E}_{\boldsymbol{\mu}_{\mathbb{S}^{D-1} \cap \boldsymbol{\mathcal{S}}}}[f_{\mathbf{b}_i}] \right) \tag{30}$$

■

A.2 Proof of Corollary 3

Corollary 3: *As $k \to \infty$, a projected subgradient algorithm consisting of the steps described in (12) and for specific step-size updates (piecewise geometrically diminishing step size rule) will almost surely converge to a matrix $\hat{\boldsymbol{B}}^* \in \mathbb{R}^{D \times c'}$ whose columns $\hat{\boldsymbol{b}}_i^*,\ i = 1, 2, \ldots, c'$ will be normal vectors of the inliers' subspace when randomly initialized with independent vectors $\boldsymbol{b}_i^0 \in \mathbb{S}^{D-1},\ i = 1, 2, \ldots, c'$ uniformly distributed over the sphere $\mathbb{S}^{D-1}$.*

Proof: The proof can be trivially obtained by noticing a) that the condition for convergence i.e., inequality (9) of the projected subgradient algorithm given in Theorem 1 becomes $\theta_i^0 < \frac{\pi}{2}$ in the continuous case (since $\eta_{\boldsymbol{X}} \to 0$, $\eta_{\boldsymbol{O}} \to 0$, $c_{\boldsymbol{X},\min} \to c_d > 0$) and b) the set of unit ℓ_2-norm vectors $\boldsymbol{b}_i^0$s, $i = 1, 2, \ldots, c'$ sampled independently by a uniform distribution over the sphere and whose principal angle θ_i^0 is $\frac{\pi}{2}$ form the inliers' subspace has measure 0. ■

A.3 Proof of Theorem 5

Lemma 4 *The PSGM iterates $\hat{\boldsymbol{b}}_i^k,\ i = 1, 2, \ldots, c',\ k = 1, 2, \ldots$ given in (12), when randomly initialized with $\hat{\boldsymbol{b}}_i^0$s, $i = 1, 2, \ldots, c'$ that are independently drawn from a spherical distribution with unit ℓ_2 norm converge almost surely to c' normal vectors of the inliers subspace $\mathcal{S}$ denoted as $\hat{\boldsymbol{b}}_i^*,\ i = 1, 2, \ldots, c'$ that are given by*

$$\hat{\boldsymbol{b}}_i^* = \frac{\mathcal{P}_{\mathcal{S}_\perp}(\hat{\boldsymbol{b}}_i^0)}{\|\mathcal{P}_{\mathcal{S}_\perp}(\hat{\boldsymbol{b}}_i^0)\|_2}, \quad i = 1, 2, \ldots, c' \tag{31}$$

Proof Let us assume $\boldsymbol{b}_i^0 = \hat{\boldsymbol{b}}_i^0$. The iterates of subgradients steps of PSGM can be written in the following form,

$$\begin{aligned}
\boldsymbol{b}_i^1 &= (1 - \mu_i^0 M c_D) \hat{\boldsymbol{b}}_i^0 - \mu_i^0 N c_d \hat{\mathbf{s}} \\
\boldsymbol{b}_i^2 &= (1 - \mu_i^1 M c_D) \hat{\boldsymbol{b}}_i^1 - \mu_i^1 N c_d \hat{\mathbf{s}} \\
\vdots &= \qquad \vdots \\
\boldsymbol{b}_i^K &= (1 - \mu_i^{K-1} M c_D) \hat{\boldsymbol{b}}_i^{K-1} - \mu_i^{K-1} N c_d \hat{\mathbf{s}}
\end{aligned} \tag{32}$$

By projecting each update of PSGM onto $\mathcal{S}_\perp$ and since $\mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{b}_i^k) = \|\boldsymbol{b}_i^k\|\mathcal{P}_{\mathcal{S}_\perp}(\hat{\boldsymbol{b}}_i^k)$ we have,

$$\mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{b}_i^{k+1}) = \frac{(1-\mu_i^k M c_D)}{\|\boldsymbol{b}_i^k\|}\mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{b}_i^k) \tag{33}$$

