Journal of Machine Learning Research 15 (2014) 749-808 Submitted 12/11; Revised 6/13; Published 2/14

A Novel M-Estimator for Robust PCA

Teng Zhang ZHANG620@UMN.EDU
Institute for Mathematics and its Applications
University of Minnesota
Minneapolis, MN 55455, USA

Gilad Lerman* LERMAN@UMN.EDU
School of Mathematics
University of Minnesota
Minneapolis, MN 55455, USA

Editor: Martin Wainwright

Abstract

We study the basic problem of robust subspace recovery. That is, we assume a data set that some of its points are sampled around a fixed subspace and the rest of them are spread in the whole ambient space, and we aim to recover the fixed underlying subspace. We first estimate "robust inverse sample covariance" by solving a convex minimization procedure; we then recover the subspace by the bottom eigenvectors of this matrix (their number correspond to the number of eigenvalues close to 0). We guarantee exact subspace recovery under some conditions on the underlying data. Furthermore, we propose a fast iterative algorithm, which linearly converges to the matrix minimizing the convex problem. We also quantify the effect of noise and regularization and discuss many other practical and theoretical issues for improving the subspace recovery in various settings. When replacing the sum of terms in the convex energy function (that we minimize) with the sum of squares of terms, we obtain that the new minimizer is a scaled version of the inverse sample covariance (when exists). We thus interpret our minimizer and its subspace (spanned by its bottom eigenvectors) as robust versions of the empirical inverse covariance and the PCA subspace respectively. We compare our method with many other algorithms for robust PCA on synthetic and real data sets and demonstrate state-of-the-art speed and accuracy.

Keywords: principal components analysis, robust statistics, M-estimator, iteratively re-weighted least squares, convex relaxation

1. Introduction

The most useful paradigm in data analysis and machine learning is arguably the modeling of data by a low-dimensional subspace. The well-known total least squares solves this modeling problem by finding the subspace minimizing the sum of squared errors of data points. This is practically done via principal components analysis (PCA) of the data matrix. Nevertheless, this procedure is highly sensitive to outliers. Many heuristics have been proposed for robust recovery of the underlying subspace. Recent progress in the rigorous study of sparsity and low-rank of data has resulted in provable convex algorithms for this purpose. Here, we propose a different rigorous and convex approach, which is a special M-estimator.

*. Gilad Lerman is the corresponding author.

Robustness of statistical estimators has been carefully studied for several decades (Huber and Ronchetti, 2009; Maronna et al., 2006). A classical example is the robustness of the geometric median (Lopuhaä and Rousseeuw, 1991). For a data set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^D$, the geometric median is the minimizer of the following function of $\mathbf{y} \in \mathbb{R}^D$:

$$\sum_{i=1}^{N} \|\mathbf{y} - \mathbf{x}_i\|, \tag{1}$$

where $\|\cdot\|$ denotes the Euclidean norm. This is a typical example of an M-estimator, that is, a minimizer of a function of the form $\sum_{i=1}^N \rho(r_i)$, where r_i is a residual of the ith data point, $\mathbf{x}_i$, from the parametrized object we want to estimate. Here, $r_i = \|\mathbf{y}-\mathbf{x}_i\|$, $\rho(x) = |x|$ and we estimate $\mathbf{y} \in \mathbb{R}^D$, which is parametrized by its D coordinates.

There are several obstacles in developing robust and effective estimators for subspaces. For simplicity, we discuss here estimators of linear subspaces and thus assume that the data is centered at the origin.[1] A main obstacle is due to the fact that the set of d-dimensional linear subspaces in $\mathbb{R}^D$, that is, the Grassmannian $\mathrm{G}(D,d)$, is not convex. Therefore, a direct optimization on $\mathrm{G}(D,d)$ (or a union of $\mathrm{G}(D,d)$ over different d's) will not be convex (even not geodesically convex) and may result in several (or many) local minima. Another problem is that extensions of simple robust estimators of vectors to subspaces (e.g., using l_1-type averages) can fail by a single far away outlier. For example, one may extend the d-dimensional geometric median minimizing (1) to the minimizer over $\mathrm{L} \in \mathrm{G}(D,d)$ of the function

$$\sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{P}_{\mathrm{L}}\mathbf{x}_i\| \equiv \sum_{i=1}^{N} \|\mathbf{P}_{\mathrm{L}^\perp}\mathbf{x}_i\|, \tag{2}$$

where $\mathrm{L}^\perp$ is the orthogonal complement of L and $\mathbf{P}_{\mathrm{L}}$ and $\mathbf{P}_{\mathrm{L}^\perp}$ are the orthogonal projections on L and $\mathrm{L}^\perp$ respectively (see, e.g., Ding et al., 2006; Lerman and Zhang, 2010). However, a single outlier with arbitrarily large magnitude will enforce the minimizer of (2) to contain it.

The first obstacle can be resolved by applying a convex relaxation of the minimization of (2) so that subspaces are mapped into a convex set of matrices (the objective function may be adapted respectively). Indeed, the subspace recovery proposed by Xu et al. (2010b) can be interpreted in this way. Their objective function has one component which is similar to (2), though translated to matrices. They avoid the second obstacle by introducing a second component, which penalizes inliers of large magnitude (so that outliers of large magnitude may not be easily identified as inliers). However, the combination of the two components involves a parameter that needs to be carefully estimated.

Here, we suggest a different convex relaxation that does not introduce arbitrary parameters and its implementation is significantly faster. However, it introduces some restrictions on the distributions of inliers and outliers. Some of these restrictions have analogs in other

1. This is a common assumption to reduce the complexity of the subspace recovery problem (Candès et al., 2011; Xu et al., 2010b, 2012; McCoy and Tropp, 2011), where McCoy and Tropp (2011) suggest centering by the geometric median. Nevertheless, our methods easily adapt to affine subspace fitting by simultaneously estimating both the offset and the shifted linear component, but the justification is a bit more complicated then.

works (see, e.g., §2.2), while others are unique to this framework (see §2.3 and the non-technical description of all of our restrictions in §1.2).

