Robust subspace recovery by geodesically convex optimization
Teng Zhang

Abstract—We introduce Tyler’s M-estimator to robustly recover the underlying linear model from a data set contaminated by outliers. We prove that the objective function of this estimator is geodesically convex on the manifold of all positive definite matrices and have a unique minimizer. Besides, we prove that when inliers (i.e., points that are not outliers) are sampled from a subspace and the percentage of outliers is bounded by some number, then under some very weak assumptions a commonly used algorithm of this estimator can recover the underlying subspace exactly. We also show that empirically this algorithm compares favorably with other convex algorithms of subspace recovery.

I. INTRODUCTION

This paper is about the following problem: suppose we are given a data set \( \mathcal{X} \) with inliers sampled from a low-dimensional linear model and some arbitrary outliers, can we recover the underlying linear model? The primary tool for this problem is Principal Component Analysis (PCA). However, PCA is very sensitive to outliers. Considering the popularity of linear modeling, an robust algorithm that find the underlying linear model will have many applications.

This work introduces Tyler’s M-estimator of covariance in [28], and proves that the objective function is geodesically convex on the manifold of all positive definite matrices. Moreover, this work proves that when inliers are sample from a subspace \( L_* \), a commonly used algorithm for this estimator finds the underlying subspace \( L_* \) under very weak conditions that almost only depend on the percentage of outliers.

A. Notation and conventions

We assume that we are given a data set \( \mathcal{X} \subset \mathbb{R}^D \) with \( N \) points. We define the projector \( \Pi_L \) as the \( D \times D \) symmetric matrix such that \( \Pi^2_L = \Pi_L \), and the range of \( \Pi_L \) is \( L \). We define \( P_L \) as the \( D \times \dim(L) \) projection matrix from \( \mathbb{R}^D \) to \( L \), or equivalently, any \( P_L \) such that \( \Pi_L = P_L P_L^T \). We use \( L^\perp \) to denote the orthogonal complement of \( L \).

We use \( \mathcal{X} \cap L \) to express the set of points that lie both in \( \mathcal{X} \) and the subspace \( L \), and use \( \mathcal{X} \setminus L \) to express the set of points that lie in \( \mathcal{X} \) but not in the subspace \( L \). We use \( |\mathcal{X}| \) to denote the cardinality of the set \( \mathcal{X} \), \( S_+(D) \) to denote the set of all \( D \times D \) semi-positive definite matrices and \( S^{++}(D) \) to denote the set of all \( D \times D \) positive definite matrices.

B. Main results

In this paper we introduce the following estimator due to Tyler [28]:

\[
\Sigma_* = \arg \min_{\Sigma: \text{tr}(\Sigma) = 1, \Sigma = \Sigma^T} F(\Sigma), \quad \text{where} \quad (I.1)
\]

where

\[
F(\Sigma) = \frac{1}{N} \sum_{x \in \mathcal{X}} \log(x^T \Sigma^{-1} x) + \frac{1}{D} \log \det(\Sigma),
\]

and we obtain \( \Sigma_* \) by the limit of the sequence \( \Sigma^{(k)} \) generated by the following iterative procedure in [28]:

\[
\Sigma^{(k+1)} = \sum_{x \in \mathcal{X}} \frac{x x^T}{x^T \Sigma^{(k)} x - 1} / \text{tr}(\sum_{x \in \mathcal{X}} \frac{x x^T}{x^T \Sigma^{(k)} x - 1}). \quad (I.2)
\]

We will explain the motivation of this estimator as an M-estimator of covariance in Section I-D and show in Section III that the objective function \( F(\Sigma) \) is geodesically convex on \( S^{++}(D) \), and under the condition (III.1) the sequence \( \Sigma^{(k)} \) generated by (I.2) converges to the unique solution of (I.1).

When the inliers lie exactly on the subspace \( L_* \), then under some weak assumptions (almost only depends on the percentage of outliers) we can recover the \( L_* \) exactly from \( \lim_{k \to \infty} \Sigma^{(k)} \), which is a singular matrix with \( L_* \) as its range.

Theorem I.1. If there exists a \( d \)-dimensional subspace \( L_* \) such that

\[
\frac{|\mathcal{X} \cap L_*|}{|\mathcal{X}|} > \frac{d}{D}, \quad (I.3)
\]

and the points in the set \( \mathcal{Y}_1 = \{P_L x : x \in \mathcal{X} \cap L_* \} \subset \mathbb{R}^d \) and \( \mathcal{Y}_0 = \{P_L x : x \in \mathcal{X} \setminus L_* \} \subset \mathbb{R}^{D-d} \) lie in general positions respectively (i.e., any \( k \) points in \( \mathcal{Y}_1 \) span a \( k \)-dimensional subspace for all \( k \leq d \) and any \( k \) points in \( \mathcal{Y}_0 \) span a \( k \)-dimensional subspace for all \( k \leq D - d \)). Then the sequence \( \Sigma^{(k)} \) generated by (I.2) converges to some \( \Sigma \) such that \( \text{im}(\Sigma) = L_* \).

The condition of “general positions” is very weak—for example, when we choose inliers arbitrarily from a uniform distribution in a ball in \( L_* \), or from Gaussian measure in \( L_* \) and choose outliers arbitrarily from a uniform measure in a ball in \( \mathbb{R}^D \), or from Gaussian measure in \( \mathbb{R}^D \) this condition holds with probability 1.

We remark that when the ambient dimension \( D \to \infty \) and the dimension of subspace \( L_* \) is kept as the constant \( d \), then \( \frac{d}{D} \) approaches 0 and the required percentage of inliers in Theorem I.1 approaches 0. This property makes Theorem I.1 particularly strong for high-dimensional data set with a low-dimensional structure.
C. Previous work

The robust estimator for covariance has been well studied in the statistical literature, which is related topic to robust linear modeling since we can recover the linear model by the principal components of the estimated covariance. The M-estimators, L-estimators, MCD/MVE estimators and Stahel-Donoho estimator have been proposed; for a complete review we refer the reader to [20, Section 6]. However most of these methods are not convex (or the convexity is not analyzed), so the algorithms are intractable (or have unknown tractability). It is possible that the only exception is M-estimators: the convergence of its algorithm has been analyzed in [13], and we will show later that as a special M-estimator, $F(\Sigma)$ is geodesically convex in the space of positive definite matrices (it is also shown in [32]).

There are other methods that recover the linear model without estimating the robust covariance, such as Projection Pursuit [19, 1, 22, Section 2], which find the principal component directly by optimization on a sphere. Another common strategy is to fit the linear model by PCA after removing possible outliers [27, 33]. However these methods are still nonconvex.

Some recent works on robust linear modeling focus on convex optimization and tractable algorithms [34, 22, 35, 17]. Similar to Theorem [11] these works provide conditions for exact subspace recovery. We remark that these conditions are more complicated than the condition in Theorem [11] since they assume an “incoherence condition” that requires the inliers to be spread out on the subspace $L_s$, which is not required in Theorem [11]. These kind of conditions are required in [34] Theorem 1], [35] (6)(7)]. In [17] Theorem 1.1 it is shown that the exact recovery holds with high probability when
\[
\frac{|X \cap L_s|}{d} > C_0 + C_1 \frac{|X \setminus L_s|}{D},
\]
which is a simple condition and very similar to the condition in [35]. However this condition is obtained under the assumption that inliers and outliers are both sampled from Gaussian distributions. In another recent work, Soltanolkotabi and Candès proved that sparse subspace clustering (SSC) algorithm [17] can recover multiple subspaces with high probability, but this theory also have probabilistic assumptions: it assumes that inliers and outliers are both i.i.d sampled from uniform measures on unit spheres [26, Theorem 1.3].

We remark that our condition (3.3) sometimes can be more restrictive than the corresponding conditions of other convex methods. For example, when outliers have small magnitude and concentrate around the origin, then the conditions in [35, Theorem 2] can tolerate more outliers. Similarly, the conditions in [34] Theorem 1] and [26 Theorem 1.3] can also tolerate more outliers than (3.5) under some settings. The advantage of our condition is that it is deterministic and simple, and empirically it is also usually less restrictive than the conditions in [35], [22, 35, 17].