We can thus easily derive the following form for $\mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{b}_i^K)$,

$$\mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{b}_i^K) = \left(\prod_{k=1}^{K-1}\frac{(1-\mu_i^k M c_D)}{\|\boldsymbol{b}_i^k\|_2}\right)\mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{b}_i^0) \tag{34}$$

We know from Theorem 1 and Corollary 4 when initialized with $\boldsymbol{b}_i^0,\ i = 1, 2, \ldots, c'$s randomly drawn according according to a spherical distribution DPCP-PSGM will almost surely converge as $K \to \infty$ to vectors $\hat{\boldsymbol{b}}_i^*,\ i = 1, 2, \ldots, c'$ i.e., $\hat{\boldsymbol{b}}_i^K \to \hat{\boldsymbol{b}}_i^*$ where $\hat{\boldsymbol{b}}_i^* \in \mathcal{S}_\perp$. Hence $\mathcal{P}_{\mathcal{S}_\perp}(\hat{\boldsymbol{b}}_i^K) \to \hat{\boldsymbol{b}}_i^*$ as $K \to \infty$. Note that from Theorem 1 we have that $\mu_i^k \neq \frac{1}{Mc_D} \quad \forall k = \{1, 2, \ldots, K\}$ hence $\prod_{k=1}^{K-1}\frac{(1-\mu_i^k M c_D)}{\|\hat{\boldsymbol{b}}_i^k\|_2} \neq 0$. From 34 and after projecting on the unit sphere and we thus have $\hat{\boldsymbol{b}}_i^* = \frac{\mathcal{P}_{\mathcal{S}_\perp}(\hat{\boldsymbol{b}}_i^0)}{\|\mathcal{P}_{\mathcal{S}_\perp}(\hat{\boldsymbol{b}}_i^0)\|_2}$. ■

Theorem 5 *Let $\hat{\boldsymbol{B}}^0 \in \mathbb{R}^{D\times c'}$ where $c' \geq c$ with c denoting the true codimension of the inliers subspace $\mathcal{S}$, consisting of unit ℓ_2 norm column vectors $\hat{\boldsymbol{b}}_i^0 \in \mathbb{S}^{D-1}, i = 1, 2, \ldots, c'$ that are independently drawn from uniform distribution over the sphere $\mathbb{S}^{D-1}$. A PSGM algorithm initialized with $\hat{\boldsymbol{B}}^0$ will almost surely converge to a matrix $\hat{\boldsymbol{B}}^*$ such that* $\operatorname{span}(\hat{\boldsymbol{B}}^*) \equiv \mathcal{S}_\perp$.

Proof From Lemma 4 we have that for each initial unit norm vector $\boldsymbol{b}_i^0$ which corresponds to the ith column of $\boldsymbol{B}^0$ will almost surely converge to $\hat{\boldsymbol{b}}_i^* = \frac{\mathcal{P}_{\mathcal{S}_\perp}(\hat{\boldsymbol{b}}_i^0)}{\|\mathcal{P}_{\mathcal{S}_\perp}(\hat{\boldsymbol{b}}_i^0)\|_2}$. We can thus write $\boldsymbol{B}^* = \mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{B}^0)\boldsymbol{\Gamma}$ where $\boldsymbol{\Gamma}$ is a full-rank diagonal matrix given be $\boldsymbol{\Gamma} = \operatorname{diag}(\frac{1}{\|\mathcal{P}_{\mathcal{S}_\perp}(\hat{\boldsymbol{b}}_1^0)\|_2}, \frac{2}{\|\mathcal{P}_{\mathcal{S}_\perp}(\hat{\boldsymbol{b}}_2^0)\|_2}, \ldots, \frac{1}{\|\mathcal{P}_{\mathcal{S}_\perp}(\hat{\boldsymbol{b}}_{c'}^0)\|_2})$. Note that $\mathcal{P}_{\mathcal{S}_\perp}$ is a linear projection and thus we can write $\mathcal{P}_{\mathcal{S}_\perp}(\boldsymbol{B}^0) = \boldsymbol{B}_{\mathcal{S}_\perp}\boldsymbol{B}_{\mathcal{S}_\perp}^\top$ where $\mathbf{B}_{\mathcal{S}_\perp}\mathbb{R}^{D\times c}$ is an orthonormal matrix which spans $\mathcal{S}_\perp$. Note that the probability of sampling a low-rank matrix $\boldsymbol{B}_0 = [\boldsymbol{b}_1^0, \boldsymbol{b}_2^0, \ldots, \boldsymbol{b}_{c'}^0]$ when columns $\boldsymbol{b}_i^0$s are randomly and independently drawn from a spherical distribution is zero. We thus have $\boldsymbol{B}^* = \boldsymbol{B}_{\mathcal{S}_\perp}\boldsymbol{B}_{\mathcal{S}_\perp}^\top\boldsymbol{B}^0\boldsymbol{\Gamma}$ with $\operatorname{rank}(\boldsymbol{B}^*) = c$. ■