1.1 Previous Work

Many algorithms (or pure estimators) have been proposed for robust subspace estimation or equivalently robust low rank approximation of matrices. Maronna (1976), Huber and Ronchetti (2009, §8), Devlin et al. (1981), Davies (1987), Xu and Yuille (1995), Croux and Haesbroeck (2000) and Maronna et al. (2006, §6) estimate a robust covariance matrix. Some of these methods use M-estimators (Maronna et al., 2006, §6) and compute them via iteratively re-weighted least squares (IRLS) algorithms, which linearly converge (Arslan, 2004). The convergence of algorithms based on other estimators or strategies is not as satisfying. The objective functions of the MCD (Minimum Covariance Determinant) and S-estimators converge (Maronna et al., 2006, §6), but no convergence rates are specified. Moreover, there are no guarantees for the actual convergence to the global optimum of these objective functions. There is no good algorithm for the MVE (Minimum Volume Ellipsoid) or Stahel-Donoho estimators (Maronna et al., 2006, §6). Furthermore, convergence analysis is problematic for the online algorithm of Xu and Yuille (1995).

Li and Chen (1985), Ammann (1993), Croux et al. (2007), Kwak (2008) and McCoy and Tropp (2011, §2) find low-dimensional projections by "Projection Pursuit" (PP), now commonly referred to as PP-PCA (the initial proposal is due to Huber, see, e.g., Huber and Ronchetti, 2009, p. 204 of first edition). The PP-PCA procedure is based on the observation that PCA maximizes the projective variance and can be implemented incrementally by computing the residual principal component or vector each time. Consequently, PP-PCA replaces this variance by a more robust function in this incremental implementation. Most PP-based methods are based on non-convex optimization and consequently lack satisfying guarantees. In particular, Croux et al. (2007) do not analyze convergence of their non-convex PP-PCA and Kwak (2008) only establishes convergence to a local maximum. McCoy and Tropp (2011, §2) suggest a convex relaxation for PP-PCA. However, they do not guarantee that the output of their algorithm coincides with the exact maximizer of their energy (though they show that the energies of the two are sufficiently close). Ammann (1993) applies a minimization on the sphere, which is clearly not convex. It iteratively tries to locate vectors spanning the orthogonal complement of the underlying subspace, that is, $D-d$ vectors for a subspace in $\mathrm{G}(D,d)$. We remark that our method also suggests an optimization revealing the orthogonal complement, but it requires a single convex optimization, which is completely different from the method of Ammann (1993).

Torre and Black (2001, 2003), Brubaker (2009) and Xu et al. (2010a) remove possible outliers, followed by estimation of the underlying subspace by PCA. These methods are highly non-convex. Nevertheless, Xu et al. (2010a) provide a probabilistic analysis for their near recovery of the underlying subspace.

The non-convex minimization of (2) as a robust alternative for principal component analysis was suggested earlier by various authors for hyperplane modeling (Osborne and Watson, 1985; Späth and Watson, 1987; Nyquist, 1988; Bargiela and Hartley, 1993), surface modeling (Watson, 2001, 2002), subspace modeling (Ding et al., 2006) and multiple subspaces modeling (Zhang et al., 2009). This minimization also appeared in a pure geometric-analytic

context of general surface modeling without outliers (David and Semmes, 1991). Lerman and Zhang (2010, 2011) have shown that this minimization can be robust to outliers under some conditions on the sampling of the data.

Ke and Kanade (2003) tried to minimize (over all low-rank approximations) the element-wise l_1 norm of the difference of a given matrix and its low-rank approximation. Chandrasekaran et al. (2011) and Candès et al. (2011) have proposed to minimize a linear combination of such an l_1 norm and the nuclear norm of the low-rank approximation in order to find the optimal low-rank estimator. Candès et al. (2011) considered the setting where uniformly sampled elements of the low-rank matrix are corrupted, which does not apply to our outlier model (where only some of the rows are totally corrupted). Chandrasekaran et al. (2011) consider a general setting, though their underlying condition is too restrictive; weaker condition was suggested by Hsu et al. (2011), though it is still not sufficiently general. Nevertheless, Chandrasekaran et al. (2011) and Candès et al. (2011) are groundbreaking to the whole area, since they provide rigorous analysis of exact low-rank recovery with unspecified rank.

Xu et al. (2010b) and McCoy and Tropp (2011) have suggested a strategy analogous to Chandrasekaran et al. (2011) and Candès et al. (2011) to solve the outlier problem. They divide the matrix $\mathbf{X}$ whose rows are the data points as follows: $\mathbf{X} = \mathbf{L} + \mathbf{O}$, where $\mathbf{L}$ is low-rank and $\mathbf{O}$ represents outliers (so that only some of its rows are non-zero). They minimize $\|\mathbf{L}\|_* + \lambda\|\mathbf{O}\|_{(2,1)}$, where $\|\cdot\|_*$ and $\|\cdot\|_{(2,1)}$ denote the nuclear norm and sum of l_2 norms of rows respectively and λ is a parameter that needs to be carefully chosen. We note that the term $\|\mathbf{O}\|_{(2,1)}$ is analogous to (2). Xu et al. (2012) have established an impressive theory showing that under some incoherency conditions, a bound on the fraction of outliers and correct choice of the parameter λ, they can exactly recover the low-rank approximation. Hsu et al. (2011) and Agarwal et al. (2012a) improved error bounds for this estimator as well as for the ones of Chandrasekaran et al. (2011) and Candès et al. (2011).