D. M-estimators

In this section we show that the estimator [11] can be considered as a special M-estimator of covariance and gives background of the current research on this estimator. We start with the motivation: it is well known that the empirical covariance is the MLE estimator for the covariance when we assume that all $x \in \mathcal{X}$ are i.i.d. drawn from a Gaussian distribution. As a natural generalization, M-estimators of covariance [19], [10], [21] consider the generalized distribution
\[
C(\rho) e^{-\rho(x^T \Sigma^{-1} x)} \sqrt{\text{det}(\Sigma)},
\]
where $C(\rho)$ is a normalization constant, chosen so that the integral of the distribution is equal to one. It is a generalization since when $\rho(x) = x$, (1.4) gives Gaussian distribution. When data points are i.i.d. sampled from the distribution (1.4), the corresponding MLE estimator of covariance is called M-estimator, which minimizes
\[
\frac{1}{N} \sum_{x \in \mathcal{X}} \rho(x^T \Sigma^{-1} x) + \frac{1}{2} \log \text{det}(\Sigma).
\]

The objective function $F(\Sigma)$ can be considered as the M-estimator when $\rho(x) = \frac{1}{2} \log(1 + \frac{1}{\sigma} x^T \Sigma^{-1} x)$. While for this choice of $\rho$ the function in (1.4) is not a distribution, it can be considered as the limit of the following multivariate student distribution as $\nu \to 0$ [20, page 187]:
\[
\Gamma(\nu/2) / \sqrt{\nu} \pi^{D/2} / \sqrt{\det(\Sigma)} \sum_{x \in \mathcal{X}} [1 + \frac{1}{\nu} x^T \Sigma^{-1} x]^{(\nu+D)/2}.
\]

Since student distribution has a heavy tail, it is expected that this estimator should be more robust to outliers.

The reason that we enforce the condition $\text{tr}(\Sigma) = 1$ in (1.1) is the scale invariance property of $F(\Sigma)$: for any constant $c > 0$ and $\Sigma \in S_{++}(D)$, we have
\[
F(c \Sigma) = F(\Sigma).
\]

This simple fact can be easily verified, but will be repetitively used in the analysis later.

Tyler and Kent investigated the estimator (1.1) implicitly, by solving the equation $F'(\Sigma) = 0$ [28, 12]. They obtained the uniqueness of the solution (up to scaling) to $F'(\Sigma) = 0$ and showed that the algorithm (1.2) converges to this solution in [12, Theorem 2], under the assumption (III.1). This result is almost equivalent Theorem (III.1) and Theorem (III.4) in this work, except that we consider the minimization of $F(\Sigma)$ directly and also show the existence of the minimizer. An interesting claim in [28] is that, this estimator is the “most robust” estimator of the scatter matrix of an elliptical distribution in the sense of minimizing the maximum asymptotic variance.

The geodesical convexity of the objective function $F(\Sigma)$ was discovered later. In [2], Auderset et al. showed that the function is geodesically convex on the space $\{\Sigma \in S_{++}(D) : \text{det}(\Sigma) = 1\}$. After finishing this work, we learned that the geodesical convexity of $F(\Sigma)$ on the space $S_{++}(D)$ was recently independently investigated by Wiesel in [32, Proposition 1]. Wiesel also extended the convex analysis to regularized Tyler’s M-estimator in [32] and generalized LSE (logarithm of a sum of exponents) functions and the estimation of Kronecker structured covariance in [30, 31].
E. Contributions and the structure of this paper

The main contribution in this work is that, we introduce Tyler’s M-estimator for subspace recovery, and justify this estimator by showing that the algorithm (12) can recover the underlying subspace exactly under rather weak assumptions on the distribution of data points. Besides, we also apply geodesic convexity and majorization-minimization argument to show the existence and the uniqueness of the minimizer, and the convergence of the algorithm. While these two facts are also observed in [22], the analysis in this paper is more careful and therefore it proves the uniqueness of the minimizer and the pointwise convergence of the algorithm.

The paper is organized as follows. In Section II we introduce the background on the geometry of $S_{++}(D)$ and geodesic convexity. With this background we prove the uniqueness of the solution to (1.1) and the convergence of the algorithm (12) in Section III. Then we perform some experiments that describe the performance of Algorithm (12) and verify Theorem (1.1) in Section IV. Technical proofs are shown in the Appendix.

II. Preliminaries

Our analysis relies on basic concepts from the geometry of $S_{++}(D)$ and the geodesic convexity. For this purpose, in Section II-A we present a brief summary of the geometry of $S_{++}(D)$ and in Section II-B we introduce the definition and some properties of geodesic convexity. For more details we refer the reader to [21], [22] on the geometry of $S_{++}(D)$ and the geodesic convexity.

A. Metric and geodesic on $S_{++}(D)$

The metric of $S_{++}(D)$ has been well studied in literature. Indeed, the trace metric in differential geometry [13] pg 326, natural metric in symmetric cone [8, 9, 5], affine-invariant metric [24], and the metric given by Fisher information matrix for Gaussian covariance matrix estimation [25] give the same metric on $S_{++}(D)$. For $\Sigma_1, \Sigma_2 \in S_{++}(D)$, these metrics are defined by:

$$\text{dist}(\Sigma_1, \Sigma_2) = \| \log(\Sigma_1^{-\frac{1}{2}} \Sigma_2 \Sigma_1^{-\frac{1}{2}}) \|_F. \quad (II.1)$$

Based on this metric, the unique geodesic $\gamma_{\Sigma_1; \Sigma_2}(t) \ (0 \leq t \leq 1)$ connecting $\Sigma_1$ and $\Sigma_2$ is given by [4, (6.11)]:

$$\gamma_{\Sigma_1; \Sigma_2}(t) = \Sigma_1^\frac{t}{2} (\Sigma_1^{-\frac{1}{2}} \Sigma_2 \Sigma_1^{-\frac{1}{2}})^t \Sigma_1^{-\frac{1}{2}}. \quad (II.2)$$

It follows that the midpoint of $\Sigma_1$ and $\Sigma_2$ is $\gamma_{\Sigma_1; \Sigma_2}(\frac{1}{2}) = \Sigma_1^\frac{1}{2} (\Sigma_1^{-\frac{1}{2}} \Sigma_2 \Sigma_1^{-\frac{1}{2}})^{\frac{1}{2}} \Sigma_1^{-\frac{1}{2}} \Sigma_2^\frac{1}{2}$. We remark that $\Sigma_1^\frac{1}{2} (\Sigma_1^{-\frac{1}{2}} \Sigma_2 \Sigma_1^{-\frac{1}{2}})^{\frac{1}{2}} \Sigma_2^\frac{1}{2}$ is also called the geometric mean of $\Sigma_1$ and $\Sigma_2$ [4, Section 4.1].

B. Geodesic convexity

Geodesic convexity is a natural generalization of the convexity to Riemannian manifolds [22, Chapter 3.2]. Given a Riemannian manifold $\mathcal{M}$ and a set $\mathcal{A} \subset \mathcal{M}$, we say a function $f : \mathcal{A} \to \mathbb{R}$ is geodesically convex, if every geodesic $\gamma_{xy}$ of $\mathcal{M}$ with endpoints $x, y \in \mathcal{A}$ (i.e., $\gamma_{xy}$ is a function from $[0, 1]$ to $\mathcal{M}$ with $\gamma_{xy}(0) = x$ and $\gamma_{xy}(1) = y$) lies in $\mathcal{A}$, and

$$f(\gamma_{xy}(t)) \leq (1 - t)f(x) + tf(y) \text{ for any } x, y \in \mathcal{A} \text{ and } 0 < t < 1. \quad (II.3)$$

Following the proof of [23] Theorem 1.1.4], for a continuous function, the geodesic midpoint convexity is equivalent to the geodesic convexity:

**Lemma II.1.** Let $f : A \to \mathbb{R}$ be a continuous function. If

$$f(\gamma_{xy}(\frac{1}{2})) \leq \frac{f(x) + f(y)}{2} \text{ for any } x \neq y \in A \quad (II.4)$$

then $f$ is a geodesically convex function.