Lemma 6 *If $\sigma_{c'}(\boldsymbol{B}^0) \geq \|\hat{\boldsymbol{\Delta}}\|_2$ then the rank of matrix $\boldsymbol{B}^0 - \hat{\boldsymbol{\Delta}}$ equals c'.*

Proof Let $\boldsymbol{A} = \boldsymbol{B}^0 - \hat{\boldsymbol{\Delta}}$. From singular value perturbation inequalities we have $|\sigma_i(\boldsymbol{A}) - \sigma_i(\boldsymbol{B}^0)| \leq \|\hat{\boldsymbol{\Delta}}\|_2$, for $i = 1, 2, \ldots, c'$. Hence it holds,

$$-\|\hat{\boldsymbol{\Delta}}\|_2 \leq \sigma_i(\boldsymbol{A}) - \sigma_i(\boldsymbol{B}^0) \tag{35}$$

If $\sigma_{c'}(\boldsymbol{A}) > 0$ the from equation 35 we get

$$\sigma_{c'}(\boldsymbol{B}^0) > \|\hat{\boldsymbol{\Delta}}\|_2 \tag{36}$$

■

A.4 Proof of Theorem 7

We first give the following Lemmas:

Lemma 7 *For the ℓ_2 norm of $\boldsymbol{e}_{\boldsymbol{O}}^{i,k}$ for any $k = 1, 2, \ldots, K$ and $i = 1, 2, \ldots, c'$ it holds,*

$$\|\boldsymbol{e}_{\boldsymbol{O}}^{i,k}\|_2 \leq \eta_{\boldsymbol{\mathcal{O}}} + c_{\boldsymbol{\mathcal{O}},\max} - c_d \tag{37}$$

Proof

$$\begin{aligned}\|\boldsymbol{e}_{\boldsymbol{O}}^{i,k}\|_2 = \|\boldsymbol{o}_b - c_d\mathbf{b}\|_2 &= \|\frac{1}{M}\boldsymbol{\mathcal{O}}\mathrm{Sgn}(\boldsymbol{\mathcal{O}}^\top\mathbf{b}) - c_d\mathbf{b}\|_2 \\ &= \|\frac{1}{M}(\mathbf{I}-\mathbf{b}\mathbf{b}^\top)\boldsymbol{\mathcal{O}}\mathrm{Sgn}(\boldsymbol{\mathcal{O}}^\top\mathbf{b}) + \frac{1}{M}\mathbf{b}\mathbf{b}^\top\boldsymbol{\mathcal{O}}\mathrm{Sgn}(\boldsymbol{\mathcal{O}}^\top\mathbf{b}) - c_d\mathbf{b}\|_2 \\ &\leq \|\frac{1}{M}(\mathbf{I}-\mathbf{b}\mathbf{b}^\top)\boldsymbol{\mathcal{O}}\mathrm{Sgn}(\boldsymbol{\mathcal{O}}^\top\mathbf{b})\|_2 + (\frac{1}{M}\|\boldsymbol{\mathcal{O}}^\top\mathbf{b}\|_1 - c_d)\|\mathbf{b}\|_2 \\ &\leq \boldsymbol{\eta}_{\boldsymbol{\mathcal{O}}} + c_{\boldsymbol{\mathcal{O}},\max} - c_d\end{aligned} \tag{38}$$

■

Lemma 8 *Let the step size of Algorithm 1(DPCP-PSGM) μ_i^k being updated following the piecewise geometrically diminishing step size rule with*

$$\beta < \left(\frac{1-\mu_i^0 M c_D}{1+\mu_i^0\left(N(\boldsymbol{\eta}_{\boldsymbol{\mathcal{X}}}+c_{\boldsymbol{\mathcal{X}},\max})+M(\boldsymbol{\eta}_{\boldsymbol{\mathcal{O}}}+c_{\boldsymbol{\mathcal{O}},\max})\right)}\right)^{K_*}.$$