In practice, the implementations by Chandrasekaran et al. (2011), Candès et al. (2011), Xu et al. (2010b) and McCoy and Tropp (2011) use the iterative procedure described by Lin et al. (2009). The difference between the objective functions of the minimizer and its estimator obtained at the kth iteration is of order $O(k^{-2})$ (Lin et al., 2009, Theorem 2.1). On the other hand, for our algorithm the convergence rate is of order $O(\exp(-ck))$ for some constant c (i.e., it r-linearly converges). This rate is the order of the Frobenius norm of the difference between the minimizer sought by our algorithm (formulated in (4) below) and its estimator obtained at the kth iteration (it is also the order of the difference of the regularized objective functions of these two matrices). Recently, Agarwal et al. (2012b) showed that projected gradient descent algorithms for these estimators obtain linear convergence rates, though with an additional statistical error.

Our numerical algorithm can be categorized as IRLS. Weiszfeld (1937) used a procedure similar to ours to find the geometric median. Lawson (1961) later used it to solve uniform approximation problems by the limits of weighted l_p-norm solutions. This procedure was generalized to various minimization problems, in particular, it is native to M-estimators (Huber and Ronchetti, 2009; Maronna et al., 2006), and its linear convergence was proved for special instances (see, e.g., Cline, 1972; Voss and Eckhardt, 1980; Chan and Mulet, 1999). Recently, IRLS algorithms were also applied to sparse recovery and matrix completion (Daubechies et al., 2010; Fornasier et al., 2011).

1.2 This Work

We suggest another convex relaxation of the minimization of (2). We note that the original minimization is over all subspaces L or equivalently all orthogonal projectors $\mathbf{P} \equiv \mathbf{P}_{\mathrm{L}^\perp}$. We can identify $\mathbf{P}$ with a $D \times D$ matrix satisfying $\mathbf{P}^2 = \mathbf{P}$ and $\mathbf{P}^T = \mathbf{P}$ (where $\cdot^T$ denotes the transpose). Since the latter set is not convex, we relax it to include all symmetric matrices, but avoid singularities by enforcing unit trace. That is, we minimize over the set:

$$\mathbb{H} := \{\mathbf{Q} \in \mathbb{R}^{D\times D} : \mathbf{Q} = \mathbf{Q}^T, \mathrm{tr}(\mathbf{Q}) = 1\} \tag{3}$$

as follows

$$\hat{\mathbf{Q}} = \operatorname*{arg\,min}_{\mathbf{Q}\in\mathbb{H}} F(\mathbf{Q}), \text{ where } F(\mathbf{Q}) := \sum_{i=1}^{N} \|\mathbf{Q}\mathbf{x}_i\|. \tag{4}$$

For the noiseless case (i.e., inliers lie exactly on L^*), we estimate the subspace L^* by

$$\hat{\mathrm{L}} := \ker(\hat{\mathbf{Q}}). \tag{5}$$

If the intrinsic dimension d is known (or can be estimate from the data), we estimate the subspace by the span of the bottom d eigenvectors of $\hat{\mathbf{Q}}$ (or equivalently, the top d eigenvectors of $-\hat{\mathbf{Q}}$). This procedure is robust to sufficiently small levels of noise. We refer to it as the Geometric Median Subspace (GMS) algorithm and summarize it in Algorithm 1. We elaborate on this scheme throughout the paper,

Algorithm 1 The Geometric Median Subspace Algorithm

Input: $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N \subseteq \mathbb{R}^D$: data, d: dimension of L^*, an algorithm for minimizing (4)
Output: $\hat{\mathrm{L}}$: a d-dimensional linear subspace in $\mathbb{R}^D$.
Steps:
- $\{\mathbf{v}_i\}_{i=1}^d =$ the bottom d eigenvectors of $\hat{\mathbf{Q}}$ (see (4))
- $\hat{\mathrm{L}} = \mathrm{Sp}(\{\mathbf{v}_i\}_{i=1}^d)$

We remark that $\hat{\mathbf{Q}}$ is semi-definite positive (we verify this later in Lemma 14). We can thus restrict $\mathbb{H}$ to contain only semi-definite positive matrices and thus make it even closer to a set of orthogonal projectors. Theoretically, it makes sense to require that the trace of the matrices in $\mathbb{H}$ is $D-d$ (since they are relaxed versions of projectors onto the orthogonal complement of a d-dimensional subspace). However, scaling of the trace in (3) results in scaling the minimizer of (4) by a constant, which does not effect the subspace recovery procedure.

We note that (4) is an M-estimator with residuals $r_i = \|\mathbf{Q}\mathbf{x}_i\|$, $1 \le i \le N$, and $\rho(x) = |x|$. Unlike (2), which can also be seen as a formal M-estimator, the estimator $\hat{\mathbf{Q}}$ is unique under a weak condition that we will state later.

We are unaware of similar formulations for the problem of robust PCA. Nevertheless, the Low-Rank Representation (LRR) framework of Liu et al. (2010, 2013) for modeling data by multiple subspaces (and not a single subspace as in here) is formally similar. LRR tries to assign to a data matrix $\mathbf{X}$, which is viewed as a dictionary of N column vectors in $\mathbb{R}^D$, dictionary coefficients $\mathbf{Z}$ by minimizing $\lambda\|\mathbf{Z}\|_* + \|(\mathbf{X}(\mathbf{I}-\mathbf{Z}))^T\|_{(2,1)}$ over all $\mathbf{Z} \in \mathbb{R}^{N\times N}$, where λ is a free parameter. Our formulation can be obtained by their formulation with

$\lambda = 0$, $\mathbf{Q} = (\mathbf{I} - \mathbf{Z})^T$ and the additional constraint $\mathrm{tr}(\mathbf{Z}) = D - 1$ (which is equivalent with the scaling $\mathrm{tr}(\mathbf{Q}) = 1$), where $\{\mathbf{x}_i\}_{i=1}^N$ are the row vectors of $\mathbf{X}$ (and not the column vectors that represent the original data points). In fact, our work provides some intuition for LRR as robust recovery of the low rank row space of the data matrix and its use (via $\mathbf{Z}$) in partitioning the column space into multiple subspaces. We also remark that a trace 1 constraint is quite natural in convex relaxation problems and was applied, for example, in the convex relaxation of sparse PCA (d'Aspremont et al., 2007), though the optimization problem there is completely different.

