Dual Principal Component Pursuit

Manolis C. Tsakiris and René Vidal
Center for Imaging Science, Johns Hopkins University
3400 N. Charles Street, Baltimore, MD, 21218, USA
m.tsakiris,rvidal@jhu.edu

Abstract

We consider the problem of outlier rejection in single subspace learning. Classical approaches work directly with a low-dimensional representation of the subspace. Our approach works with a dual representation of the subspace and hence aims to find its orthogonal complement. We pose this problem as an ℓ_1-minimization problem on the sphere and show that, under certain conditions on the distribution of the data, any global minimizer of this non-convex problem gives a vector orthogonal to the subspace. Moreover, we show that such a vector can still be found by relaxing the non-convex problem with a sequence of linear programs. Experiments on synthetic and real data show that the proposed approach, which we call Dual Principal Component Pursuit (DPCP), outperforms state-of-the art methods, especially in the case of high-dimensional subspaces.

1. Introduction

Principal Component Analysis (PCA) is one of the oldest [16, 11] and most fundamental techniques in data analysis, enjoying ubiquitous applications in modern science and engineering [12]. Given a data matrix $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{D \times L}$ of L data points of dimension D, PCA gives a closed form solution to the problem of fitting, in the Euclidean sense, a d-dimensional linear subspace to the columns of $\boldsymbol{\mathcal{X}}$. Even though the optimization problem associated with PCA is non-convex, it does admit a simple solution by means of the Singular Value Decomposition (SVD) of $\boldsymbol{\mathcal{X}}$. In fact, the d-dimensional subspace $\hat{\mathcal{V}}$ of $\mathbb{R}^D$ that is closest to the column span of $\boldsymbol{\mathcal{X}}$ is precisely the subspace spanned by the first d left singular vectors of $\boldsymbol{\mathcal{X}}$.

Using $\hat{\mathcal{V}}$ as a model for the data is meaningful when the data are known to have an approximately linear structure of underlying dimension d, i.e. they lie close to a d-dimensional subspace $\mathcal{V}$. In practice, the principal components of $\boldsymbol{\mathcal{X}}$ are known to be well-behaved under mild levels of noise, i.e., the angle between $\hat{\mathcal{V}}$ and $\mathcal{V}$ is relatively small and more importantly $\hat{\mathcal{V}}$ is optimal when the noise is Gaussian [12]. However, in the presence of even a few outliers in $\boldsymbol{\mathcal{X}}$, i.e., points whose angle from the underlying ground truth subspace $\mathcal{V}$ is large, the angle between $\mathcal{V}$ and its estimate $\hat{\mathcal{V}}$ will in general be large. This is to be expected since, by definition, the principal components are orthogonal directions of maximal correlation with *all* the points of $\boldsymbol{\mathcal{X}}$. This phenomenon, together with the fact that outliers are almost always present in real datasets, has given rise to the important problem of outlier detection in PCA.

Traditional outlier detection approaches come from robust statistics and include *Influence-based Detection*, *Multivariate Trimming*, *M-Estimators*, *Iteratively Weighted Recursive Least Squares* and *Random Sampling Consensus* (RANSAC) [12]. These methods are usually based on non-convex optimization problems, admit limited theoretical guarantees and have high computational complexity; for example, in the case of RANSAC many trials are required. Recently, two attractive methods have appeared [23, 19] with tight connections to *compressed sensing* [3] and *low-rank representation* [14]. Both of these methods are based on convex optimization problems and admit theoretical guarantees and efficient implementations. Remarkably, the self-expressiveness method of [19] does not require an upper bound on the number of outliers as the method of [23] does. However, they are both guaranteed to succeed only in the low-rank regime: the dimension d of the underlying subspace $\mathcal{V}$ associated to the inliers should be small compared to the ambient dimension D.

In this paper we adopt a *dual* approach to the problem of robust PCA in the presence of outliers, which allows us to transcend the low-rank regime of modern methods such as [23, 19]. The key idea of our approach comes from the fact that, in the absence of noise, the inliers lie inside any hyperplane $\mathcal{H}_1 = \mathrm{Span}(\boldsymbol{b}_1)^\perp$ that contains the underlying linear subspace $\mathcal{V}$. This suggests that, instead of attempting to fit directly a low-dimensional linear subspace to the entire data set, as done e.g. in [23], we can search for a hyperplane $\mathcal{H}_1$ that contains as many points of the dataset as possible. When the inliers are in general position inside the subspace, and the outliers are in general position outside the subspace, this hyperplane will ideally contain the entire set of inliers together with possibly a few outliers. After removing the points that do not lie in that hyperplane, the robust PCA problem is reduced to one with a potentially much smaller outlier percentage than in the original dataset. In fact, the number of outliers in the new dataset will be at most $D-2$, an upper bound that can be used to dramatically facilitate the outlier detection process using existing methods. We think of the direction $\boldsymbol{b}_1$ of the normal to the hyperplane $\mathcal{H}_1$ as a *dual principal component* of $\boldsymbol{\mathcal{X}}$, as ideally it is an element of $\mathcal{V}^\perp$. Naturally, one can continue by finding a second dual principal component by searching for a hyperplane $\mathcal{H}_2 = \operatorname{Span}(\boldsymbol{b}_2)^\perp$, with $\boldsymbol{b}_2 \perp \boldsymbol{b}_1$, that contains as many points as possible from $\boldsymbol{\mathcal{X}} \cap \mathcal{H}_1$, and so on, leading to a *Dual Principal Component Analysis* of $\boldsymbol{\mathcal{X}}$.

We pose the problem of searching for such hyperplanes as an ℓ_0 cosparsity-type problem, which we relax to a non-convex ℓ_1 problem on the sphere. We provide theoretical guarantees under which every global solution of that problem is a dual principal component. More importantly, we relax this non-convex optimization problem to a sequence of linear programming problems, which, after a finite number of steps, yields a dual principal component. Experiments on synthetic data demonstrate that the proposed method is able to handle more outliers and higher dimensional subspaces than the state-of-the-art methods [23, 19].

2. Problem Formulation

We begin by establishing our data model in Section 2.1, then we formulate our DPCP problem conceptually and computationally in Sections 2.2 and 2.3, respectively.

2.1. Data Model

We employ a deterministic noise-free data model, under which the inliers consist of N points $\boldsymbol{\mathcal{X}} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_N] \in \mathbb{R}^{D \times N}$ that lie in the intersection of the unit sphere $\mathbb{S}^{D-1}$ with an unknown proper subspace $\mathcal{V}$ of $\mathbb{R}^D$ of unknown dimension d. Accordingly, the outliers consist of M arbitrary points $\boldsymbol{\mathcal{O}} = [\boldsymbol{o}_1, \dots, \boldsymbol{o}_M] \in \mathbb{R}^{D \times M}$ that lie on $\mathbb{S}^{D-1}$. The dataset, that we assume given, is $\tilde{\boldsymbol{\mathcal{X}}} = [\boldsymbol{\mathcal{X}} \ \boldsymbol{\mathcal{O}}]\boldsymbol{\Gamma} \in \mathbb{R}^{D \times L}$, where $L = N + M$ and $\boldsymbol{\Gamma}$ is some permutation, indicating that the partition of the columns of $\tilde{\boldsymbol{\mathcal{X}}}$ into $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{O}}$ is unknown. We further assume that the columns of $\tilde{\boldsymbol{\mathcal{X}}}$ are in *general position* in the following sense: First, any d-tuple of inliers and any D-tuple of outliers is linearly indepenent. Second, for any $\{\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_{D-1}\} \subset \tilde{\boldsymbol{\mathcal{X}}}$, of which at most $d-1$ come from $\boldsymbol{\mathcal{X}}$, the hyperplane of $\mathbb{R}^D$ spanned by $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_{D-1}$ does not contain any of the remaining points.