III. Property of the Objective Function and the Algorithm

In this section we study the properties of the objective function $F(\Sigma)$ and the algorithm in (12). We show that under a very mild assumption, the solution to (1.1) is unique and the sequence $\Sigma^{(k)}$ converges to the solution. We will also discuss Theorem (1.1) the empirical algorithm and some implementation issues in this section.

A. Uniqueness of the solution

We first show the existence and the uniqueness of the solution to (1.1) under a rather weak assumption.

**Theorem III.1.** If for any linear subspace $L$ we have

$$\frac{|\chi \cap L|}{N} < \frac{\dim(L)}{D}, \quad (III.1)$$

then the solution of (1.1) exists and is unique.

Indeed, for real data sets that contain noise, (III.1) is usually satisfied if the dimension is smaller than the number of points: in noisy data set generally any $d$-dimensional linear subspace only contains at most $d$ points.

An important remark is that the condition (III.1) is incompatible with the condition on the percentage of inliers in Theorem (1.1). Indeed, the solution to (1.1) does not exists: one may verify that $F\left(\frac{11}{10}H_{\chi_{\Sigma_1}^{(l)} + \epsilon I} + \frac{9}{10}H_{\chi_{\Sigma_2}^{(l)} + \epsilon I}\right)$ converges to $-\infty$ as $\epsilon \to 0$, while $\frac{11}{10}H_{\chi_{\Sigma_1}^{(l)} + \epsilon I} + \frac{9}{10}H_{\chi_{\Sigma_2}^{(l)} + \epsilon I}$ converges to a singular matrix where $F(\Sigma)$ is undefined.

The proof of Theorem III.1 depends on the following two lemmas, whose proofs will be presented in the appendix. In general, Lemma III.2 guarantees the uniqueness of the solution and Lemma III.3 guarantees the existence of the solution. While Lemma III.2 is also proved to [22] Proposition 1], additionally we show the condition for the equality in Lemma III.2 which implies the uniqueness of the minimizer of (1.1).

**Lemma III.2.** $F(\Sigma)$ is geodesically convex on the manifold $S_{++}(D)$. That is, for any $\Sigma_1$ and $\Sigma_2 \in S_{++}(D)$, we have

$$F(\Sigma_1) + F(\Sigma_2) \geq 2F(\frac{1}{2}(\Sigma_1^{-\frac{1}{2}} \Sigma_2 \Sigma_1^{-\frac{1}{2}})^{\frac{1}{2}} \Sigma_1^{-\frac{1}{2}} \Sigma_2^\frac{1}{2}). \quad (III.2)$$

When $\text{span}\{\chi\} = \mathbb{R}^D$, the equality in (III.2) holds if and only if $\Sigma_1 = c \Sigma_2$.
Lemma III.3. Under the condition III.1, we have
\[ F(\Sigma) \to \infty \text{ as } \lambda_{\min}(\Sigma) \to 0. \] (III.3)

Here \( \lambda_{\min}(\Sigma) \) is the smallest eigenvalue of \( \Sigma \).

Now we are ready to prove Theorem III.1.

Proof: We first prove the uniqueness of the solution to (I.1). If \( \Sigma_1 \neq \Sigma_2 \) are both solutions to (I.1), then apply III.2 and the scale invariance (I.6), we have
\[ F(\Sigma_1) \leq F(\Sigma_2) = F(\Sigma_2), \]
for
\[ \Sigma_3 = \frac{\Sigma_1^{+} (\Sigma_1^{+} - \Sigma_2^{+} \Sigma_1^{+} - \Sigma_1^{+})}{\operatorname{tr}(\Sigma_1^{+} (\Sigma_1^{+} - \Sigma_2^{+} \Sigma_1^{+} - \Sigma_1^{+}) + \Sigma_1^{+})}. \]

Since \( \Sigma_1 \) and \( \Sigma_2 \) are both minimizers to \( F(\Sigma) \), we have
\[ F(\Sigma_3) = F(\Sigma_1) = F(\Sigma_1), \]
by the condition of equality in Lemma III.2 (the assumption \( \text{span}\{X\} = \mathbb{R}^D \) in Lemma III.2 holds; otherwise III.1 does not hold for \( L = \text{span}\{X\} \)), we have \( \Sigma_1 = \epsilon \Sigma_2. \) However we have \( \operatorname{tr}(\Sigma_1) = \operatorname{tr}(\Sigma_2) \), therefore \( \Sigma_1 = \Sigma_2 \), which contradicts our assumption, and we prove the uniqueness of the solution to (I.1).

Now we prove the existence of the solution. First, there exists a sequence \( \{\Sigma_k\}_{k \geq 1} \subset \{\Sigma \in S_+^D(D) : \operatorname{tr}(\Sigma) = 1\} \) such that \( F(\Sigma_k) \) converges to
\[ \inf_{\operatorname{tr}(\Sigma) = 1} \{\Sigma \in S_+^D(D) : F(\Sigma)\}. \]
By compactness there is a converging subsequence of \( \{\Sigma_k\} \), and by Lemma III.3 this subsequence does not converge to a singular matrix and therefore the subsequence converges to some matrix \( \Sigma_0 \in S_+^D(D) \). By continuity of \( F(\Sigma) \) we obtain \( F(\Sigma_0) = \inf_{\operatorname{tr}(\Sigma) = 1} \{\Sigma \in S_+^D(D) : F(\Sigma)\} \) and therefore \( \Sigma_0 \) is a solution to (I.1).

B. Convergence of algorithm

In this section we prove the convergence of the sequence \( (\Sigma^k) \) generated by (I.2) under the assumption III.1, and we will also discuss its connection to Theorem III.1 which is about the convergence of the sequence \( (\Sigma^k) \) under another assumption.

We begin with the motivation of the procedure (I.2). If we set the derivative of \( F(\Sigma) \) with respect to \( \Sigma^{-1} \) to be 0, we have
\[ \frac{d}{d\Sigma^{-1}} F(\Sigma) = \frac{1}{N} \sum_{x \in X} x x^T x^{-1} x - \frac{1}{D} \Sigma = 0 \]
and
\[ \Sigma = \frac{D}{N} \sum_{x \in X} x x^T x^{-1} x. \]

Since we minimize \( F(\Sigma) \) under the assumption \( \operatorname{tr}(\Sigma) = 1 \), we have
\[ \Sigma^k = \frac{D}{N} \sum_{x \in X} x x^T x^{-1} x \quad \sum_{x \in X} x x^T \Sigma^{-1} x = \frac{D}{\text{tr} \left( \sum_{x \in X} x x^T \Sigma^{-1} x \right)} \]
whose RHS is the update formula (I.2). Therefore we have the motivation that the

Theorem III.4 shows that \( \Sigma^k \) converges to the solution to (I.1) under the assumption III.1. Similar to Section II, it uses the majorization-minimization argument. However the analysis here is more complete in the sense that it proves the convergence of the sequence \( \Sigma^k \), while the argument in II only leads to the convergence of the objective function \( F(\Sigma^k) \).

Theorem III.4. When the condition III.1 holds, the sequence \( \Sigma^k \) generated by (I.2) converges to the unique solution to (I.1).

This theorem also implies that the condition \( \frac{\lambda_{\min}(\Sigma)}{\lambda_{\max}(\Sigma)} < \frac{d}{D} \) in Theorem III.1 is almost necessary. Indeed, if \( \frac{\lambda_{\min}(\Sigma)}{\lambda_{\max}(\Sigma)} < \frac{d}{D} \), then the condition III.1 is usually satisfied, and by Theorem III.1 the solution to (I.1) exists (and by definition nonsingular) and by Theorem III.4 \( \Sigma^k \) converges to this nonsingular matrix. Therefore we can not recovery \( L_\ast \) by its range. This also shows a phase transition phenomenon on \( \frac{\lambda_{\min}(\Sigma)}{\lambda_{\max}(\Sigma)} = \frac{d}{D} \).