Our formulation is rather simple and intuitive, but results in the following fundamental contributions to robust recovery of subspaces:

1. We prove that our proposed minimization can achieve exact recovery under some assumptions on the underlying data (which we clarify below) and without introducing an additional parameter.

2. We propose a fast iterative algorithm for achieving this minimization and prove its linear convergence.

3. We demonstrate the state-of-the-art accuracy and speed of our algorithm when compared with other methods on both synthetic and real data sets.

4. We establish the robustness of our method to noise and to a common regularization of IRLS algorithms.

5. We explain how to incorporate knowledge of the intrinsic dimension and also how to estimate it empirically.

6. We show that when replacing the sum of norms in (4) by the sum of squares of norms, then the modified minimizer $\hat{\mathbf{Q}}$ is a scaled version of the empirical inverse covariance. The subspace spanned by the bottom d eigenvectors is clearly the d-dimensional subspace obtained by PCA. The original minimizer of (4) can thus be interpreted as a robust version of the inverse covariance matrix.

7. We show that previous and well-known M-estimators (Maronna, 1976; Huber and Ronchetti, 2009; Maronna et al., 2006) do not solve the subspace recovery problem under a common assumption.

1.3 Exact Recovery and Conditions for Exact Recovery by GMS

In order to study the robustness to outliers of our estimator for the underlying subspace, we formulate the exact subspace recovery problem (see also Xu et al. 2012). This problem assumes a fixed d-dimensional linear subspace L^*, inliers sampled from L^* and outliers sampled from its complement; it asks to recover L^* as well as identify correctly inliers and outliers.

In the case of point estimators, like the geometric median minimizing (1), robustness is commonly measured by the breakdown point of the estimator (Huber and Ronchetti, 2009; Maronna et al., 2006). Roughly speaking, the breakdown point measures the proportion of

arbitrarily large observations (that is, the proportion of "outliers") an estimator can handle before giving an arbitrarily large result.

In the case of estimating subspaces, we cannot directly extend this definition, since the set of subspaces, that is, the Grassmannian (or unions of it), is compact, so we cannot talk about "an arbitrarily large result", that is, a subspace with arbitrarily large distance from all other subspaces. Furthermore, given an arbitrarily large data point, we can always form a subspace containing it; that is, this point is not arbitrarily large with respect to this subspace. Instead, we identify the outliers as the ones in the compliment of L^* and we are interested in the largest fraction of outliers (or smallest fraction of inliers per outliers) allowing exact recovery of L^*. Whenever an estimator can exactly recover a subspace under a given sampling scenario we view it as robust and measure its effectiveness by the largest fraction of outliers it can tolerate. However, when an estimator cannot exactly recover a subspace, one needs to bound from below the distance between the recovered subspace and the underlying subspace of the model. Alternatively, one would need to point out at interesting scenarios where exact recovery cannot even occur in the limit when the number of points approaches infinity. We are unaware of other notions of robustness of subspace estimation (but of robustness of covariance estimation, which does not apply here; see, for example, §6.2.1 of Maronna et al. 2006).

In order to guarantee exact recovery of our estimator we basically require three kinds of restrictions on the underlying data, which we explain here on a non-technical level (technical discussion appears in §2). First of all, the inliers need to permeate through the whole underlying subspace L^*, in particular, they cannot concentrate on a lower dimensional subspace of L^*. Second of all, outliers need to permeate throughout the whole complement of L^*. This assumption is rather restrictive and its violation is a failure mode of the algorithm. We thus show that this failure mode does not occur when the knowledge of d is used appropriately. We also suggest some practical methods to avoid this failure mode when d is unknown (see §5.1). Third of all, the "magnitude' of outliers needs to be restricted. We may initially scale all points to the unit sphere in order to avoid extremely large outliers. However, we still need to avoid outliers concentrating along lines, which may have an equivalent effect of a single arbitrarily large outlier. Figure 1 (which appears later in §2) demonstrates cases where these assumptions are not satisfied.

The failure mode discussed above occurs in particular when the number of outliers is rather small and the dimension d is unknown. While we suggest some practical methods to avoid it (see §5.1), we also note that there are many modern applications with high percentages of outliers, where this failure mode may not occur. In particular, computer vision data often contain high percentages of outliers (Stewart, 1999; Chin et al., 2012). However, such data usually involve multiple geometric models, in particular, multiple underlying linear subspaces. We believe that the robust subspace modeling is still relevant to these kinds of data. First of all, robust single subspace strategies can be well-integrated into common schemes of modeling data by multiple subspaces. For example, the K-flats algorithm is based on repetitive clustering and single subspace modeling per cluster (Tipping and Bishop, 1999; Bradley and Mangasarian, 2000; Tseng, 2000; Ho et al., 2003; Zhang et al., 2009, 2012) and the LBF and SLBF algorithms use local subspace modeling (Zhang et al., 2010, 2012). Second of all, some of the important preprocessing tasks in computer vision require single subspace modeling. For example, in face recognition, a preprocessing step

requires efficient subspace modeling of images of the same face under different illuminating conditions (Basri and Jacobs, 2003; Basri et al., 2011). There are also problems in computer vision with more complicated geometric models and large percentage of corruption, where our strategies can be carefully adapted. One important example is the synchronization problem, which finds an important application in Cryo-EM. The goal of this problem is to recover rotation matrices $R_1, \ldots, R_N \in SO(3)$ from noisy and mostly corrupted measurements of $R_i^{-1}R_j$ for some values of $1 \leq i, j \leq N$. Wang and Singer (2013) adapted ideas of both this work and Lerman et al. (2012) to justify and implement a robust solution for the synchronization problem.