2.2. Conceptual Formulation

Notice that in our data model we have made no assumption about the dimension of $\mathcal{V}$: indeed, $\mathcal{V}$ can be anything from a line to a $(D-1)$-dimensional hyperplane. Ideally, we would like to be able to partition the columns of $\tilde{\boldsymbol{\mathcal{X}}}$ into those that lie in $\mathcal{V}$ and those that don't. But under such generality, this is not a well-posed problem since $\boldsymbol{\mathcal{X}}$ lies inside every subspace that contains $\mathcal{V}$, which in turn may contain some elements of $\boldsymbol{\mathcal{O}}$. In other words, given $\tilde{\boldsymbol{\mathcal{X}}}$ and without any other a-priori knowledge, it may be impossible to correctly partition $\tilde{\boldsymbol{\mathcal{X}}}$ into $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{O}}$. Instead, we formulate the following well-posed problem:

Problem 1 *Partition the columns of $\tilde{\boldsymbol{\mathcal{X}}} \in \mathbb{R}^{D \times L}$ into two groups, such that one of the groups is a subset of $\tilde{\boldsymbol{\mathcal{X}}}$ with maximal cardinality, with respect to the property of lying inside a $(D-1)$-dimensional hyperplane of $\mathbb{R}^D$.*

The usefulness of this formulation is that for large values of $\gamma := M/N$, where known methods for outlier detection in PCA fail, one of the groups, say $\tilde{\boldsymbol{\mathcal{X}}}_1$ will contain the entire $\boldsymbol{\mathcal{X}}$ together with precisely $D-d-1$ columns of $\boldsymbol{\mathcal{O}}$, while the other group, say $\tilde{\boldsymbol{\mathcal{X}}}_2$, will contain the remaining $M-(D-d-1)$ columns of $\boldsymbol{\mathcal{O}}$. Note that the first group is *structured* in the sense that it must lie in a hyperplane and so in general $\dim \operatorname{Span}(\tilde{\boldsymbol{\mathcal{X}}}_1) = D-1$. Having the partition $\tilde{\boldsymbol{\mathcal{X}}} = \tilde{\boldsymbol{\mathcal{X}}}_1 \cup \tilde{\boldsymbol{\mathcal{X}}}_2$, we can reject the unstructured group $\tilde{\boldsymbol{\mathcal{X}}}_2$ and reconsider the Robust PCA problem on the group $\tilde{\boldsymbol{\mathcal{X}}}_1$. But now the number of outliers has decreased from γN to $D-d-1$. In fact, we can use the upper bound $D-2$ on the number of outliers to dramatically facilitate the outlier detection process using other existing methods.

2.3. Computational Formulation

A natural approach towards solving Problem 1 is to solve

$$\min_{\boldsymbol{b}} ||\tilde{\boldsymbol{\mathcal{X}}}^\top \boldsymbol{b}||_0 \text{ s.t. } \boldsymbol{b} \neq 0. \tag{1}$$

The idea behind (1) is that a hyperplane $\mathcal{H} = \operatorname{Span}(\boldsymbol{b})^\perp$ contains a maximal number of columns of $\tilde{\boldsymbol{\mathcal{X}}}$ if and only if $\tilde{\boldsymbol{\mathcal{X}}}^\top \boldsymbol{b}$ is as sparse as possible. Since (1) is intractable, consider

$$\min_{\boldsymbol{b}} ||\tilde{\boldsymbol{\mathcal{X}}}^\top \boldsymbol{b}||_1 \text{ s.t. } \|\boldsymbol{b}\|_2 = 1. \tag{2}$$

Notice that the objective in (2) is convex, while the constraint $\boldsymbol{b} \in \mathbb{S}^{D-1}$ is non-convex, thus leading to a non-smooth and non-convex optimization problem.

Problem 2 *When is every global solution $\boldsymbol{b}^*$ of* (2) *orthogonal to* $\operatorname{Span}(\boldsymbol{\mathcal{X}})$*? How can we efficiently solve* (2)*?*

In this paper, we propose to relax (2) by a sequence of linear programs of the form

$$\boldsymbol{n}_{k+1} := \operatorname*{argmin}_{\boldsymbol{b}^\top \hat{\boldsymbol{n}}_k = 1} \left\| \tilde{\boldsymbol{\mathcal{X}}}^\top \boldsymbol{b} \right\|_1, \tag{3}$$

where $\boldsymbol{n}_0$ is some arbitrary vector and $\hat{\cdot}$ indicates normalization to unit ℓ_2-norm. We naturally ask:

Problem 3 *Under what conditions does the sequence of* (3) *converge to a vector $\hat{\boldsymbol{n}}_\infty$ that is orthogonal to* $\operatorname{Span}(\boldsymbol{\mathcal{X}})$*?*

3. Related Work

In this section, we aim to familiarize the reader with the state-of-the-art of outlier detection in modern single subspace learning (Section 3.1), as well as give a brief overview (Section 3.2) of existing work, that relates technically to the problems of interest of this paper, i.e. problems (2) and (3).

3.1. Outlier Rejection in PCA

One of the oldest and most popular outlier detection methods in PCA is *Random Sampling Consensus (RANSAC)* [12]. The idea behind RANSAC is simple: alternate between randomly sampling $\hat{d}$ points from the dataset and computing a subspace model for these points, until a model is found that fits a maximal number of points in the entire dataset within some error ε. RANSAC is usually characterized by high performance, when not both $\hat{d}$ and the oultier percentage are large; otherwise it requires a high computational time, particularly when d is unknown and $\hat{d}$ is allowed to vary, since exponentially many trials are required in order to sample outlier-free subsets, and thus obtain reliable models. Moreover, its performance is very sensitive on in the input parameters $\hat{d}$ and ε.

Among many other outlier detection methods (see Section 1), in the remaining of this section we will focus on the modern low-rank/sparse-representation theoretic methods of [23] and [19], which we will later use experimentally to compare against our proposed method.