For simplicity in the proof we define the operator \( T : S_+^D \to S_+^D \) as
\[ T(\Sigma) = \sum_{x \in X} x x^T x^{-1} x / \text{tr} \left( \sum_{x \in X} x x^T x^{-1} x \right). \] (III.4)

The main ingredient for the proof is the observation that \( \Sigma^{k+1} = T(\Sigma^k) \) can be considered as a minimizer of a majorization function \( G(\Sigma, \Sigma^k) \) over \( F(\Sigma) \) such that \( G(\Sigma, \Sigma^k) \geq F(\Sigma) \) and \( G(\Sigma_0, \Sigma^k) = F(\Sigma^k) \). In this sense it can be considered as an algorithm with the majorization-minimization (MM) principle III.1. We remark that similar observations are also used in the proof of the convergence in other iteratively reweighted least square (IRLS) algorithms such as [14], [6], [55], [17].

When the condition in Theorem III.1 holds, the assumption III.1 is violated, and by our analysis in Section III.4 the solution to (I.1) does not exist. However, Theorem III.1 shows that the sequence \( \Sigma^k \) still converges and it converges to a singular matrix. Due to the complexity we put its proof in Appendix B.

C. Empirical algorithm and implementation issues

Since the solution to (I.1) can be considered a robust estimator of covariance, empirically we can simply recover the underlying \( d \)-dimensional subspace by the span of its top \( d \) eigenvectors. Our empirical algorithm in summarized in Algorithm I.1.

In each iteration the major computational cost is due to the calculation of the inverse of \( \Sigma^k \) and the calculation of \( \sum_{x \in X} x x^T \Sigma^{k-1} x \), therefore the cost is in the order of \( O(N D^2) \) when \( N \geq D \). We will show later in Section IV-C that the algorithm exhibit linear convergence. In implementation we stop the algorithm after \( k \)-th iteration when
\[ \frac{\| \Sigma^k - \Sigma^{k-1} \|_F}{\| \Sigma^k \|_F} < 10^{-8}. \]

In this paragraph we describe an empirical problem where the algorithm breaks down at some iteration step, and describe a way to overcome it. If the condition in Theorem III.1 holds, \( \lambda_{\min}(\Sigma^k) \) converges to 0 as \( k \to \infty \) and it is nonzero for each \( k \). However in implementation, due to the rounding error,
In some experiments we also add a Gaussian noise to each of the point. We use uniform distribution in outliers, to show that our algorithm allows anisotropic outliers.

\[ \text{outliers i.i.d. from the uniform distribution in the cube} \]

\[ \text{Gaussian distribution} \]

Data generated from the following model. We choose a

\[ \text{Algorithm 1} \]

Empirical algorithm for recovering a \( d \)-dimensional subspace

**Input:** \( \mathcal{X} \in \mathbb{R}^D \): data set, \( d \): dimension of the subspace.

**Output:** \( L_\star \): a \( d \)-dimensional linear subspace.

**Steps:**
1. Initialization: \( \Sigma^{(0)} = I \).
2. Repeat (1)–(2) until convergence:
   - \( k = k + 1 \),
   - \( \Sigma^{(k+1)} = \sum_{x \in \mathcal{X}} \frac{x x^T}{\text{tr}(\sum_{x \in \mathcal{X}} x x^T \Sigma^{(k)} - x x^T)} \).
3. Let \( \Sigma_* \) be the limit of the sequence \( \Sigma^{(k)} \), and let \( L_* \) be the span of top \( d \) eigenvectors of \( \Sigma_* \).

When \( k \) is very large and \( \lambda_{\min}(\Sigma^{(k)}) \) is very close to zero, the calculated \( \Sigma^{(k)} \) could be a non-positive matrix or a matrix with imaginary part and the convergence of Algorithm 1 fails. Therefore in implementation we check the value of \( \min_{x \in \mathcal{X}} x^T \Sigma^{(k)} - x x^T \) in each iteration, and stop the algorithm when it is negative or has imaginary part. We remark that this breakdown will not happen for real data sets or synthetic data sets with noise, since in these cases \( \Sigma^{(k)} \) converges to a nonsingular positive matrix and the rounding error will not make \( \Sigma^{(k)} \) a non-positive matrix or a matrix with imaginary part.

**D. Discussion on spherical projection**

A simple and powerful method to enhance the robustness of an algorithm to outliers is to preprocess the data set by projecting the data points to a unit sphere. Empirically it enhances the robustness of PCA and Reaper algorithms significantly (Section 5). Therefore a natural question is that whether it can also be applied in our algorithm.

Interestingly, spherical projection has been implicitly applied in the objective function \( F(\Sigma) \) and our algorithm: one may verify that the magnitude of any point in \( \mathcal{X} \) does not impact the solution to Figure 1, or the update formula Figure 2.

**IV. Numerical Experiments**

In this section, we present some numerical experiments on Algorithm 1 to obtain the empirical performance of this algorithm. We also show that our algorithm outperforms other convex algorithms of robust PCA by a real data set.

**A. Model for simulation**

In Sections IV-B–IV-D we apply our algorithm on the data generated from the following model. We choose a \( d \)-dimensional subspace \( L_\star \), sample \( N_1 \) points i.i.d. from the Gaussian distribution \( N(0, \Pi_{L_\star}) \) on \( L_\star \), and sample \( N_0 \) outliers i.i.d. from the uniform distribution in the cube \( [0, 1]^D \).

In some experiments we also add a Gaussian noise \( N(0, \epsilon^2 I) \) to each of the point. We use uniform distribution in \( [0, 1]^D \) for outliers, to show that our algorithm allows anisotropic outliers.

In this section we use the model in Section IV-A choose \( D = 10 \) or 50, \( d = 5 \), \( N_0 = 100 \) and different values of \( N_1 \) (2 to 20 for \( D = 50 \) and 80 to 120 for \( D = 10 \)). The mean recovery error \( \| \Pi_L - \Pi_{L_\star} \|_F \) over 20 experiments is recorded in Figure 1 where \( \hat{L} \) is obtained by the Algorithm 1 and \( L_\star \) is the true underlying subspace. Theorem II guarantees that \( \| \Pi_L - \Pi_{L_\star} \|_F \) for \( N_1 > 100 \) when \( D = 10 \) and \( N_1 > 10 \) when \( D = 50 \), and this is verified in this experiment. When \( D = 50 \) and \( N_1 = 11 \) there is a small nonzero recovery error, which seems to contradicts Theorem II. But we remark that when \( D = 50 \) and \( N_1 = 11 \) the convergence is slow, and we stop the algorithm at the 1000-th iteration without really converging to the solution to Figure 1. We expect that exact recovery of the subspace \( L_\star \) could be obtained after larger number of iterations in Algorithm 1.
In this section we show that empirically the algorithm converges linearly. In the left figure in Figure 2 we show the convergence rate for simulated data sets with $D = 10$, $d = 5$, $N_0 = 100$ and $N_1 = 80, 100, 120$. Additional we add a Gaussian noise with $\varepsilon = 0.01$. The $x$-axis represent the number of iterations $k$ and the $y$-axis is $\|\Sigma^{(k)} - \Sigma^{(K)}\|_F$, where $K$ is the total number of iterations in Algorithm 1. In the right figure we show a different convergent rate: for two different simulations with no noise we plot $\|\Pi_{L_k} - \Pi_{L_*}\|_F$ with respect to the number of iterations $k$, where $L_k$ is the span of first $d$ eigenvectors of $\Sigma^{(k)}$. We use the settings $(N_1, N_0, D, d) = (120, 100, 10, 5)$ and $(20, 100, 50, 5)$ since by Theorem 2 we have $\lim_{k \to \infty} \|\Pi_{L_k} - \Pi_{L_*}\|_F = 0$. From the right figure in Figure 2 we see that the recovery error also converges linearly.

**D. Robustness to noise**

In this section we investigate the empirically robustness of our algorithm to noise by simulated data set sampled according to section 4.1 with $(N_1, N_0, D, d) = (120, 100, 10, 5)$, and different size of noise $\varepsilon$. We use this setting since when $\varepsilon = 0$, we recover the subspace exactly. We record the recovery error in Figure 3 with respect to the size of noise. In this experiment the recovery error depends linearly on the size of noise $\varepsilon$. Indeed, we consider a theory that explain the performance of Algorithm 1 under some noise of small size as an interesting future question.