For the spectral norm of $\hat{\boldsymbol{\Delta}}$ it holds $\|\hat{\boldsymbol{\Delta}}\|_2 < \sqrt{c'}\kappa(\boldsymbol{\eta}_{\boldsymbol{\mathcal{O}}}+c_{\boldsymbol{\mathcal{O}},\max}-c_d)$ where $\kappa = \max_i \frac{M\mu_i^0}{\beta^{K_0/K_*}(1-r_i)}$ and $r_i = \frac{\left(1+\mu_i^0(N(\boldsymbol{\eta}_{\boldsymbol{\mathcal{X}}}+c_{\boldsymbol{\mathcal{X}},\max})+M(\boldsymbol{\eta}_{\boldsymbol{\mathcal{O}}}+c_{\boldsymbol{\mathcal{O}},\max}))\right)}{1-\mu_i^0 M c_D}\beta^{1/K_*}$.

Proof We first bound the ℓ_2 norm of vectors $\mathbf{b}_i^j$s. We have that $\forall i = 1,2,\ldots,c'$ and $j = 1,2,\ldots,K$ it holds

$$\boldsymbol{b}_i^{j+1} = \hat{\boldsymbol{b}}_i^j - \mu_i^j\left(\boldsymbol{\mathcal{X}}\mathrm{Sgn}(\boldsymbol{\mathcal{X}}^\top\hat{\boldsymbol{b}}_i^j) + \boldsymbol{\mathcal{O}}\mathrm{Sgn}(\boldsymbol{\mathcal{O}}^\top\hat{\boldsymbol{b}}_i^j)\right) \tag{39}$$

We define the quantities

$$\boldsymbol{\eta}_{\boldsymbol{\mathcal{X}}} \coloneqq \max_{\mathbf{b}\in\mathbb{S}^{D-1}} \frac{1}{N}\|(\mathcal{P}_{\mathcal{S}} - \hat{\mathbf{b}}_i^j\hat{\mathbf{b}}_i^{j,\top})\boldsymbol{\mathcal{X}}\mathrm{Sgn}(\boldsymbol{\mathcal{X}}^\top\hat{\mathbf{b}}_i^j)\|_2 \tag{40}$$

$$c_{\boldsymbol{X},\max} \coloneqq \max_{\mathbf{b}\in\mathbb{S}^{D-1}} \frac{1}{N}\|\boldsymbol{\mathcal{X}}^\top\hat{\mathbf{b}}_i^j\|_1 \tag{41}$$

$\|\hat{\mathbf{b}}_i^j\|_2 = 1$ hence

$$\begin{aligned}\|\boldsymbol{b}_i^{j+1}\|_2 &\leq 1 + \mu_i^j\|\boldsymbol{\mathcal{X}}\mathrm{Sgn}(\boldsymbol{\mathcal{X}}^\top\hat{\boldsymbol{b}}_i^j) + \boldsymbol{\mathcal{O}}\mathrm{Sgn}(\boldsymbol{\mathcal{O}}^\top\hat{\boldsymbol{b}}_i^j\|_2 \\ &\leq 1 + \mu_i^j\left(\|\boldsymbol{\mathcal{X}}\mathrm{Sgn}(\boldsymbol{\mathcal{X}}^\top\hat{\boldsymbol{b}}_i^j)\|_2 + \|\boldsymbol{\mathcal{O}}\mathrm{Sgn}(\boldsymbol{\mathcal{O}}^\top\hat{\mathbf{b}}_i^j)\|_2\right) \\ &\leq 1 + \mu_i^j(\|(\mathcal{P}_{\mathcal{S}} - \hat{\mathbf{b}}_i^j\hat{\boldsymbol{b}}_i^{j,\top})\boldsymbol{\mathcal{X}}\mathrm{Sgn}(\boldsymbol{\mathcal{X}}^\top\hat{\boldsymbol{b}}_i^j)\|_2 + \|(\hat{\mathbf{b}}_i^j\hat{\boldsymbol{b}}_i^{j,\top})\boldsymbol{\mathcal{X}}\mathrm{Sgn}(\boldsymbol{\mathcal{X}}^\top\hat{\boldsymbol{b}}_i^j)\|_2 \\ &+ \|\left(1-\hat{\boldsymbol{b}}_i^j\hat{\mathbf{b}}_i^{j,\top}\right)\boldsymbol{\mathcal{O}}\mathrm{Sgn}(\boldsymbol{\mathcal{O}}^\top\hat{\mathbf{b}}_i^j)\|_2 + \|\hat{\boldsymbol{b}}_i^j\hat{\mathbf{b}}_i^{j,\top}\boldsymbol{\mathcal{O}}\mathrm{Sgn}(\boldsymbol{\mathcal{O}}^\top\hat{\boldsymbol{b}}_i^j)\|_2) \\ &\leq 1 + \mu_i^j\left(N(\boldsymbol{\eta}_{\boldsymbol{\mathcal{X}}}+c_{\boldsymbol{\mathcal{X}},\max})+M(\boldsymbol{\eta}_{\boldsymbol{\mathcal{O}}}+c_{\boldsymbol{\mathcal{O}},\max})\right)\end{aligned}$$