The first method [23], referred to as L21, is a variation of the *Robust PCA* algorithm of [13, 2], which computes a $(\ell_* + \ell_{21})$-norm decomposition[1] of the data matrix, instead of the $(\ell_* + \ell_1)$-decomopsition in [2]. More specifically, L21 solves the convex optimization problem

$$\min_{\boldsymbol{L},\boldsymbol{E}:\, \tilde{\boldsymbol{\mathcal{X}}}=\boldsymbol{L}+\boldsymbol{E}} \left\|\boldsymbol{L}\right\|_* + \lambda \left\|\boldsymbol{E}\right\|_{21}. \tag{4}$$

It is shown in [23] that, under certain conditions, the optimal solution to this problem is of the form $\boldsymbol{L} = [\boldsymbol{\mathcal{X}}\ \boldsymbol{0}_{D\times M}]\boldsymbol{\Gamma}$ and $\boldsymbol{E} = [\boldsymbol{0}_{D\times N}\ \boldsymbol{\mathcal{O}}]\boldsymbol{\Gamma}$. That is, the nonzero columns of the $\boldsymbol{L}$ matrix give the inliers and the nonzero columns of the $\boldsymbol{E}$ matrix give the outliers. However, the theoretical conditions require the intrinsic dimension $d = \dim \mathcal{V}$ and the outlier percentage to be small enough.

The second method that we consider, referred to as SE, is based on the *self-expressiveness* property of the data matrix, a notion popularized by the work of [4, 5] in the area of subspace clustering [22]. More specifically, if a column of $\tilde{\boldsymbol{\mathcal{X}}}$ is an inlier, then it can in principle be expressed as a linear combination of d other columns of $\tilde{\boldsymbol{\mathcal{X}}}$, which are inliers. If the column is instead an outlier, then it will in principle be expressible as a linear combination of not less than D other columns. To encourage each point to express itself as a linear combination of the smallest number of other data points, the following convex optimization problem is solved:

$$\min_{\boldsymbol{C}} \left\|\boldsymbol{C}\right\|_1, \ \ s.t.\ \tilde{\boldsymbol{\mathcal{X}}} = \tilde{\boldsymbol{\mathcal{X}}}\boldsymbol{C},\ \operatorname{Diag}(\boldsymbol{C}) = \boldsymbol{0}. \tag{5}$$

If d is small enough with respect to D, an element is declared as an outlier if the ℓ_1 norm of its coefficient vector in $\boldsymbol{C}$ is large; see [19] for an explicit formula. SE admits theoretical guarantees [19] and efficient ADMM implementations [5]. However, as it is clear from its description, it is expected to succeed only when d is sufficiently small. In contrast though to L21, SE has the remarkable property that it can, in principle, handle an arbitrary number of outliers.

3.2. Connections with Compressed Sensing and Dictionary Learning

Problems of the form

$$\min_{\boldsymbol{b}} ||\boldsymbol{\Omega b}||_0 \ \text{ s.t. }\ \boldsymbol{b} \neq 0, \tag{6}$$

and variants of its relaxations have appeared on several occasions and in diverse contexts in the literature, but are much less understood than the now classic sparse [1] and cosparse [15] problems of the form

$$\min_{\boldsymbol{x}} ||\boldsymbol{x}||_0 \ \text{ s.t. }\ \boldsymbol{Ax} = \boldsymbol{b} \tag{7}$$

$$\min_{\boldsymbol{x}} ||\boldsymbol{\Omega x}||_0 \ \text{ s.t. }\ \boldsymbol{Ax} = \boldsymbol{b}, \tag{8}$$

respectively. The main source of difficulty is that, in contrast to (8), obtaining tight convex relaxations of (6) is a hard problem. One of the first instances where (6) was considered was in the context of blind source separation [24], where it was proposed to relax it with the problem

$$\min_{\boldsymbol{b}} ||\boldsymbol{\Omega b}||_1 \ \text{ s.t. }\ ||\boldsymbol{b}||_2 \geq 1. \tag{9}$$

This is still a non-convex problem, and a heuristic based on quadratic programming was proposed to solve it.

It was not until very recently, that the convex relaxation

$$\min_{\boldsymbol{b}} ||\boldsymbol{\Omega b}||_1 \ \text{ s.t. }\ \boldsymbol{b}^\top \boldsymbol{w} = 1 \tag{10}$$

was proposed, with $\boldsymbol{w}$ taken to be a row or a sum of two rows of $\boldsymbol{\Omega}$, and theorems of correctness were given in the context of dictionary learning [20]. Notice that our proposed convex relaxations (3) can be seen as a generalization of (10). In the context of finding the sparsest vector in a subspace, which is intrinsically related to dictionary learning, an alternating direction minimization scheme was proposed in [17, 18] to solve a relaxation of the form

$$\min_{\boldsymbol{b},\boldsymbol{x}:\, ||\boldsymbol{b}||_2=1} ||\boldsymbol{\Omega b} - \boldsymbol{x}||_2^2 + \lambda \left\|\boldsymbol{x}\right\|_1. \tag{11}$$

[1] Here ℓ_* denotes the nuclear norm of a matrix, i.e., the sum of its singular values, and ℓ_{21} is defined as the sum of the Euclidean norms of the columns of a matrix.

Remarkably, under some mild conditions, this was shown to converge with high probability to a global solution of

$$\min_{\boldsymbol{b}} \|\boldsymbol{\Omega b}\|_1 \text{ s.t. } \|\boldsymbol{b}\|_2 = 1. \tag{12}$$

The geometry of (12) was further studied in a probabilistic framework in the recent [21], after replacing the ℓ_1-norm with a smooth surrogate.

4. Theoretical Analysis

In this section we state and discuss our main theoretical results[2], regarding problems (2) and (3). Before doing so though, we need to introduce additional notation and draw some interesting connections with the field of numerical integration on the sphere (Section 4.1).

4.1. An Integration Perspective

To begin with, for a vector $\boldsymbol{b} \in \mathbb{S}^{D-1}$, denote by $f_{\boldsymbol{b}} : \mathbb{S}^{D-1} \to \mathbb{R}^+$ the function $\boldsymbol{y} \mapsto \left|\boldsymbol{b}^\top \boldsymbol{y}\right|$. Then given a set of L points $\boldsymbol{Y} \subset \mathbb{S}^{D-1}$, the quantity

$$\frac{1}{L}\left\|\boldsymbol{Y}^\top \boldsymbol{b}\right\|_1 = \frac{1}{L}\sum_{j=1}^{L}\left|\boldsymbol{b}^\top \boldsymbol{y}_j\right| = \frac{1}{L}\sum_{j=1}^{L} f_{\boldsymbol{b}}(\boldsymbol{y}_j) \tag{13}$$

is a discrete approximation of the integral

$$\int_{\boldsymbol{y}\in\mathbb{S}^{D-1}} f_{\boldsymbol{b}}(\boldsymbol{y}) d\mu = \int_{\boldsymbol{y}\in\mathbb{S}^{D-1}} |\boldsymbol{y}^\top \boldsymbol{b}| d\mu, \tag{14}$$

where μ is the uniform measure on $\mathbb{S}^{D-1}$ and c_D is the mean height of the unit hemisphere of $\mathbb{R}^D$, given in closed form by

$$c_D = \frac{(D-2)!!}{(D-1)!!} \cdot \left\{ \begin{array}{ll} \frac{2}{\pi} & \text{if } D \text{ even} \\ 1 & \text{if } D \text{ odd} \end{array} \right. , \tag{15}$$

where the double factorial is defined as

$$k!! := \left\{ \begin{array}{ll} k(k-2)(k-4)\cdots 4\cdot 2 & \text{if } k \text{ even} \\ k(k-2)(k-4)\cdots 3\cdot 1 & \text{if } k \text{ odd} \end{array} \right. \tag{16}$$

A useful fact is that c_D is a decreasing function of D and in fact tends to zero as D goes to infinity.