1.4 Recent Subsequent Work

In the case where d is known, Lerman et al. (2012) followed this work and suggested a tight convex relaxation of the minimization of (31) over all projectors $\mathbf{P}_{\mathrm{L}^\perp}$ of rank d. Their optimizer, which they refer to as the REAPER (of the needle-in-haystack problem) minimize the same function $F(\mathbf{Q})$ (see (4)) over the set

$$\mathbb{H}' = \{\mathbf{Q} \in \mathbb{R}^{D\times D} : \mathbf{Q} = \mathbf{Q}^T, \mathrm{tr}(\mathbf{Q}) = 1, \|\mathbf{Q}\| \leq \frac{1}{D-d}\}.$$

They estimate the underlying subspace by the bottom d eigenvectors of the REAPER. The new constraints in $\mathbb{H}'$ result in more elegant conditions for exact recovery and tighter probabilistic theory (due to the tighter relaxation). Since d is known the failure mode of GMS mentioned above is avoided. Their REAPER algorithm for computing the REAPER is based on the IRLS procedure of this paper with additional constraints, which complicate its analysis. The algorithmic and theoretical developments of Lerman et al. (2012) are based on the ones here.

While the REAPER framework applies a tighter relaxation, the GMS framework still has several advantages over the REAPER framework. First of all, in various practical situations the dimension of the data is unknown and thus REAPER is inapplicable. On the other hand, GMS can be used for dimension estimation, as we demonstrate in §6.3. Second of all, the GMS algorithm is faster than REAPER (the REAPER requires additional eigenvalue decomposition of a $D \times D$ matrix at each iteration of the IRLS algorithm). Furthermore, we present here a complete theory for the linear convergence of the GMS algorithm, where the convergence theory for the REAPER algorithm is currently incomplete. Third of all, when the failure mode mentioned above is avoided, the empirical performances of REAPER and GMS are usually comparable (while GMS is faster). At last, GMS and REAPER have different objectives with different consequences. REAPER aims to find a projector onto the underlying subspace. On the other hand, GMS aims to find a "generalized inverse covariance" (see §3.3) and is formally similar to other M-estimators (see §3.1 and §3.2). Therefore, the eigenvalues and eigenvectors of the GMS estimator (i.e., the "generalized inverse covariance") can be interpreted as robust eigenvalues and eigenvectors of the empirical covariance (see §6.3 and §6.5).

1.5 Structure of This Paper

In §2 we establish exact and near subspace recovery via the GMS algorithm. We also carefully explain the common obstacles for robust subspace recovery and the way they are

handled by previous rigorous solutions (Candès et al., 2011; Chandrasekaran et al., 2011; Xu et al., 2012) as well as our solution. Section 3 aims to interpret our M-estimator in two different ways. First of all, it shows a formal similarity to a well-known class of M-estimators (Maronna, 1976; Huber and Ronchetti, 2009; Maronna et al., 2006), though clarifies the difference. Those estimators aims to robustly estimate the sample covariance. However, we show there that unlike our M-estimator, they cannot solve the subspace recovery problem (under a common assumption). Second of all, it shows that non-robust adaptation of our M-estimator provides both direct estimation of the inverse covariance matrix as well as convex minimization equivalent to the non-convex total least squares (this part requires full rank data and thus a possible initial dimensionality reduction but without any loss of information). We thus interpret (4) as a robust estimation of the inverse covariance. In §4 we propose an IRLS algorithm for minimizing (4) and establish its linear convergence. Section 5 discusses practical versions of the GMS procedure that allow more general distributions than the ones guaranteed by the theory. One of these versions, the Extended GMS (EGMS) even provides robust alternative to principal components. In §6 we demonstrate the state-of-the-art accuracy and speed of our algorithm when compared with other methods on both synthetic and real data sets and also numerically clarify some earlier claims. Section 7 provides all details of the proofs and §8 concludes with brief discussion.

2. Exact and Near Subspace Recovery by GMS

We establish exact and near subspace recovery by the GMS algorithm. In §2.1 we formulate the problems of exact and near subspace recovery. In §2.2 we describe common obstacles for solving these problems and how they were handled in previous works; in §2.3 we formulate some conditions that the data may satisfy; whereas in §2.4 we claim that these conditions are sufficient to avoid the former obstacles, that is, they guarantee exact recovery (see Theorem 1); We also propose weaker conditions for exact recovery and demonstrate their near-tightness in §2.4.1. Section 2.5 describes a simple general condition for uniqueness of GMS (beyond the setting of exact recovery). Section 2.6 establishes (with some specified limitations) unique exact recovery with high probability under basic probabilistic models (see Theorems 4 and 5); it also covers cases with asymmetric outliers. At last, §2.7 and §2.8 establish results for near recovery under noise and under regularization respectively.

2.1 Problem Formulation

Let us repeat the formulation of the exact subspace recovery problem, which we motivated in §1.2 as a robust measure for the performance of our estimator. We assume a linear subspace $\mathrm{L}^* \in \mathrm{G}(D,d)$ and a data set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$, which contains inliers sampled from L^* and outliers sampled from $\mathbb{R}^D \setminus \mathrm{L}^*$. Given the data set $\mathcal{X}$ and no other information, the objective of the exact subspace recovery problem is to exactly recover the underlying subspace L^*.

In order to make the problem well-defined, one needs to assume some conditions on the sampled data set, which may vary with the proposed solution. We emphasize that this is a formal mathematical problem, which excludes some ambiguous scenarios and allows us to determine admissible distributions of inliers and outliers.

In the noisy case (where inliers do not lie on L^*, but perturbed by noise), we ask about near subspace recovery, that is, recovery up to an error depending on the underlying noise level. We argue below that in this case additional information on the model is needed. Here we assume the knowledge of d, though under some assumptions we can estimate d from the data (as we demonstrate later). We remark that exact asymptotic recovery under some conditions on the noise distribution is way more complicated and is discussed in another work (Coudron and Lerman, 2012).