**E. Faces in a Crowd**

In this section we test our algorithm on the experiment of “Faces in a Crowd” in [17, Section 5.4].

The goal of this experiment is to show that our algorithm can be used to robustly learn the structure of face images. Linear modeling is applicable here since the images of the faces from the same person lies on a 9-dimensional subspace [3]. In this experiment we learn the subspace from a data set that contains 32 face images of a person from the Extended Yale Face Database [16] and 400 random images from the BACKGROUND/Google folder of the Caltech101 database [9]. The images are converts to grayscale and downsample to 20x20 pixels. We preprocess the images by subtracting their Euclidean median, apply Algorithm 1 to this data set to obtain a 9-dimensional subspace, and we use 32 other images from the same person to test how the learned subspace fits these images.

This experiment is also used in [17, Section 5.4], therefore we only compare our algorithm with S-Reaper, which has been shown to perform better than PCA, spherical PCA, LLD and Reaper algorithms. PCA algorithm is still included for comparison since it is the basic technique in linear modeling. Figure 5 shows five images and their projections to the 9-dimensional subspace fitted by PCA, S-reaper and our algorithm (which is labeled as “M-estimator” due to the argument in Section 1-D) respectively. Figure 5 shows that our algorithm visually performs better than S-Reaper, especially...
ordered distances of the 32 test images to the fitted subspace. This observation can also be quantitatively verified by checking the distances of 32 test images to the fitted subspace by PCA, S-reaper and our algorithm, which is shown in Figure IV-E. The subspace generated by our algorithm has smaller distances to the test images, which explain the better performance of our algorithm in Figure IV-E.

For the test images. This observation can also be quantitatively verified by checking the distances of 32 test images to the fitted subspace by PCA, S-reaper and our algorithm, which is shown in Figure IV-E. The subspace generated by our algorithm has smaller distances to the test images, which explain the better performance of our algorithm in Figure IV-E.

Besides, in this experiment our algorithm performs much faster than S-Reaper; our algorithm costs 4.4 seconds on a machine with Intel Core 2 Duo CPU at 3.00GHz and 6GB memory, while S-reaper cost 40 seconds. It is expected since there is an additional eigenvalue decomposition in each iteration of the S-Reaper algorithm.

V. Discussion

In this paper we have investigated an M-estimator for covariance estimation, and proved that this estimator can find the underlying subspace exactly under a rather weak assumption. We also demonstrated the virtue of this method by experiments on simulated data sets and real data sets.

An open question is that, if we can have a theoretical guarantee on the robustness of our algorithm to noise and therefore verify the empirical performance in Section IV-D, we find it difficult to apply the commonly used perturbation analysis in [32] Section 2.7] or [34] Theorem 2], which are based on the size of the perturbation of the objective function, since the objective function \( F(\Sigma) \) at a singular matrix is undefined.

An interesting direction is to extend the idea of geodesical convexity to other problems. Euclidean metric between matrices is usually used and under this metric the set of all positive definite matrices is considered as a cone. However in this work we consider the set of all positive matrices as a manifold and use the Riemannian metric between matrices. It turns out that while \( F(\Sigma) \) in nonconvex in Euclidean metric, it is convex in Riemannian metric, and this formulation is more powerful than similar formulations that are convex in Euclidean metric [35], [17]. It would be interesting if there are other optimization problems with the property of geodesical convexity.

VI. Acknowledgement

The author would like to thank Michael McCoy for reading an earlier version of this manuscript and for helpful comments. The author is grateful to Lek Heng Lim for introducing the book [4] and helpful discussions.

VII. Appendix

A. Proof of Lemma III.2

Proof: Geodesical convexity of \( F(\Sigma) \) follows from (III.2) and Lemma III.1. Therefore we only need to prove (III.2) for geodesic convexity.

We will prove (III.2) by showing that, if \( \Sigma_1 \in S_{++}(D) \) is the geometric mean of \( \Sigma_1, \Sigma_2 \in S_{++}(D) \), then we have

\[
\ln(\det(\Sigma_1)) + \ln(\det(\Sigma_2)) = 2 \ln(\det(\Sigma_3)),
\]

and

\[
\ln(x^T \Sigma_1 x) + \ln(x^T \Sigma_2 x) \geq 2 \ln(x^T \Sigma_3 x).
\]

We start with the proof of (VII.1). Use (II.2) with \( t = \frac{1}{2} \), we have

\[
\Sigma_3 \Sigma_1^{-1} \Sigma_3 = \Sigma_1^{-\frac{1}{2}} (\Sigma_1^{-\frac{1}{2}} \Sigma_3 \Sigma_1^{-\frac{1}{2}})^{\frac{1}{2}} \Sigma_1^{-\frac{1}{2}} \Sigma_1^{-1} \Sigma_2^{-\frac{1}{2}} (\Sigma_2^{-\frac{1}{2}} \Sigma_3 \Sigma_2^{-\frac{1}{2}})^{\frac{1}{2}} \Sigma_2^{-\frac{1}{2}}
\]

(VII.3)

Using (VII.3), (VII.1) can be proved as follows:

\[
\det(\Sigma_3) = \det(\Sigma_3^{\Sigma_1^{-1} \Sigma_3}) = \det(\Sigma_3) \det(\Sigma_1^{-1}) \det(\Sigma_3)
\]

\[
= \det(\Sigma_3)^2 / \det(\Sigma_1).
\]

To prove (VII.2), we let the SVD decomposition of \( \Sigma_1^{-\frac{1}{2}} \Sigma_2 \Sigma_1^{-\frac{1}{2}} = U_0 \Sigma_0 U_0^T \) and define \( \hat{x} = U_0 \Sigma_0^T \hat{x} \), then we have \( x^T \Sigma_1 x = \hat{x}^T \hat{x} \), \( x^T \Sigma_2 x = \hat{x}^T \Sigma_0 \hat{x} \), and \( x^T \Sigma_3 x = \hat{x}^T \hat{x} \).
\(\hat{x}^T \Sigma_0^{-\frac{1}{2}} \hat{x}\). Assuming that \(\Sigma_0\) is a diagonal matrix with diagonal entries \(\sigma_1, \sigma_2, \ldots, \sigma_p\) and \(\hat{x} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_p)^T\), then \(\text{VII.2}\) is equivalent to

\[
\sum_{i=1}^{p} \sigma_i^2 \hat{x}_i^2 \geq \left(\sum_{i=1}^{p} \frac{1}{\sigma_i^2} \hat{x}_i^2\right)^2,
\]

which can be verified by Cauchy-Schwartz inequality. Therefore \(\text{VII.2}\) is proved.

Finally we find the condition such that the equality in \(\text{III.2}\) holds. By its proof of geodesic convexity we know that it holds only when the equality \(\text{VII.2}\) holds for any \(x \in X\).

By the condition of equality in Cauchy-Schwartz inequality, we have that the equality in \(\text{III.2}\) only holds when for any \(1 \leq i \leq D\) (here \(i\) is the index of coordinates) such that \(\hat{x}_i \neq 0, \sigma_i = c\) for some \(c \in \mathbb{R}\). When \(\Sigma_1 \neq c \Sigma_2, \sigma_i\) is not the same number for all \(1 \leq i \leq D\). Therefore there exists \(1 \leq i \leq D\) such that \(\hat{x}_i = 0\). That is, there exists a hyperplane in \(\mathbb{R}^D\) such that \(\hat{x}\) lies on it. Since \(\hat{x}\) is a linear transformation of \(x\), when \(\text{VII.2}\) holds for any \(x \in X\), then there exists a hyperplane such that it contains \(x \in X\), which contradicts our assumption that \(\text{span}\{X\} = \mathbb{R}^D\).

\[\blacksquare\]

\[\text{B. Proof of Theorem \text{III.3}}\]

First we will prove that the operator \(T\) is monotone with respect to the objective function \(F(x) = F(T(\Sigma)) \leq F(\Sigma), \) and the equality holds for \(\Sigma \in S_{++}(D)\) only when \(T(\Sigma) = \Sigma\).