Now, observe that because of the symmetry of $\mathbb{S}^{D-1}$, the integral in (14) does not depend on $\boldsymbol{b}$. However, the *integration error*

$$\left| c_D - \frac{1}{L}\sum_{j=1}^{L} f_{\boldsymbol{b}}(\boldsymbol{y}) \right| \tag{17}$$

does depend both on the direction of $\boldsymbol{b}$ as well as the distribution of the points $\boldsymbol{Y}$ on $\mathbb{S}^{D-1}$. It is clear though, that the more uniformly the points are distributed, the smaller will be the dependence of the integration error on the direction of $\boldsymbol{b}$. We note here that the notion of uniform point set distribution on the sphere is a non-trivial one. In a deterministic setting, this is an active subject of study in the fields of combinatorial geometry and numerical integration on the sphere [9, 8]. A widely used measure of the uniformity of a point set on the sphere is the so-called *point set discrepancy* $\mathcal{D}_L^S(\boldsymbol{Y})$ of the set, which can be defined in terms of spherical harmonics as

$$\mathcal{D}_L^S(\boldsymbol{Y}) := \sup_{m\geq 1} \frac{1}{m^D} \max_{i=1,\ldots,Z(D,m)} \left| \frac{1}{L}\sum_{j=1}^{L} S_{m,i}(\boldsymbol{y}_j) \right|, \tag{18}$$

where $Z(D,m)$ is the dimension of the vector space of spherical harmonics of order m, and $S_{m,i}$ is the i-th basis element. It is then a fact that the integration error is small if and only if $\mathcal{D}_L^S(\boldsymbol{Y})$ is small.

As before, for any $\boldsymbol{b} \in \mathbb{S}^{D-1}$ we define a vector valued function $\boldsymbol{f_b} : \mathbb{S}^{D-1} \to \mathbb{R}^D$ by $\boldsymbol{y} \xmapsto{\boldsymbol{f_b}} \text{Sign}(\boldsymbol{b}^\top \boldsymbol{y})\boldsymbol{y}$. Note that the image of $\boldsymbol{f_b}$ is $\mathbb{S}^{D-1} \cup \boldsymbol{0}$ and that points that are orthogonal to $\boldsymbol{b}$ are mapped to $\boldsymbol{0}$. Moreover,

Lemma 1 $\int_{\boldsymbol{y}\in\mathbb{S}^{D-1}} \text{Sign}(\boldsymbol{b}^\top \boldsymbol{y})\boldsymbol{y} d\mu = c_D\, \boldsymbol{b}, \;\; \forall \boldsymbol{b} \in \mathbb{S}^{D-1}$.

This result suggests that the quantity $\boldsymbol{y_b} := \frac{1}{L}\sum_{j=1}^{L} \text{Sign}(\boldsymbol{b}^\top \boldsymbol{y}_j)\boldsymbol{y}_j$ can be interpreted as a discrete approximation of the integral $\int_{\boldsymbol{y}\in\mathbb{S}^{D-1}} \boldsymbol{f_b}(\boldsymbol{y}) d\mu$ and so the more uniformly distributed are the points $\boldsymbol{Y}$, the closer $\boldsymbol{y_b}$ is to the quantity $c_D\, \boldsymbol{b}$.

The above discussion motivates defining the quantities $\epsilon_{\mathcal{O}}$ and $\epsilon_{\mathcal{X}}$, to capture the uniformity of outliers and inliers, respectively:

$$\epsilon_{\mathcal{O}} := \max_{\boldsymbol{b}\in\mathbb{S}^{D-1}} \|c_D\, \boldsymbol{b} - \boldsymbol{o_b}\|_2, \tag{19}$$

$$\boldsymbol{o_b} := \frac{1}{M}\sum_{j=1}^{M} \text{Sign}(\boldsymbol{b}^\top \boldsymbol{o}_j)\boldsymbol{o}_j, \tag{20}$$

$$\epsilon_{\mathcal{X}} := \max_{\boldsymbol{v}\in\mathbb{S}^{D-1}\cap\mathcal{V}} \|c_d\, \boldsymbol{v} - \boldsymbol{\chi_v}\|_2 \tag{21}$$

$$\boldsymbol{\chi_v} := \frac{1}{N}\sum_{j=1}^{N} \text{Sign}(\boldsymbol{v}^\top \boldsymbol{x}_j)\boldsymbol{x}_j. \tag{22}$$

4.2. The Non-Convex Problem

Before we consider the *discrete* non-convex problem (2), it is instructive to examine its continuous counterpart

$$\min_{\boldsymbol{b}^\top\boldsymbol{b}=1} \; M\int_{\boldsymbol{o}\in\mathbb{S}^{D-1}} \left|\boldsymbol{b}^\top \boldsymbol{o}\right| d\mu + \\ +N\int_{\boldsymbol{x}\in\mathcal{V}\cap\mathbb{S}^{D-1}} \left|\boldsymbol{b}^\top \boldsymbol{x}\right| d\sigma, \tag{23}$$

[2] All proofs are omitted due to space limitations.

where σ is the uniform measure on $\mathcal{V} \cap \mathbb{S}^{D-1}$. Of course this problem is only of theoretical interest and serves in establishing a first intuition for the idea behind (2). In fact,

Theorem 1 *Any global solution to problem* (23) *must be orthogonal to* $\mathcal{V}$.

The proof of the above theorem follows easily from the symmetry of the sphere, since the first integral appearing in (23) does not depend on $\boldsymbol{b}$, while the second integral depends only on the angle of $\boldsymbol{b}$ from $\mathcal{V}$.