Due to equation 42 and since μ_i^j follows a non-increasing path as $j \to k$, the scalar term $\prod_{j=0}^k \frac{\|\mathbf{b}_i^j\|_2}{(1-\mu_i^j M c_D)}$ is bounded above as follows,

$$\prod_{j=0}^k \frac{\|\boldsymbol{b}_i^j\|_2}{(1-\mu_i^j M c_D)} \leq \left(\frac{\left(1+\mu_i^0\left(N(\boldsymbol{\eta}_{\boldsymbol{\mathcal{X}}}+c_{\boldsymbol{\mathcal{X}},\max})+M(\boldsymbol{\eta}_{\boldsymbol{\mathcal{O}}}+c_{\boldsymbol{\mathcal{O}},\max})\right)\right)}{1-\mu_i^0 M c_D}\right)^k \tag{42}$$

We now focus on the geometrically diminishing step size rule given in equation 10. We have $\mu_i^k = \mu_i^0\beta^{\lfloor k-K_0/K_*\rfloor+1} < \mu_i^0\beta^{(k-K_0)/K_*}$ for $k \geq K_0$ and $\mu_i^0\beta^{(k-K_0)/K_*} > \mu_i^0$ for $k < K_0$. Hence

we can get the following upper bound

$$\lim_{K\to\infty}\sum_{k=0}^{K-1}\prod_{j=0}^{k}\frac{\|\boldsymbol{b}_i^j\|_2}{(1-\mu_i^j Mc_D)}\mu_i^k M <$$
$$M\lim_{K\to\infty}\sum_{k=0}^{K-1}\left(\frac{\left(1+\mu_i^0\left(N(\boldsymbol{\eta}_{\boldsymbol{\mathcal{X}}}+c_{\boldsymbol{\mathcal{X}},\max})+M(\boldsymbol{\eta}_{\boldsymbol{\mathcal{O}}}+c_{\boldsymbol{\mathcal{O}},\max})\right)\right)}{1-\mu_i^0 Mc_D}\right)^k\mu_i^0\beta^{(k-K_0)/K_*}$$
$$\equiv\lim_{K\to\infty}\sum_{k=0}^{K-1}M\frac{1}{\beta^{K_0/K_*}}\mu_i^0\left(\frac{\left(1+\mu_i^0\left(N(\boldsymbol{\eta}_{\boldsymbol{\mathcal{X}}}+c_{\boldsymbol{\mathcal{X}},\max})+M(\boldsymbol{\eta}_{\boldsymbol{\mathcal{O}}}+c_{\boldsymbol{\mathcal{O}},\max})\right)\right)}{1-\mu_i^0 Mc_D}\beta^{1/K_*}\right)^k$$