We prove it by constructing the following majorization function over \(F(\Sigma)\):

\[
G(\Sigma, \Sigma^*) = \left\{ \frac{1}{N} \sum_{x \in X} \frac{x^T x}{x^T \Sigma^{-1} x} \right\} + \frac{1}{D} \log \det(\Sigma) + C.
\]

When \(C\) is well chosen such that \(G(\Sigma^*, \Sigma^*) = F(\Sigma^*)\). The fact

\[
G(\Sigma, \Sigma^*) \geq F(\Sigma)
\]

can be proved by checking the first and the second derivative of \(G(\Sigma, \Sigma^*) - F(\Sigma)\) with respect to \(\Sigma^{-1}\).

It is easy to verify the unique minimizer of \(G(\Sigma, \Sigma^*)\) is

\[
\Sigma = \frac{D}{N} \sum_{x \in X} \frac{x^T x}{x^T \Sigma^*^{-1} x},
\]

which is a scaled version of \(T(\Sigma^*)\). Therefore we prove the monotonicity of \(T\) as follows:

\[
F(T(\Sigma^*)) = F(\Sigma) \leq G(\Sigma, \Sigma^*) \leq G(\Sigma^*, \Sigma^*) = F(\Sigma^*).
\]

Because of the uniqueness of the minimizer of \(G(\Sigma, \Sigma^*)\), the equality in the second inequality of \(\text{VII.5}\) holds only when \(\Sigma = \Sigma^*\). Since \(\Sigma = cT(\Sigma^*)\) and \(\text{tr}(\Sigma^*) = \text{tr}(T(\Sigma^*)) = 1\), the equality in \(\text{VII.5}\) only holds when \(T(\Sigma^*) = \Sigma^*\).

Therefore the sequence \(F(\Sigma^{(k)}\) is monotone, and any accumulation points of the sequence \(\{\Sigma^{(k)}\}\), \(\Sigma\), satisfies \(F(T(\Sigma)) = F(\Sigma)\) and therefore \(T(\Sigma) = \Sigma\).

Applying \(T(\Sigma) = \Sigma\), we have

\[
\Sigma \sum_{x \in X} \frac{x^T x}{x^T \Sigma^{-1} x} = cI, \quad \text{for some } c \in \mathbb{R}.
\]

Let \(A = \log(\Sigma^{-1})\), applying \(\log \det(\Sigma) = -\text{tr}(A)\) and \(\frac{d}{dA} \exp(A) = \exp(A)\), the derivative of \(F(\Sigma)\) with respect to \(A\) is

\[
\frac{d}{dA} F(\Sigma) = \frac{1}{N} \sum_{x \in X} \frac{x^T x}{x^T \Sigma^{-1} x} - \frac{1}{D} I.
\]

Since the set \(\{\Sigma : A = \log(\Sigma^{-1})\} = \{\Sigma : \text{tr}(A) = 1\}\), applying \(\text{VII.6}\) the derivative of \(F(\Sigma)\) with respect to \(A\) in the set \(\{\Sigma : \det(\Sigma) = 1\}\) is 0 at \(c_0\), where \(c_0\) is a number chosen such that \(\det(c_0 \Sigma) = 1\). Since both the set \(\det(\Sigma) = 1\) and \(F(\Sigma)\) are geodesically convex (see \(\text{VII.1}\) for the convexity of the set), \(c_0\Sigma\) is the unique minimizer of \(F(\Sigma)\) in the set \(\{\Sigma : \det(\Sigma) = 1\}\). Applying the scale invariance of \(F(\sigma)\) in \(\text{I.6}\), \(\Sigma\) is the unique solution in the set \(\{\Sigma : \text{tr}(\Sigma) = 1\}\), which means that \(\Sigma\) is also the unique solution to \(\text{I.1}\).

\[\text{C. Proof of Lemma \text{III.3}}\]

Proof: If Lemma \(\text{III.3}\) does not hold, then there exists a sequence \(\Sigma_m\) such that it converges some \(\Sigma \in S_{++}(D) \setminus S_{++}(D)\), and the sequence \(F(\Sigma_m)\) is bounded. WLOG we assume that \(\lambda_j(\Sigma_m)\) and \(v_j(\Sigma_m)\) also converge for any \(1 \leq j \leq p\), where \(\lambda_j(\Sigma)\) and \(v_j(\Sigma)\) are the \(j\)-th eigenvalue and eigenvector of \(\Sigma\). This can be assumed since any sequence has a subsequence satisfying this property (eigenvectors and eigenvalues of \(\Sigma_m\) lie in a compact space).

We prove \(\text{III.3}\) by induction on the ambient dimension \(D\). When \(D = 2\), we have \(\dim(\ker(\Sigma)) = 1\), and

\[
\begin{align*}
F(\Sigma_m) \\
&\geq \frac{1}{N} \sum_{x \in X \setminus \ker(\Sigma)} \left( \log(\lambda_2(\Sigma_m)) + 2 \log(x^T v_2(\Sigma_m)) \right) \\
&\quad + \frac{1}{2} \log(\lambda_1(\Sigma_m)) + \frac{1}{2} \log(\lambda_1(\Sigma_m)).
\end{align*}
\]

When \(x \notin \ker(\Sigma)\), we have \(\lim \inf_{m \to \infty} x^T v_2(\Sigma_m) > 0\), therefore the term \(\log(x^T v_2(\Sigma_m))\) is bounded from below. Applying the assumption that \(\lambda_1(\Sigma_m)\) are bounded from below, \(\frac{1}{2} \log(\lambda_1(\Sigma_m))\) is also bounded from below. Applying the assumption \(\dim(X \setminus \ker(\Sigma)) > \frac{1}{2}\) and \(\lim_{m \to \infty} \lambda_2(\Sigma_m) = 0\), the RHS of \(\text{VII.7}\) converges to \(+\infty\), which is a contradiction to the assumption that \(F(\Sigma_m)\) is bounded, and therefore \(\text{III.3}\) is proved.

If \(\text{III.3}\) holds for the case \(\dim(x) < D_0\), then we will prove \(\text{III.3}\) for \(\dim(x) = D_0\). By the assumption on the convergence of eigenvectors and eigenvalues of \(\Sigma^{(k)}\), to prove \(\text{III.3}\) it is equivalent to prove that

\[
F'(\Sigma_m') \to \infty \text{ as } m \to \infty,
\]

where \(\Sigma_m' = P_L^T \Sigma_m P_L, \quad L = \ker(\Sigma), \quad d_0 = \dim(L)\) and \(F' : S_{++}(d_0) \to \mathbb{R}\) is defined by

\[
F'(\Sigma) = \frac{1}{N} \sum_{x \in X} \log((P_L^T x)^T \Sigma^{-1} P_L^T x) + \frac{1}{D_0} \log \det(\Sigma).
\]
An important observation is that \( \lim_{m \to \infty} \text{tr}(\Sigma'_m) = 0 \). Combine it with \( \frac{|X \setminus \hat{L}|}{N} > \frac{d_0}{D_0} \), we have
\[
\lim_{m \to \infty} F'(\Sigma'_m) - F'(\hat{\Sigma}'_m) = \left( \frac{d_0}{D_0} - \frac{|X \setminus \hat{L}|}{N} \right) \lim_{m \to \infty} \log \text{tr}(\Sigma'_m) = \infty.
\] (VII.9)

When \( \hat{\Sigma}'_m \) converges to a nonsingular matrix \( \Sigma' \),
\[
\lim_{m \to \infty} F'(\Sigma'_m) = F'(\Sigma') = C
\] (VII.10)
for some constant \( C \), and when \( \hat{\Sigma}'_m \) converges to a singular matrix, by induction
\[
\lim_{m \to \infty} F'(\Sigma'_m) = \infty.
\] (VII.11)
Combining (VII.9), (VII.10) and (VII.11), (VII.8) is proved and therefore Lemma III.3 is proved by induction.

\[ \square \]

D. Proof of Theorem [L]

The roadmap of the proof is as follows. We denote the set of outliers by \( X_0 = X \setminus L_*) and let \( X_1 = X \cap L_* \), and let \( N_1 = |X_1| \) and \( N_0 = |X_0| \). Assume that the solutions of (1.1) for the set \( X_1 \) and \( X_0 \) are \( I_d/d \) and \( I_{|D-d|/(D-d)} \) respectively, then we will prove that
\[
\lim_{k \to \infty} \Sigma^{(k)} = \frac{1}{d} \Pi_L, \tag{VII.12}
\]
which implies Theorem [L].