Theorem 1 suggests that under sufficiently well-distributed point sets of inliers and outliers, any global solution to the discrete problem (2) should *also* be orthogonal to the span of the inliers. Before stating the precise result, we need one last piece of notation:

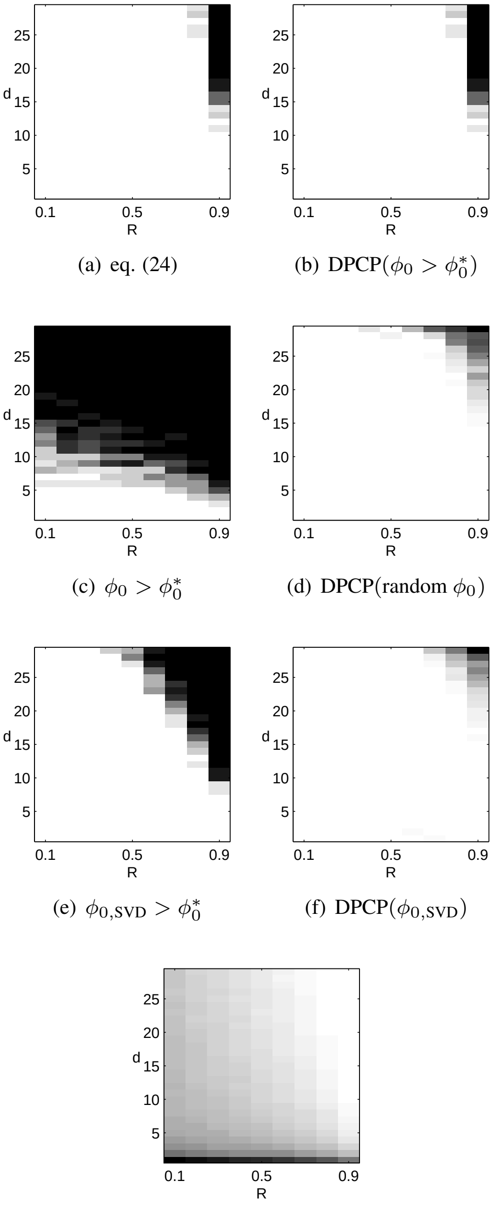
Definition 1 *For a set* $\boldsymbol{Y} = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_L] \subset \mathbb{S}^{D-1}$ *and integer* K*, define* $\mathcal{R}_{\boldsymbol{Y},K}$ *to be the maximum circumradius among all polytopes* $\operatorname{Conv}\left(\pm\boldsymbol{y}_{j_1} \pm \boldsymbol{y}_{j_2} \pm \cdots \pm \boldsymbol{y}_{j_K}\right)$*, where* $j_1, \dots, j_K$ *are distinct integers in* $[L]$*, and* $\operatorname{Conv}(\cdot)$ *indicates the convex hull operator.*

Theorem 2 *Suppose that the quantity* $\gamma := \frac{M}{N}$ *satisfies*

$$\gamma < \min\left\{\frac{c_d - \epsilon_{\boldsymbol{\mathcal{X}}}}{2\,\epsilon_{\boldsymbol{\mathcal{O}}}}, \frac{c_d - \epsilon_{\boldsymbol{\mathcal{X}}} - \left(\mathcal{R}_{\boldsymbol{\mathcal{O}},K_1} + \mathcal{R}_{\boldsymbol{\mathcal{X}},K_2}\right)/N}{\epsilon_{\boldsymbol{\mathcal{O}}}}\right\}, \tag{24}$$

for all positive integers K_1, K_2 *such that* $K_1 + K_2 \leq D - 1, K_2 \leq d - 1$. *Then any global solution* $\boldsymbol{b}^*$ *to* (2) *will be orthogonal to* $\operatorname{Span}(\boldsymbol{\mathcal{X}})$.

Towards interpreting this result, consider first the asymptotic case where we allow N and M to go to infinity, while keeping the ratio γ constant. Under point set uniformity, i.e. under the hypothesis that $\lim_{N\to\infty} D_N^S(\boldsymbol{\mathcal{X}}) = 0$ and $\lim_{M\to\infty} D_M^S(\boldsymbol{\mathcal{O}}) = 0$, we will have that $\lim_{N\to\infty} \epsilon_{\boldsymbol{\mathcal{X}}} = 0$ and $\lim_{M\to\infty} \epsilon_{\boldsymbol{\mathcal{O}}} = 0$, in which case (24) is satisfied. This suggests the interesting fact that when the number of inliers is a linear function of the number of outliers, then (2) will always give a normal to the inliers even for arbitrarily large number of outliers and irrespectively of the subspace dimension d. Along the same lines, for a given γ and under the point set uniformity hypothesis, we can always increase the number of inliers and outliers (thus decreasing $\epsilon_{\boldsymbol{\mathcal{X}}}$ and $\epsilon_{\boldsymbol{\mathcal{O}}}$), while keeping γ constant, until (24) is satisfed, once again indicating that (2) is possible to yield a normal to the space of inliers irrespectively of their intrinsic dimension.



(a) eq. (24) (b) DPCP($\phi_0 > \phi_0^*$)

(c) $\phi_0 > \phi_0^*$ (d) DPCP(random ϕ_0)

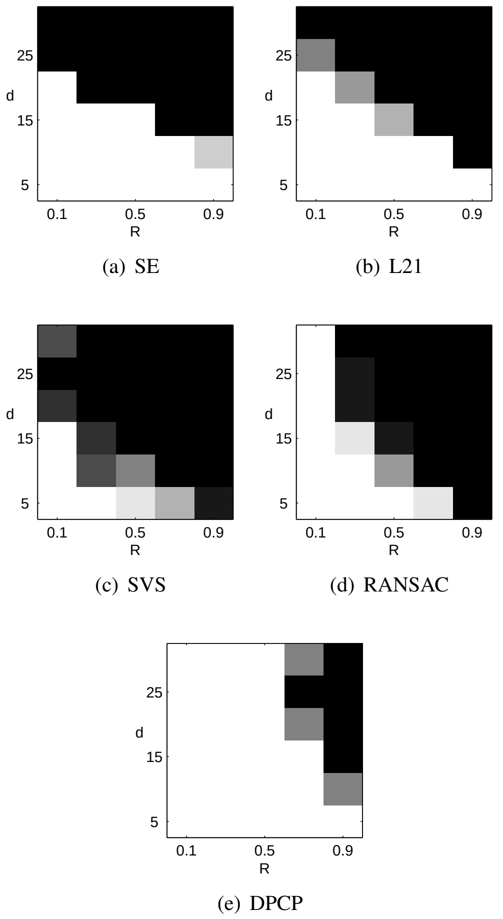
(e) $\phi_{0,\text{SVD}} > \phi_0^*$ (f) DPCP($\phi_{0,\text{SVD}}$)

(g) ϕ_0^*

Figure 1. Experimental analysis of DPCP (see subsection 6.1). (a): Empirical probability of (24) being true. (b): Angle from inlier space of the vector $\boldsymbol{b}$ computed by DPCP, with $\hat{\boldsymbol{n}}_0$ initialized with $\phi_0 > \phi_0^*$, when (24) is true. (c): Empirical probability of a random $\hat{\boldsymbol{n}}_0$ satisfying $\phi_0 > \phi_0^*$. (d): As in (b) but with $\hat{\boldsymbol{n}}_0$ initialized at random. (e): As in (c) but with $\hat{\boldsymbol{n}}_0$ initialized with SVD. (f): As in (d) but with $\hat{\boldsymbol{n}}_0$ initialized with SVD. (g): ϕ_0^* as given by (26).