The series $S=\sum_{k=0}^{K-1}\left(\frac{\left(1+\mu_i^0(N(\boldsymbol{\eta}_{\boldsymbol{\mathcal{X}}}+c_{\boldsymbol{\mathcal{X}},\max})+M(\boldsymbol{\eta}_{\boldsymbol{\mathcal{O}}}+c_{\boldsymbol{\mathcal{O}},\max}))\right)}{1-\mu_i^0 Mc_D}\beta^{1/K_*}\right)^k$ is geometric and if

$$\beta<\left(\frac{1-\mu_i^0 Mc_D}{\left(1+\mu_i^0\left(N(\boldsymbol{\eta}_{\boldsymbol{\mathcal{X}}}+c_{\boldsymbol{\mathcal{X}},\max})+M(\boldsymbol{\eta}_{\boldsymbol{\mathcal{O}}}+c_{\boldsymbol{\mathcal{O}},\max})\right)\right)}\right)^{K_*} \tag{43}$$

it converges as $K\to\infty$ to $\frac{1}{1-r_i}$ where $r_i=\frac{\left(1+\mu_i^0(N(\boldsymbol{\eta}_{\boldsymbol{\mathcal{X}}}+c_{\boldsymbol{\mathcal{X}},\max})+M(\boldsymbol{\eta}_{\boldsymbol{\mathcal{O}}}+c_{\boldsymbol{\mathcal{O}},\max}))\right)}{1-\mu_i^0 Mc_D}\beta^{1/K_*}$.

Let us now bound the ℓ_2 norms of the columns of $\hat{\boldsymbol{\Delta}}$. From Lemma 7 we have $\|\boldsymbol{e}_{\boldsymbol{\mathcal{O}}}^{i,k}\|_2\leq\boldsymbol{\eta}_{\boldsymbol{\mathcal{O}}}+c_{\boldsymbol{\mathcal{O}},\max}-c_d$. We can easily thus derive that $\|\hat{\boldsymbol{\delta}}_i\|_2\leq\kappa(\boldsymbol{\eta}_{\boldsymbol{\mathcal{O}}}+c_{\boldsymbol{\mathcal{O}},\max}-c_d)$. For the spectral norm of $\hat{\boldsymbol{\Delta}}$ we thus have

$$\begin{aligned}\|\hat{\boldsymbol{\Delta}}\|_2=\sup_{\boldsymbol{x}\neq\boldsymbol{0}}\frac{\|\hat{\boldsymbol{\Delta}}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}=\sup_{\boldsymbol{x}\neq\boldsymbol{0}}\frac{\|\sum_{i=1}^{c'}\hat{\boldsymbol{\delta}}_i x_i\|_2}{\|\boldsymbol{x}\|_2}\leq\sup_{\boldsymbol{x}\neq\boldsymbol{0}}\frac{\sum_{i=1}^{c'}\|\hat{\boldsymbol{\delta}}_i\|_2|x_i|}{\|\boldsymbol{x}\|_2}\\ \leq\max_i\|\hat{\boldsymbol{\delta}}_i\|_2\sup_{\boldsymbol{x}\neq\boldsymbol{0}}\frac{\|\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_2}\leq\kappa(\boldsymbol{\eta}_{\boldsymbol{\mathcal{O}}}+c_{\boldsymbol{\mathcal{O}},\max}-c_d)\sqrt{c'}\end{aligned} \tag{44}$$

Where the last inequality arises since $\|\boldsymbol{x}\|_1\leq\sqrt{c'}\|\boldsymbol{x}\|_2$. ■

We then give the Theorem.

Theorem 9 *(Theorem 5.58 of Vershynin (2010)) Let $\boldsymbol{B}$ be a $D\times d$ matrix ($D\geq d$) whose columns $\boldsymbol{b}_i$ are independent sub-gaussian isotropic random vectors in $\mathbb{R}^D$ with $\|\boldsymbol{b}_i\|_2=\sqrt{D}$ almost surely. Then for every $t\geq 0$ the inequality*

$$\sqrt{D}-C\sqrt{d}-t\leq\sigma_{\min}(\boldsymbol{B})\leq\sigma_{\max}(\boldsymbol{B})\leq\sqrt{D}+C\sqrt{d}+t \tag{45}$$

with probability at least $1-2\exp(-ct^2)$, where $C=C'_k, c=c'_K>0$ depend only on the subgaussian norm $K=\max_j\|\boldsymbol{b}_i\|_{\psi_2}$ of the columns.

The proof of Theorem 7 follows next.