WLOG we can make these assumptions on the solutions of (1.1) for the set \( X_1 \) and \( X_0 \) since points in \( X_1 \) and \( X_0 \) lie in general positions, and applying Theorem III.1 to the solution to (1.1) for the set \( X_1 \) and \( X_0 \) are nonsingular. Assuming the solution of (1.1) for the set \( X_1 \) and \( X_0 \) are \( \Sigma_1 \) and \( \Sigma_2 \) respectively, then the following set \( \mathcal{X}' \), which is a linear transformation of \( X \), satisfies that the solution to (1.1) for the set \( X'_1 \) and \( X'_0 \) (they are generated by \( \mathcal{X}' \)) are \( I_d/d \) and \( I_{D-d}/(D-d) \):
\[
\mathcal{X}' = \{ \sqrt{\Sigma_1} \Pi_L x + \sqrt{\Sigma_2} \Pi_L x : x \in \mathcal{X} \}.
\]

The algorithm in (2.2) for \( \mathcal{X}' \) converges to \( \frac{1}{d} \Pi_L \), then by linear transformation the algorithm for \( X \) converges to \( P_L^* \Sigma_1 P_L^* \), whose range is also in \( L_* \). Therefore to prove Theorem [L] we only need to prove (2.2).

Now we start to prove (VII.12). Using the update formula in (2.2) and the assumption that the solutions of (1.1) to \( X_1 \) and \( X_0 \) are \( I_d/d \) and \( I_{D-d}/(D-d) \) respectively, we have
\[
\frac{\sum_{x \in X_1} P_L^T x x^T P_L^*}{\text{tr}\left( \sum_{x \in X_1} P_L^T x x^T P_L^* \right)} = \frac{1}{d} I_d.
\] (VII.13)

and
\[
\frac{\sum_{x \in X_0} P_L^T x x^T P_L^*}{\text{tr}\left( \sum_{x \in X_0} P_L^T x x^T P_L^* \right)} = \frac{1}{D-d} I_{|D-d|}.
\] (VII.14)

By checking the trace of the numerator of the LHS in (VII.13) and (VII.14) we have
\[
\sum_{x \in X_1} P_L^T x x^T P_L^* \left\| P_{L^*} x \right\|^2 = \frac{N_1}{d} I_d \tag{VII.15}
\]

and
\[
\sum_{x \in X_0} P_L^T x x^T P_L^* \left\| P_{L^*} x \right\|^2 = \frac{N_0}{D-d} I_{D-d} \tag{VII.16}
\]
Applying (VII.15) and (VII.16) we have
\[
\lambda_{\min}(P_{L^*}, \sum_{x \in X} x x^T P_{L^*}^*) \geq \lambda_{\min}(P_{L^*}, \sum_{x \in X} x x^T P_{L^*}^*) \|x\|^2 P_{L^*}^*
\]
\[
= \frac{N_1}{d} \lambda_{\min}(P_{L^*}, \Sigma P_{L^*})
\]

and
\[
\lambda_{\max}(P_{L^*}, \sum_{x \in X} x x^T P_{L^*}^*) \leq \lambda_{\max}(P_{L^*}, \sum_{x \in X} x x^T P_{L^*}^*) \|x\|^2 P_{L^*}^*
\]
\[
= \frac{N_0}{D-d} \lambda_{\max}(P_{L^*}, \Sigma P_{L^*})
\]
Combining them with the definition of the operator \( T \) in (III.4), we have
\[
\frac{\lambda_{\min}(P_{L^*}, T(\Sigma) P_{L^*})}{\lambda_{\max}(P_{L^*}, T(\Sigma) P_{L^*})} \geq \frac{\lambda_{\min}(P_{L^*}, \Sigma P_{L^*})}{\lambda_{\max}(P_{L^*}, \Sigma P_{L^*})},
\]
where \( \alpha = \frac{N_1(D-d)}{N_0 d} > 1 \) (it follows from the assumption \( \frac{|X \setminus L_*|}{|X|} > \frac{d}{D} \)).

Therefore
\[
\lim_{k \to \infty} \frac{\lambda_{\min}(P_{L^*}, \Sigma^{(k)} P_{L^*})}{\lambda_{\max}(P_{L^*}, \Sigma^{(k)} P_{L^*})} \geq \lim_{k \to \infty} \alpha^{k-1} \frac{\lambda_{\min}(P_{L^*}, \Sigma^{(1)} P_{L^*})}{\lambda_{\max}(P_{L^*}, \Sigma^{(1)} P_{L^*})} = \infty.
\] (VII.17)

Since \( \text{tr}(\Sigma^{(k)}) = 1 \) for all \( k > 0 \), we have
\[
\lim_{k \to \infty} \frac{\lambda_{\max}(P_{L^*}, \Sigma^{(k)} P_{L^*})}{\lambda_{\min}(P_{L^*}, \Sigma^{(k)} P_{L^*})} = 0
\] (VII.18)
Combining (VII.18) and the fact that \( \Sigma^{(k)} \) is positive semidefinite,
\[
\lim_{k \to \infty} \frac{\lambda_{\max}(P_{L^*}, \Sigma^{(k)} P_{L^*})}{\lambda_{\min}(P_{L^*}, \Sigma^{(k)} P_{L^*})} = 0.
\] (VII.19)
Since we already obtained (VII.18) and (VII.19), in order to prove (VII.12) we only need to prove that \( P_L^* \Sigma P_{L^*} \) converges to \( I_d/d \). Applying (VII.13) we have
\[
\lambda_{\max}(P_{L^*}, \sum_{x \in X} x x^T P_{L^*}) \leq \sum_{x \in X} \lambda_{\max}(P_{L^*}, x x^T P_{L^*}) + \lambda_{\max}(P_{L^*}, x x^T P_{L^*}) \|x\|^2 P_{L^*}^*
\]
\[
\leq \lambda_{\max}(P_{L^*}, \Sigma P_{L^*}) \sum_{x \in X} \|x\|^2 + \frac{N_1}{d} \lambda_{\max}(P_{L^*}, \Sigma P_{L^*})
\]
\[
\lambda_{\min}(P_L^T \sum_{x \in A} xx^T P_L) \geq \lambda_{\min}(\sum_{x \in A} P_L^T xx^T P_L) = \sum_{x \in A} \lambda_{\min}(P_L^T xx^T P_L) \\
\geq \frac{N_1}{d} \lambda_{\min}(P_L^T \Sigma P_L).
\]

Therefore
\[
\frac{\lambda_{\max}(P_L^T T(\Sigma) P_L)}{\lambda_{\min}(P_L^T T(\Sigma) P_L)} \leq \frac{\lambda_{\max}(P_L^T \Sigma P_L)}{\lambda_{\min}(P_L^T \Sigma P_L)}
\]
\[
= \frac{d \lambda_{\max}(P_L^T \Sigma P_L)}{N_1 \lambda_{\min}(P_L^T \Sigma P_L)}. \tag{VII.20}
\]

Now we will prove that
\[
\lambda_{\min}(P_L^T \Sigma(k) P_L) > c \text{ for all } k \geq 1 \text{ for some } c > 0. \tag{VII.21}
\]

If (VII.21) does not hold then there exists a subsequence \(\Sigma(k)\) such that
\[
\lim_{k \rightarrow \infty} \lambda_{\min}(P_L^T \Sigma(k) P_L) = 0.
\]
Applying (VII.18), (VII.19) and the induction argument in the proof of Lemma [III.3] we have \(\lim_{k \rightarrow \infty} F(\Sigma(k)) = \infty\), which contradicts the monotone property of the algorithm in (VII.5). Therefore (VII.21) is proved.