4.3. The Sequence of Convex Relaxations

In this section we consider the sequence of convex relaxations (3); in particular, there are two important issues to be addressed. First, note that relaxing the constraint $\boldsymbol{b}^\top \boldsymbol{b} = 1$ in (2) with a linear constraint $\boldsymbol{b}^\top \hat{\boldsymbol{n}} = 1$ as in (10), has already been found to be of limited theoretical guarantees [20]. So it is natural to ask whether the idea of considering a sequence of such relaxations $\boldsymbol{b}^\top \hat{\boldsymbol{n}}_k = 1,\ k = 0, 1, \dots$ has an intrinsic merit or not, irrespectively of the data distribution. For example, if the data is *perfectly well distributed*, yet the sequence does not yield vectors orthogonal to the



(a) SE

(b) L21

(c) SVS

(d) RANSAC

(e) DPCP

Figure 2. Outlier/Inlier separation for the 5 compared methods.

inlier space, then we will know that a-priori the method is limited. Fortunately, this is not the case: when the data is perfectly well distributed, i.e. when we restrict our attention to the continuous analog of (3), given by

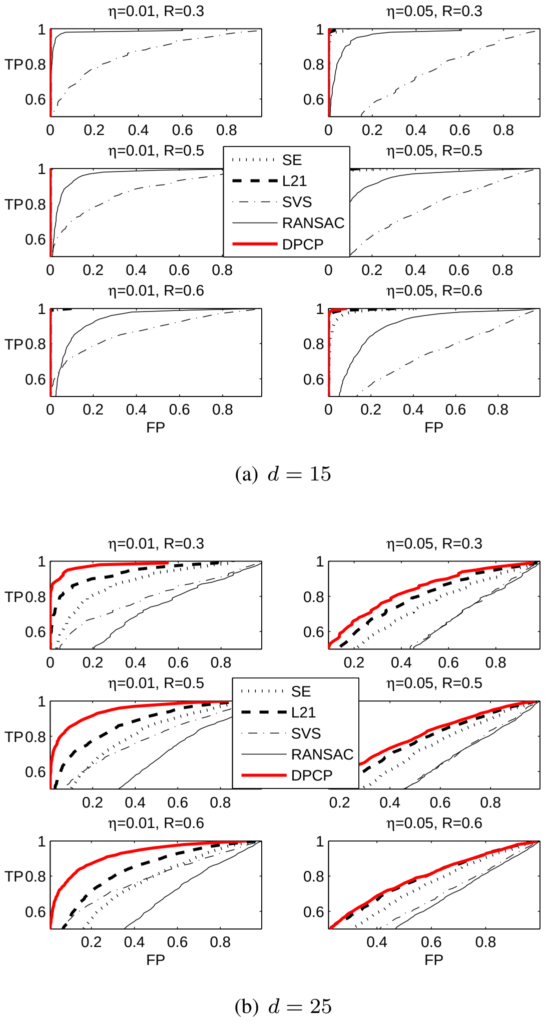
$$\boldsymbol{n}_{k+1} = \operatorname*{argmin}_{\boldsymbol{b}^\top \hat{\boldsymbol{n}}_k = 1} \left[M \int_{\boldsymbol{o} \in \mathbb{S}^{D-1}} \left| \boldsymbol{b}^\top \boldsymbol{o} \right| d\mu + N \int_{\boldsymbol{x} \in \mathcal{V} \cap \mathbb{S}^{D-1}} \left| \boldsymbol{b}^\top \boldsymbol{x} \right| d\sigma \right], \qquad (25)$$

then the sequence $\{\boldsymbol{n}_k\}$ achieves the property of interest:

Theorem 3 *Consider the sequence of vectors $\{\boldsymbol{n}_k\}$ generated by recursion (25), where $\hat{\boldsymbol{n}}_0 \in \mathbb{S}^{D-1}$ is arbitrary. Let $\{\phi_k\}$ be the corresponding sequence of angles from $\mathcal{V}$. Then $\lim_{k\to\infty} \phi_k = \frac{\pi}{2}$, provided that $\boldsymbol{n}_0 \notin \mathcal{V}$.*

This result suggests that relaxing $\boldsymbol{b}^\top \boldsymbol{b} = 1$ with the sequence $\boldsymbol{b}^\top \hat{\boldsymbol{n}}_k = 1, \; k \geq 0$ is intrinsically the right idea.

The second issue is how the distribution of the data affects the ability of this sequence of relaxations to give vectors orthogonal to $\mathcal{V}$. The answer is given by Theorem



(a) $d = 15$

(b) $d = 25$

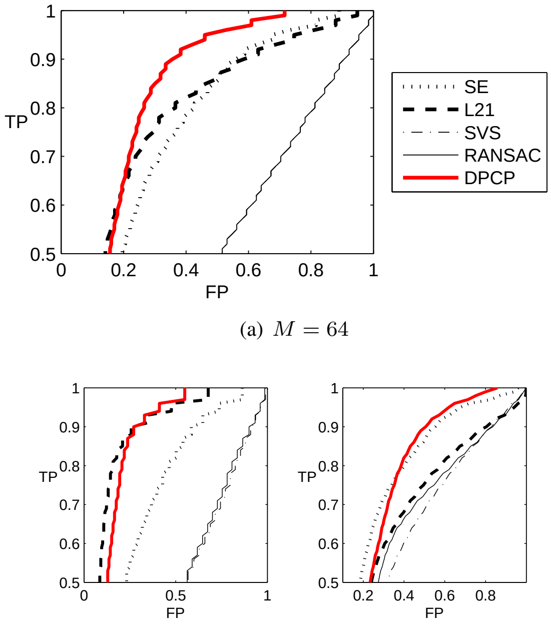
Figure 3. ROC curves as functions of noise percentage η, outlier percentage R, and subspace dimension d.

4, which says that when the angle between $\boldsymbol{n}_0$ and $\mathcal{V}$ is large enough and the data points are well distributed, the sequence (3) will consist of vectors orthogonal to the inlier space, for sufficiently large indices k.

Theorem 4 *Let ϕ_0 be the angle between $\boldsymbol{n}_0$ and $\mathcal{V}$. Suppose that condition (24) on the outlier ratio γ holds true and consider the vector sequence $\{\hat{\boldsymbol{n}}_k\}$ generated by recursion (3). Then after a finite number of terms $\hat{\boldsymbol{n}}_0, \ldots, \hat{\boldsymbol{n}}_K$, for some K, every term of $\{\hat{\boldsymbol{n}}_k\}$ will be orthogonal to* $\operatorname{Span}(\mathcal{X})$*, providing that*

$$\phi_0 > \cos^{-1}\left(\frac{c_d - \epsilon_{\boldsymbol{\mathcal{X}}} - 2\,\gamma\,\epsilon_{\boldsymbol{\mathcal{O}}}}{c_d + \epsilon_{\boldsymbol{\mathcal{X}}}}\right) =: \phi_0^*. \qquad (26)$$

First note that if (24) is true, then the expression of (26) always defines an angle between 0 and $\pi/2$. Second, Theorem 4 can be interpreted using the same asymptotic arguments as Theorem 2. In particular, notice that the lower



(a) $M = 64$

(b) $M = 32$

(c) $M = 128$

Figure 4. ROC curves for different number of outliers. Inliers are face images of a single individual and outliers are images chosen randomly among different object categories.

bound on the angle ϕ_0 tends to zero as M, N go to infinity with γ constant. Note also that this result does not show convergence of the sequence $\hat{\boldsymbol{n}}_k$: it only shows that this sequence will eventually satisfy the desired property of being orthogonal to the space of inliers; a convergence result remains yet to be established.