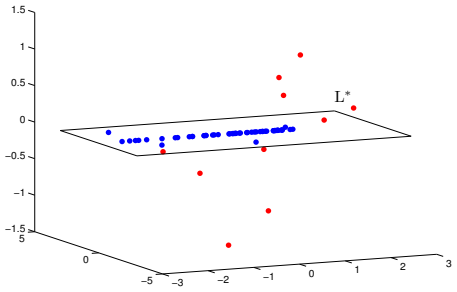
2.2 Common Difficulties with Subspace Recovery

We introduce here three typical enemies of subspace recovery and exemplify them in Figure 1. We also explain how they are handled by the previous convex solutions for exact recovery of subspaces as well as low-rank matrices (Chandrasekaran et al., 2011; Candès et al., 2011; Xu et al., 2012).

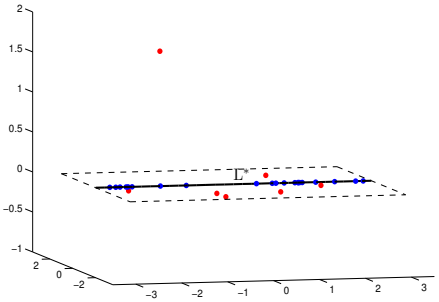
A type 1 enemy occurs when the inliers are mainly sampled from a subspace $\mathrm{L}' \subset \mathrm{L}^*$. In this case, it seems impossible to recover L^*. We would expect a good algorithm to recover L' (instead of L^*) or a subspace containing it with slightly higher dimension (see for example Figure 1(a)). Chandrasekaran et al. (2011), Candès et al. (2011) and Xu et al. (2012) have addressed this issue by requiring incoherence conditions for the inliers. For example, if m and $N-m$ points are sampled from L' and $L^* \setminus L'$ respectively, then the incoherency condition of Xu et al. (2012) requires that $\mu \geq N/(\dim(\mathrm{L}^*) \cdot (N-m))$, where μ is their incoherency parameter. That is, their theory holds only when the fraction of points sampled from $L^* \setminus L'$ is sufficiently large.

A type 2 enemy occurs when the outliers are mainly sampled from a subspace $\tilde{\mathrm{L}}$ such that $\dim(\tilde{\mathrm{L}} \oplus \mathrm{L}^*) < D$. In this case $\mathrm{L}^* \oplus \tilde{\mathrm{L}}$ can be mistakenly identified as the low-rank subspace (see for example Figure 1(b)). This is a main issue when the intrinsic dimension is unknown; if on the other hand the intrinsic dimension is known, then one can often overcome this enemy. Candès et al. (2011) handle it by assuming that the distribution of corrupted elements is uniform. Chandrasekaran et al. (2011) address it by restricting their parameter μ (see their main condition, which is used in Theorem 2 of their work, and their definition of μ in (1.2) of their work) and consequently limit the values of the mixture parameter (denoted here by λ). On the other hand, Xu et al. (2012) use the true percentage of outliers to infer the right choice of the mixture parameter λ. That is, they practically invoke model selection (for estimating this percentage) in order to reject $\tilde{\mathrm{L}} \oplus \mathrm{L}^*$ and choose the true model, which is L^*.

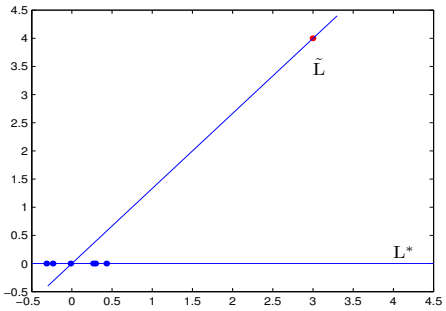
A type 3 enemy occurs due to large magnitudes of outliers. For example, a single outlier with arbitrarily large magnitude will be contained in the minimizer of (2), which will thus be different than the underlying subspace (see for example Figure 1(c)). Also, many outliers with not-so-small magnitudes that lie around a fixed line may have the effect of a single large outlier (see for example Figure 1(d)). This enemy is avoided by Chandrasekaran et al. (2011), Candès et al. (2011) and Xu et al. (2012) by the additional mixture component of nuclear norm, which penalizes the magnitude (or combined magnitude) of the supposed inliers (so that outliers of large magnitude may not be easily identified as inliers). It is interesting to note that if the rank is used instead of the nuclear norm (as sometimes advocated), then it will not resolve this issue.



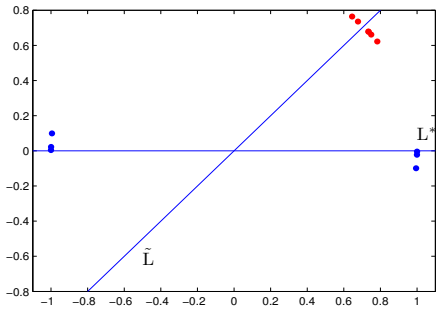
(a) Example of a type 1 enemy: L^* is a plane represented by a rectangle, "inliers" (in L^*) are colored blue and "outliers" (in $\mathbb{R}^3 \setminus L^*$) red. Most inliers lie on a line inside L^*. It seems unlikely to distinguish between inliers, which are not on "the main line", and the outliers. It is thus likely to recover the main line instead of L^*.



(b) Example of a type 2 enemy: L^* is a line represented by a black line segment, "inliers" (in L^*) are colored blue and "outliers" (in $\mathbb{R}^3 \setminus L^*$) red. All outliers but one lie within a plane containing L^*, which is represented by a dashed rectangle. There seems to be stronger distinction between the points on this plane and the isolated outlier than the original inliers and outliers. Therefore, an exact recovery algorithm may output this plane instead of L^*.