Theorem 7 *Let $\mathbf{B}^0\in\mathbb{R}^{D\times c'}$ with columns randomly sampled from a unit ℓ_2 norm spherical distribution where $c'\geq c$ with c denoting the true codimension of the inliers subspace $\mathcal{S}$ that satisfies Assumption 1. If*

$$1-C_1\sqrt{\frac{c'}{D}}-\frac{\epsilon}{\sqrt{D}}>\sqrt{c'}\kappa(\boldsymbol{\eta}_{\boldsymbol{\mathcal{O}}}+c_{\boldsymbol{\mathcal{O}},\max}-c_d) \tag{46}$$

where $\kappa=\max_i\frac{M\mu_i^0}{\beta^{K_0/K_*}(1-r_i)}$ and $r_i=\frac{\left(1+\mu_i^0(N(\boldsymbol{\eta}_{\boldsymbol{\mathcal{X}}}+c_{\boldsymbol{\mathcal{X}},\max})+M(\boldsymbol{\eta}_{\boldsymbol{\mathcal{O}}}+c_{\boldsymbol{\mathcal{O}},\max}))\right)}{1-\mu_i^0 Mc_D}\beta^{1/K_*}$ then with probability at least $1-2\exp(-\epsilon^2C_2)$ (where C_1,C_2 are absolute constants), Algorithm 1 with a geometrically diminishing step size rule will converge to a matrix $\hat{\boldsymbol{B}}^*$ such that $\operatorname{span}(\hat{\boldsymbol{B}}^*)\equiv\mathcal{S}_\perp$.

Proof By Assumption 1 we have that all columns of $\boldsymbol{B}_0$ will satisfy the sufficient condition for convergence of DPCP-PSGM (Algorithm 1) to a normal vector of $\mathcal{S}$. From Lemma 4 and we get the

Table 1: Results on Washinghton DC AVIRIS hyperspectral image

Methods	**F1-scores**	
	r = 80%	r = 90%
DPCP-PSGM (unknown c)	0.994	0.993
RSGM (unknown c)	0	0
DPCP-IRLS (unknown c)	0	0
RSGM ($c = 5$)	0.999	0.993
DPCP-IRLS $c = 5$	1	0.995

inequality $\sigma_{c'}(\boldsymbol{B}_0) < \|\hat{\boldsymbol{\Delta}}\|_2$ which ensures full-rankness of $\hat{\boldsymbol{B}}^*$, which is the key ingredient in order to prove that $\mathrm{span}(\hat{\boldsymbol{B}}^*) = \mathcal{S}_\perp$. We can then Use Theorem 7 for matrix $\boldsymbol{B}_0$. Note that columns of $\boldsymbol{B}_0$ are drawn independently and are uniformly distributed on the unit sphere. Hence, columns of $\boldsymbol{B}_0$ are sampled by subgaussian distribution and the LHS of the inequality of the theorem appears if we scale with $\frac{1}{\sqrt{D}}$ so that to create unit-norm columns and use LHS of the inequality of Theorem 7. The RHS of the inequality is due to the upper bound of $\|\hat{\boldsymbol{\Delta}}\|_2$ as stated in Lemma 8. The absolute costants C_1, C_2 depend only the subgaussian norm of the uniform istribution (they is no dependency on the dimensions of the problem).

■

B EXPERIMENTAL DETAILS AND ADDITIONAL MATERIAL

All experiments were conducted on a MacBook Pro 2.6GhZ 6-Core Intel Core i7, memory 16GB 2667 Mhz DDR using Matlab2019B. For computational purposes and in order to avoid fine-tuning of the piecewise geometrically diminishing (PGD) step size, the modified backtracking line-search (MBLS) step-size rule was adopted for DPCP-PSGM as proposed in Zhu et al. (2018).