5. Dual Principal Component Pursuit

So far we have established a mechanism of obtaining an element $\boldsymbol{b}_1$ of $\mathcal{V}^\perp$, where $\mathcal{V} = \operatorname{Span}(\boldsymbol{\mathcal{X}})$: run the sequence of linear programs (3) until the function $\left\|\tilde{\boldsymbol{\mathcal{X}}}^\top \hat{\boldsymbol{b}}_k\right\|_1$ converges within some small ϵ; then assuming no pathological point set distributions, any vector $\hat{\boldsymbol{n}}_k$ can be taken as $\boldsymbol{b}_1$. There are two possibilities: either $\mathcal{V}$ is a hyperplane of dimension $D-1$ or $\dim \mathcal{V} < D-1$. In the first case, $\boldsymbol{b}_1$ is the unique up to scale element of $\mathcal{V}^\perp$, which proves that in this case the sequence of (3) in fact converges. In such a case, we can identify our subspace model with the hyperplane defined by the normal $\boldsymbol{b}_1$. Next, if $\dim \mathcal{V} < D-1$, we can proceed to find a second element $\boldsymbol{b}_2$ of $\mathcal{V}^\perp$ that is orthogonal to $\boldsymbol{b}_1$ and so on. This naturally leads to the *Dual Principal Component Pursuit* shown in Algorithm 1.

A few comments are in order. In Algorithm 1, c is an estimate for the codimension $D-d$ of the inlier subspace $\operatorname{Span}(\boldsymbol{\mathcal{X}})$. If c is rather large, then in the computation of each $\boldsymbol{b}_i$, it is more efficient to reduce the coordinate representation of the data by replacing $\tilde{\boldsymbol{\mathcal{X}}}$ with $\pi_i(\tilde{\boldsymbol{\mathcal{X}}})$, where $\pi_i : \mathbb{R}^D \to \mathbb{R}^{D-(i-1)}, i \geq 2$, is the orthogonal projection onto $\operatorname{Span}(\boldsymbol{b}_1, \dots, \boldsymbol{b}_{i-1})^\perp$, and solve the linear program in step 8 in the projected space.

Algorithm 1 Dual Principal Component Pursuit

```
1:  procedure DPCP($\tilde{\boldsymbol{\mathcal{X}}}, c, \epsilon, T_{\max}$)
2:      $\mathcal{B} \leftarrow \emptyset$;
3:      for $i = 1 : c$ do
4:          $k \leftarrow 0; \Delta\mathcal{J}_0 \leftarrow \infty$;
5:          $\boldsymbol{n}_0 \leftarrow \operatorname{argmin}_{\hat{\boldsymbol{n}}:\|\hat{\boldsymbol{n}}\|_2=1,\, \hat{\boldsymbol{n}}\perp\boldsymbol{b}_1,\dots,\boldsymbol{b}_{i-1}} \left\|\tilde{\boldsymbol{\mathcal{X}}}^\top \hat{\boldsymbol{n}}\right\|_2$;
6:          while $k \leq T_{\max}$ and $\Delta\mathcal{J}_0 > \epsilon$ do
7:              $k \leftarrow k+1$;
8:              $\boldsymbol{n}_k \leftarrow \operatorname{argmin}_{\boldsymbol{n}:\boldsymbol{n}^\top\hat{\boldsymbol{n}}_{k-1}=1,\boldsymbol{n}\perp\mathcal{B}} \left\|\tilde{\boldsymbol{\mathcal{X}}}^\top \boldsymbol{n}\right\|_1$;
9:              $\Delta\mathcal{J}_k \leftarrow \left\|\tilde{\boldsymbol{\mathcal{X}}}^\top \hat{\boldsymbol{n}}_{k-1}\right\|_1 - \left\|\tilde{\boldsymbol{\mathcal{X}}}^\top \hat{\boldsymbol{n}}_k\right\|_1$;
10:         end while
11:         $\boldsymbol{b}_i \leftarrow \hat{\boldsymbol{n}}_k$;
12:         $\mathcal{B} \leftarrow \mathcal{B} \cup \{\boldsymbol{b}_i\}$;
13:     end for
14:     return $\mathcal{B}$;
15: end procedure
```

Notice further how the algorithm initializes $\boldsymbol{n}_0$: This is effectively the right singular vector of $\pi_i(\tilde{\boldsymbol{\mathcal{X}}})^\top$, that corresponds to the smallest singular value. As it will be demonstrated in Section 6, this choice has the effect that the angle of $\boldsymbol{n}_0$ from the inlier subspace is typically large, in particular, larger than the smallest initial angle (26) required for the success of the principal component pursuit of (3).

6. Experiments

In this section we investigate experimentally the proposed DPCP Alg. 1. Using both synthetic (subsection 6.1) and real data (subsection 6.2), we compare DPCP to the three methods SE, L21 and RANSAC discussed in Section 3.1 as well as to the method of eq. (11) discussed in Section 3.2, which we will refer to as SVS (*Sparsest Vector in a Subspace*). The parameters of the methods are set to fixed values, chosen such that the methods work well across all tested dimension and outlier configurations. In particular, we use $\alpha_{SE} = 100, \tau_{L21} = 100$ and $\lambda_{L21} = 3/(7\sqrt{M})$; see [19] and [23] for details. Regarding DPCP, we fix $T_{max} = 10, \epsilon = 10^{-6}$, and unless otherwise noted, we set c equal to the true codimension of the subspace.

6.1. Synthetic Data

To begin with, we evaluate the performance of DPCP in the absence of noise, for various subspace dimensions $d = 1 : 1 : 29$ and outlier percentages $R := M/(M+N) = 0.1 : 0.1 : 0.9$. We fix the ambient dimension $D = 30$, sample $N = 200$ inliers uniformly at random from $\mathcal{V} \cap \mathbb{S}^{D-1}$