(c) Example 1 of a type 3 enemy: The inliers (in blue) lie on the line L^* and there is a single outlier (in red) with relatively large magnitude. An exact recovery algorithm can output the line $\tilde{L}$ (determined by the outlier) instead of L^*. If the data is normalized to the unit circle, then any reasonable robust subspace recovery algorithm can still recover L^*.



(d) Example 2 of a type 3 enemy: Points are normalized to lie on the unit circle, inliers (in blue) lie around the line L^* and outliers (in red) concentrate around another line, $\tilde{L}$. A subspace recovery algorithm can output $\tilde{L}$ instead of L^*.

Figure 1: Enemies of the mathematical formulation of exact subspace recovery.

Another issue for our mathematical problem of exact subspace recovery is whether the subspace obtained by a proposed algorithm is unique. Many of the convex algorithms depend on convex l_1-type methods that may not be strictly convex. But it may still happen that in the setting of pure inliers and outliers and under some conditions avoiding the three types of enemies, the recovered subspace is unique (even though it may be obtained by several non-unique minimizers). This is indeed the case in Chandrasekaran et al. (2011), Candès et al. (2011), Xu et al. (2012) and our own work. Nevertheless, uniqueness of our minimizer (and not the recovered subspace) is important for analyzing the numerical algorithm approximating it and for perturbation analysis (e.g., when considering near recovery with noisy data). It is also helpful for practically verifying the conditions we will propose for exact recovery. Uniqueness of the minimizer (and not just the subspace) is also important in Chandrasekaran et al. (2011) and Candès et al. (2011) and they thus established conditions for it.

At last, we comment that subspace recovery with unknown intrinsic dimension may require a model selection procedure (possibly implicitly). That is, even though one can provide a theory for exact subspace recovery (under some conditions), which might be stable to perturbations, in practice, some form of model selection will be necessary in noisy cases. For example, the impressive theories by Chandrasekaran et al. (2011) and Xu et al. (2012) require the estimation of the mixture parameter λ. Xu et al. (2012) propose such an estimate for λ, which is based on knowledge of the data set (e.g., the distribution of corruptions and the fraction of outliers). However, we noticed that in practice this proposal did not work well (even for simple synthetic examples), partly due to the fact that the deduced conditions are only sufficient, not necessary and there is much room left for improvement. The theory by Candès et al. (2011) specified a choice for λ that is independent of the model parameters, but it applies only for the special case of uniform corruption without noise; moreover, they noticed that other values of λ could achieve better results.

2.3 Conditions for Handling the Three Enemies

We introduce additional assumptions on the data to address the three types of enemies. We denote the sets of exact inliers and outliers by $\mathcal{X}_1$ and $\mathcal{X}_0$ respectively, that is, $\mathcal{X}_1 = \mathcal{X} \cap \mathrm{L}^*$ and $\mathcal{X}_0 = \mathcal{X} \setminus \mathrm{L}^*$. The following two conditions simultaneously address both type 1 and type 3 enemies:

$$\min_{\mathbf{Q}\in\mathbb{H},\mathbf{Q}\mathbf{P}_{\mathrm{L}^{*\perp}}=\mathbf{0}} \sum_{\mathbf{x}\in\mathcal{X}_1} \|\mathbf{Q}\mathbf{x}\| > \sqrt{2} \min_{\mathbf{v}\in\mathrm{L}^{*\perp},\|\mathbf{v}\|=1} \sum_{\mathbf{x}\in\mathcal{X}_0} |\mathbf{v}^T\mathbf{x}|, \tag{6}$$

$$\min_{\mathbf{Q}\in\mathbb{H},\mathbf{Q}\mathbf{P}_{\mathrm{L}^{*\perp}}=\mathbf{0}} \sum_{\mathbf{x}\in\mathcal{X}_1} \|\mathbf{Q}\mathbf{x}\| > \sqrt{2} \max_{\mathbf{v}\in\mathrm{L}^{*},\|\mathbf{v}\|=1} \sum_{\mathbf{x}\in\mathcal{X}_0} |\mathbf{v}^T\mathbf{x}|. \tag{7}$$

A lower bound on the common LHS of both (6) and (7) is designed to avoid type 1 enemies. This common LHS is a weak version of the permeance statistics, which was defined in (3.1) of Lerman et al. (2012) as follows:

$$\mathcal{P}(\mathrm{L}^*) := \min_{\substack{\mathbf{u}\in\mathrm{L}^* \\ \|\mathbf{u}\|=1}} \sum_{\mathbf{x}\in\mathcal{X}_1} |\mathbf{u}^T\mathbf{x}|.$$

Similarly to the permeance statistics, it is zero if and only if all inliers are contained in a proper subspace of L^*. Indeed, if all inliers lie in a subspace $\mathrm{L}' \subset \mathrm{L}^*$, then this common LHS is zero with the minimizer $\mathbf{Q} = \mathbf{P}_{\mathrm{L}'^{\perp} \cap \mathrm{L}^*} / \operatorname{tr}(\mathbf{P}_{\mathrm{L}'^{\perp} \cap \mathrm{L}^*})$. Similarly, if it is zero, then $\mathbf{Q}\mathbf{x} = \mathbf{0}$ for any $\mathbf{x} \in \mathcal{X}_1$ and for some $\mathbf{Q}$ with kernel containing $\mathrm{L}^{*\perp}$. This is only possible when $\mathcal{X}_1$ is contained in a proper subspace of L^*. Similarly to the permeance statistics, if the inliers nicely permeate through L^*, then this common LHS clearly obtain large values.

The upper bounds on the RHS's of (6) and (7) address two complementing type 3 enemies. If $\mathcal{X}_0$ contains few data points of large magnitude, which are orthogonal to L^*, then the RHS of (6) may be too large and (6) may not hold. If on the other hand $\mathcal{X}_0$ contains few data points with large magnitude and a small angle with L^*, then the RHS of (7) will be large so that (7) may not hold. Conditions (6) and (7) thus complete each other.