Applying (VII.22), there exists a constant \(C > 0\) such that
\[
\lambda_{\max}(P_L^T \Sigma(k) P_L) \leq C \alpha^{-k}. \tag{VII.22}
\]

Now we prove the existence of \(\lim_{k \rightarrow \infty} \frac{\lambda_{\max}(P_L^T \Sigma(k) P_L)}{\lambda_{\min}(P_L^T \Sigma(k) P_L)}\). If it does not exist, then there exists \(\varepsilon > 0\) such that for any sufficiently large \(K_0\), there exists \(k_0 > k_2 > K_0\) such that
\[
\frac{\lambda_{\max}(P_L^T \Sigma(k) P_L)}{\lambda_{\min}(P_L^T \Sigma(k) P_L)} > \varepsilon.
\]

Summing (VII.21) for \(\Sigma = \Sigma(k), \Sigma(k+1), \ldots, \Sigma(k+1)\), and apply (VII.21) and (VII.22), we have the contradiction for sufficiently large \(K_0\).

Next we will prove
\[
\lim_{k \rightarrow \infty} \frac{\lambda_{\max}(P_L^T \Sigma(k) P_L)}{\lambda_{\min}(P_L^T \Sigma(k) P_L)} = 1 \tag{VII.23}
\]

by contradiction, i.e., by assuming
\[
\lim_{k \rightarrow \infty} \frac{\lambda_{\max}(P_L^T \Sigma(k) P_L)}{\lambda_{\min}(P_L^T \Sigma(k) P_L)} = c_0 > 1.
\]

Since the sequence \(\Sigma(k)\) lies in compact space, there is a subsequence \(\{\Sigma(k_j)\}_{j \geq 1}\) converging to \(\Sigma\) with
\[
\frac{\lambda_{\max}(P_L^T \Sigma P_L)}{\lambda_{\min}(P_L^T \Sigma P_L)} = c_0 > 1. \tag{VII.24}
\]

Applying (VII.18) and (VII.19) we have \(\Pi_L \Sigma \Pi_L = \Sigma\). By simple calculation this property also holds for \(T^{n_0}(\Sigma)\) for any \(n\). Therefore the update \(T^{n_0}(\Sigma)\) can be considered as an update only depends on the set \(\mathcal{Y}_1\). Then by using Theorem [III.3] to the set \(\mathcal{Y}_1\) we have
\[
\lim_{n \rightarrow \infty} T^{n_0}(\Sigma) = \frac{1}{d} \Pi_L .
\]

therefore for any \(\varepsilon_1 > 0\), there exists some \(n_0 > 0\) such that
\[
\frac{\lambda_{\max}(P_L^T T^{n_0}(\Sigma) P_L)}{\lambda_{\min}(P_L^T T^{n_0}(\Sigma) P_L)} < 1 + \varepsilon_1. \tag{VII.25}
\]

Using the continuity of the mapping \(T^{n_0}\), for any \(n > 0\) there exists \(\varepsilon_2 > 0\) such that
\[
\frac{\lambda_{\max}(P_L^T T^{n_0}(\Sigma) P_L)}{\lambda_{\min}(P_L^T T^{n_0}(\Sigma) P_L)} < \eta,
\]
when \(\|\Sigma - \hat{\Sigma}\| < \varepsilon_2\).

Choose \(j_0\) large enough such that \(\|\Sigma(k_{j_0}) - \hat{\Sigma}\| < \varepsilon_2\), then applying (VII.25) and (VII.26) with \(\Sigma = \Sigma(k_{j_0})\) we obtain
\[
\frac{\lambda_{\max}(P_L^T \Sigma P_L)}{\lambda_{\min}(P_L^T \Sigma P_L)} < 1 + \varepsilon_1 + \eta. \tag{VII.27}
\]

Summing (VII.20) with \(\Sigma = \Sigma(k)\) for all \(k \geq k_{j_0} + n_0\), applying (VII.21) and (VII.22) we obtain that for some \(C_1 > 0\),
\[
\varepsilon_0 = \lim_{k \rightarrow \infty} \frac{\lambda_{\max}(P_L^T \Sigma(k) P_L)}{\lambda_{\min}(P_L^T \Sigma(k) P_L)}
\]
\[
= \frac{\lambda_{\max}(P_L^T \Sigma(k) P_L)}{\lambda_{\min}(P_L^T \Sigma(k) P_L)} + C_1 \alpha^{-k_{j_0} - n_0}
\]
\[
< 1 + C_1 \alpha^{-k_{j_0} - n_0} + \varepsilon_1 + \eta.
\]

Since we can choose \(\varepsilon_1, \eta\) arbitrarily small and \(k_{j_0}, n_0\) arbitrarily large, (VII.28) is a contradiction to (VII.24). Therefore (VII.23) is proved. Combining (VII.23) with (VII.18) and (VII.19) and notice that \(\text{tr}(\Sigma(k)) = 1\) for all \(k > 0\), we proved (VII.12).
11

[13] J. T. Kent and D. E. Tyler. Redescending M-estimates of multivariate location and scatter. *The Annals of Statistics*, 19(4):pp. 2102–2119, 1991.

[14] H. W. Kuhn. A note on Fermat’s problem. *Mathematical Programming*, 4:98–107, 1973. 10.1007/BF01584648.

[15] S. Lang. *Fundamentals of differential geometry*. Number v. 160 in Graduate texts in mathematics. Springer, 1999.

[16] K. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 27(5):684–698, 2005.

[17] G. Lerman, M. McCoy, J. A. Tropp, and T. Zhang. Robust computation of linear models, or how to find a needle in a haystack. Submitted February 2012. Available at http://arxiv.org/abs/1010.2471.

[18] G. Li and Z. Chen. Projection-Pursuit Approach to Robust Dispersion Matrices and Principal Components: Primary Theory and Monte Carlo. *Journal of the American Statistical Association*, 80(391):759–766, 1985.

[19] R. A. Maronna. Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4(1):pp. 51–67, 1976.

[20] R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 2006. Theory and methods.

[21] R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust statistics: Theory and methods*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 2006.

[22] M. McCoy and J. A. Tropp. Two proposals for robust PCA using semidefinite programming. *Elec. J. Stat.*, 5:1123–1160, 2011.

[23] C. Niculescu and L. Persson. *Convex functions and their applications: a contemporary approach*. Number v. 13 in CMS books in mathematics. Springer, 2006.

[24] X. Pennec, P. Fillard, and N. Ayache. A riemannian framework for tensor computing. *International Journal of Computer Vision*, 66:41–66, 2006. 10.1007/s11263-005-3222-z.

[25] S. Smith. Covariance, subspace, and intrinsic cramer-rao bounds. *Signal Processing, IEEE Transactions on*, 53(5):1610 – 1630, may 2005.

[26] M. Soltanolkotabi and E. J. Candès. A geometric analysis of subspace clustering with outliers. *CoRR*, abs/1112.4258, 2011.

[27] F. D. L. Torre and M. J. Black. A framework for robust subspace learning. *International Journal of Computer Vision*, 54:117–142, 2003. 10.1023/A:1023709501986.

[28] D. E. Tyler. A distribution-free m-estimator of multivariate scatter. *The Annals of Statistics*, 15(1):pp. 234–251, 1987.

[29] C. Udriste. *Convex functions and optimization methods on Riemannian manifolds*. Mathematics and its applications. Kluwer Academic Publishers, 1994.

[30] A. Wiesel. Geodesic convexity and covariance estimation. Submitted to *IEEE Trans. on Signal Processing*.

[31] A. Wiesel. On the convexity in kronecker structured covariance estimation. To be presented in SSP 2012.

[32] A. Wiesel. Unified framework to regularized covariance estimation in scaled gaussian models. *Signal Processing, IEEE Transactions on*, 60(1):29 –38, jan. 2012.

[33] H. Xu, C. Caramanis, and S. Mannor. Principal Component Analysis with Contaminated Data: The High Dimensional Case. In *Conference on Learning Theory (COLT 2010)*. 2010.

[34] H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via Outlier Pursuit. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2496–2504. 2010.

[35] T. Zhang and G. Lerman. A novel M-estimator for robust PCA. *preprint*, 2011. arXiv:1112.4863.