B.1 OUTLIERS PURSUIT IN WASHINGTON DC MALL AVIRIS HSI

Hyperspectral images (HSIs) provide rich spectral information as compared to RGB images capturing a wide range of the electromagnetic spectrum. Washington DC Mall AVIRIS HSI contains contiguous spectral bands captured at 0.4 to 2.4μm region of visible and infrared spectrum, Giampouras et al. (2019). In this experiments we randomly choose 10 out of its 210 spectral bands. Due to high coherence in the both the spectral and the spatial domain, pixels of HSIs admit representations in low-dimensional subspaces. Here, we use a 100$\times$100 segment of the hyperspectral image selecting randomly 10 out of its $D = 210$ spectral bands. We form a matrix $\tilde{\boldsymbol{\mathcal{X}}}$ of size 10×10000 whose columns correspond to different points in the 10-dimensional ambient space. Then we corrupt columns of $\tilde{\boldsymbol{\mathcal{X}}}$ by replacing them with outliers that are generated uniformly at random with unit ℓ_2 norm for two different outliers' ratios i.e., $r = 0.8$ and $r = 0.9$. In the corrupted $\tilde{\boldsymbol{\mathcal{X}}}$, the remaining clear pixels are considered as the inliers. Table 1 displays the F1 scores obtained by DPCP-PSGM, RSGM and DPCP-IRLS algorithm. The latter two algorithms are evaluated in two scenarios: a) codimension is initialized $c' = 5$ and b) $c' = 10$. Given the singular value distribution of the initial image, we infer that the dimension d of the inliers' subspace is less or equal than 5.

Hence, $c' = 5$ (recall $c = D - d$) is close to the true codimension value while $c' = 10$ is an overestimate thereof. From Table 1, we can see that the proposed DPCP-PSGM succeeds in both outliers' ratios regardless its unawareness of the true codimension value. On the other hand, DPCP-IRLS and RSGM fail when initialized with $c = 10$ and this is attributed to the restrictions induced due to the orthogonality constraints they both impose. In Fig. 4 we provide annotated versions of the clean HSI, its corrupted by outliers version for outliers' ratio $r = 90\%$, and the annotated outliers as recovered by the proposed DPCP-PSGM, RSGM, RSGM with $c' = 5$ and DPCP-IRLS with $c' = 5$.

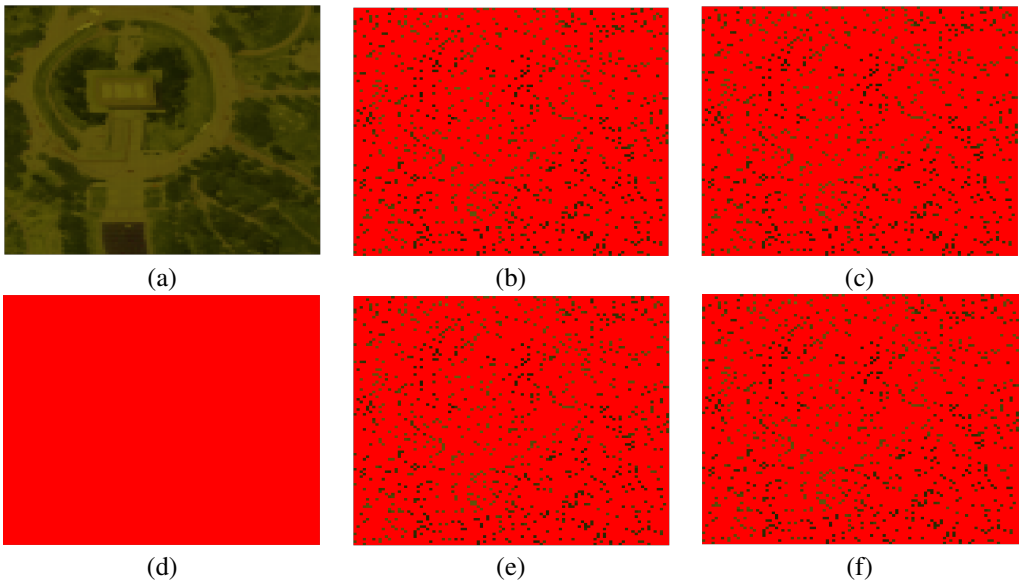

Figure 4: (a) False RGB color image of the clean version of Washington Mall AVIRIS HSI, (b) corrupted by outliers depicted with red and inliers correpsonding the non-red pixels (c) annotated outliers as recoverd by the proposed DPCP-PSGM method initialized with $c' = 10$ (d) RSGM with $c' = 10$, (e) RSGM with $c' = 5$ and (f) DPCP-IRLS with $c' = 5$.