and M outliers uniformly at random from $\mathbb{S}^{D-1}$. We are interested in examining the ability of DPCP to recover a single normal vector ($c = 1$) to the subspace, by means of recursion (3). The results are shown in Fig. 1 for 10 independent trials. Fig. 1(a) shows whether the theoretical conditions of (24) are satisfied or not. In checking these conditions, we estimate the abstract quantities $\epsilon_{\boldsymbol{\mathcal{O}}}, \epsilon_{\boldsymbol{\mathcal{X}}}, \mathcal{R}_{\boldsymbol{\mathcal{O}},K_1}, \mathcal{R}_{\boldsymbol{\mathcal{X}},K_2}$ by Monte-Carlo simulation. Whenever these conditions are satisfied, we choose $\boldsymbol{b}_0$ in a controlled fashion, so that its angle ϕ_0 from the subspace is larger than the minimal angle ϕ_0^* of (26), and then we run DPCP; if the conditions are not true, we do not run DPCP and report a 0. Fig 1(b) shows the angle of $\boldsymbol{b}_{10}$ from the subspace. We see that whenever (24) is true, DPCP returns a normal after only 10 iterations. Fig 1(c) shows that if we initialize $\boldsymbol{b}_0$ randomly, then its angle ϕ_0 from the subspace becomes less than the minimal angle ϕ_0^*, as d increases. Even so, Fig. 1(d) shows that DPCP still yields a numerical normal, except for the regime where both d and R are very high. Notice that this is roughly the regime where we have no theoretical guarantees in Fig. 1(a). Fig. 1(e) shows that if we initialize $\boldsymbol{b}_0$ as the right singular vector of $\tilde{\boldsymbol{\mathcal{X}}}^\top$ corresponding to the smallest singular value, then $\phi_0 > \phi_0^*$ is true for most cases, and the corresponding performance of DPCP in Fig. 1(f) improves further. Finally, Fig. 1(g) plots ϕ_0^*. We see that for very low d this angle is almost zero, i.e. DPCP does not depend on the initialization, even for large R. As d increases though, so does ϕ_0^*, and in the extreme case of the upper rightmost regime, where d and R are very high, ϕ_0^* is close to 90^o, indicating that DPCP will succeed only if $\boldsymbol{b}_0$ is very close to $\mathcal{V}^\perp$.

Next, for the same range of R and d, and still in the absence of noise, we examine the potential of each of SE, L21, SVS, RANSAC and DPCP to perfectly distinguish outliers from inliers. Note that each of these methods returns a *signal* $\boldsymbol{\alpha} \in \mathbb{R}_+^{N+M}$, which can be thresholded for the purpose of declaring outliers and inliers. For SE, $\boldsymbol{\alpha}$ is the ℓ_1-norm of the columns of the coefficient matrix $\boldsymbol{C}$, while for L21 it is the ℓ_2-norm of the columns of $\boldsymbol{E}$. Since RANSAC, SVS and DPCP directly return subspace models, for these methods $\boldsymbol{\alpha}$ is simply the distances of all points to the estimated subspace model. In Fig. 2 we depict success versus failure, where success is interpreted as the existence of a threshold on $\boldsymbol{\alpha}$ that perfectly separates outliers and inliers. As expected, the low-rank methods SE and L21 can not cope with large dimensions even in the presence of $10 - 20\%$ outliers. As expected, RANSAC is very successful irrespectively of dimension, when R is small, since the probability of sampling outlier-free subsets is high. But as soon as R increases, its performance drops dramatically. Moving on, SVS is the worst performing method, which we attribute to its approximate nature. Remarkably, DPCP performs perfectly irrespectively of dimension for up to 50% outliers. Note that we use the true codimension c of the subspace as input to DPCP; this is to ascertain the true limits of the method. Certainly, in practice only an estimate for c can be used. As we have observed from experiments, the performance of DPCP typically does not change much if the codimension is underestimated; however performance can deteriorate significantly if the true c is overestimated. Moreover, we note that while SE, L21 and SVS are extremely fast, as they rely on ADMM implementations, DPCP is much slower, even if we use an optimizer such as Gurobi [10]. Speeding up DPCP is the subject of current research.

Finally, in Fig. 3 we show ROC curves associated with the thresholding of $\boldsymbol{\alpha}$ for varying levels of noise and outliers. When d is small, Fig. 3(a) shows that SE, L21 and DPCP are equally robust giving perfect separation between outliers and inliers, while SVS and RANSAC perform poorly. Interestingly, for large d (Fig. 3(a)), DPCP gives considerably less False Positives (FP) than all other methods across all cases, indicating once again its unique property of being able to handle large subspace dimensions in the presence of many outliers.

6.2. Real Data

In this subsection we consider an outlier detection scenario in PCA using real images. The inliers are taken to be all $N = 64$ face images of a single individual from the Extended Yale B dataset [7], while the M outliers are randomly chosen from Caltech101 [6]. All images are cropped to size 48×42 as was done in [5]. For a fair comparison, we run SE on the raw 2016-dimensional data, while all other methods use projected data onto dimension $D = 50$. Since it is known that face images of a single individual under different lighting conditions lie close to an approximately 9-dimensional subspace [5], we choose the codimension parameter of DPCA to be $c = 41$. We perform 10 independent trials for each individual across all 38 individuals for a different number of outliers $M = 32, 64, 128$ and report the ensemble ROC curves in Fig. 4. As is evident, DPCA is the most robust among all methods.

7. Conclusions

We presented *Dual Principal Component Pursuit (DPCP)*, a novel ℓ_1 outlier detection method, which is based on solving an ℓ_1 problem on the sphere by linear programs over a sequence of tangent spaces on the sphere. DPCP is able to handle subspaces of as low codimension as 1 in the presence of as many outliers as 50%. Future research will be concerned with speeding up the method as well as extending it to multiple subspaces and other types of data corruptions, such as missing entries and entry-wise errors.

Acknowledgement

This work was supported by grant NSF 1447822.

References

[1] A.M. Bruckstein, D.L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51(1):34–81, Feb. 2009.

[2] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3), 2011.

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

[4] E. Elhamifar and R. Vidal. Robust classification using structured sparse representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

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

[6] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.*, 106(1):59–70, April 2007.

[7] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23(6):643–660, 2001.

[8] P. J. Grabner, B. Klinger, and R.F. Tichy. Discrepancies of point sequences on the sphere and numerical integration. *Mathematical Research*, 101:95–112, 1997.

[9] P. J. Grabner and R.F. Tichy. Spherical designs, discrepancy and numerical integration. *Math. Comp.*, 60(201):327–336, 1993.

[10] Inc. Gurobi Optimization. Gurobi optimizer reference manual, 2015.

[11] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 1933.

[12] I. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 2nd edition, 2002.

[13] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 61, 2009.

[14] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *International Conference on Machine Learning*, 2010.

[15] S. Nam, M.E. Davies, M. Elad, and R. Gribonval. The cosparse analysis model and algorithms. *Applied and Computational Harmonic Analysis*, 34(1):30–56, 2013.

[16] K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh and Dublin Philosphical Magazine and Journal of Science*, 2:559–572, 1901.

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

[18] Q. Qu, J. Sun, and J. Wright. Finding a sparse vector in a subspace: Linear sparsity using alternating directions. *CoRR*, abs/1412.4659, 2014.

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

[20] D.A. Spielman, H. Wang, and J. Wright. Exact recovery of sparsely-used dictionaries. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 3087–3090. AAAI Press, 2013.

[21] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere. *CoRR*, abs/1504.06785, 2015.

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

[23] H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via Outlier Pursuit. In *Neural Information Processing Systems*, 2010.

[24] M. Zibulevsky, B. Pearlmutter, et al. Blind source separation by sparse decomposition in a signal dictionary. *Neural computation*, 13(4):863–882, 2001.