The RHS of condition (7) is similar to the linear structure statistics (for L^*), which was defined in (3.3) of Lerman et al. (2012). The linear structure statistics uses an l_2 average of dot products instead of the l_1 average used here and was applied in this context to $\mathbb{R}^D$ (instead of L^*) in Lerman et al. (2012). Similarly to the linear structure statistics, the RHS of (7) is large when outliers either have large magnitude or they lie close to a line (so that their combined contribution is similar to an outlier with a very large magnitude as exemplified in Figure 1(d)). The RHS of condition (7) is a very weak analog of the linear structure statics of $\mathrm{L}^{*\perp}$ since it uses a minimum instead of a maximum. There are some significant outliers within $\mathrm{L}^{*\perp}$ that will not be avoided by requiring (7). For example, if the codimension of L^* is larger than 1 and there is a single outlier with an arbitrary large magnitude orthogonal to L^*, then the RHS of (7) is zero.

The next condition avoids type 2 enemies and also significant outliers within $\mathrm{L}^{*\perp}$ (i.e., type 3 enemies) that were not avoided by condition (7). This condition requires that any minimizer of the following oracle problem

$$\hat{\mathbf{Q}}_0 := \underset{\mathbf{Q} \in \mathbb{H}, \mathbf{Q}\mathbf{P}_{\mathrm{L}^*} = \mathbf{0}}{\arg\min} F(\mathbf{Q}) \tag{8}$$

satisfies

$$\operatorname{rank}(\hat{\mathbf{Q}}_0) = D - d. \tag{9}$$

We note that the requirement $\mathbf{Q}\mathbf{P}_{\mathrm{L}^*} = \mathbf{0}$ is equivalent to the condition $\ker(\mathbf{Q}) \supseteq \mathrm{L}^*$ and therefore the rank of the minimizer is at most $D - d$. Enforcing the rank of the minimizer to be exactly $D - d$ restricts the distribution of the projection of $\mathcal{X}$ onto $\mathrm{L}^{*\perp}$. In particular, it avoids its concentration on lower dimensional subspaces and is thus suitable to avoid type 2 enemies. Indeed, if all outliers are sampled from $\tilde{\mathrm{L}} \subset \mathrm{L}^{*\perp}$, then any $\mathbf{Q} \in \mathbb{H}$ with $\ker(\mathbf{Q}) \supset \tilde{\mathrm{L}} + \mathrm{L}^*$ satisfies $F(\mathbf{Q}) = 0$ and therefore it is a minimizer of the oracle problem (4), but it contradicts (9).

We note that this condition also avoids some type 3 enemies, which were not handled by conditions (6) and (7). For example, any $D - d - 1$ outliers with large magnitude orthogonal to L^* will not be excluded by requiring (6) or (7), but will be avoided by (9).

This condition is restrictive though, especially in very high ambient dimensions. Indeed, it does not hold when the number of outliers is smaller than $D - d$ (since then the outliers are sampled from some $\tilde{\mathrm{L}}$ with $\dim(\tilde{\mathrm{L}} \oplus \mathrm{L}^*) < D$). We thus explain in §5.2 and §5.2.1 how to avoid this condition when knowing the dimension. We also suggest in §5.1 some practical

solutions to overcome the corresponding restrictive lower bound on the number of outliers when the dimension is unknown.

Example 1 *We demonstrate the violation of the conditions above for the examples depicted in Figure 1. The actual calculations rely on ideas explained in §2.4.1.*

For the example in Figure 1(a), which represents a type 1 enemy, both conditions (6) *and* (7) *are violated. Indeed, the common LHS of* (6) *and* (7) *is* 5.69*, whereas the RHS of* (6) *is* 8.57 *and the RHS of* (7) *is larger than* 10.02 *(this lower bound is obtained by substituting* $\mathbf{v} = [0, 1, 0]$ *in the RHS of* (7)*; note that* $\mathbf{v}$ *is a unit vector in* L^**).*

For the example in Figure 1(b), which represents a type 2 enemy, condition (9) *is violated. Indeed, we obtained numerically a solution* $\hat{\mathbf{Q}}_0$ *with* $\mathrm{rank}(\hat{\mathbf{Q}}_0) = 1 \neq D - d = 2$ *(one can actually prove in this case that* $\hat{\mathbf{Q}}_0$ *is the projector onto the orthogonal complement of the plane represented by the dashed rectangle).*

For the example in Figure 1(c), which represents a type 3 enemy, both conditions (6) *and* (7) *are violated. Indeed, the common LHS of* (6) *and* (7) *is* 1.56 *and the RHS's of* (6) *and* (7) *are* 5.66 *and* 4.24 *respectively. However, if we normalize all points to lie on the unit circle, then this enemy can be overcome. Indeed, for the normalized data, the common LHS of* (6) *and* (7) *is* 6 *and the RHS's of* (6) *and* (7) *are* 1.13 *and* 0.85 *respectively.*

For the example in Figure 1(d), which also represents a type 3 enemy, both conditions (6) *and* (7) *are violated. Indeed, the LHS of* (6) *and* (7) *are* 5.99 *and the RHS's of* (6) *and* (7) *are* 6.91 *and* 7.02 *respectively.*

2.4 Exact Recovery Under Combinatorial Conditions

We show that the minimizer of (4) solves the exact recovery problem under the above combinatorial conditions.

Theorem 1 *Assume that* $d, D \in \mathbb{N}$*,* $d < D$*,* $\mathcal{X}$ *is a data set in* $\mathbb{R}^D$ *and* $\mathrm{L}^* \in \mathrm{G}(D, d)$*. If conditions* (6)*,* (7) *and* (9) *hold (w.r.t.* $\mathcal{X}$ *and* L^**), then any minimizer of* (4)*,* $\hat{\mathbf{Q}}$*, recovers the subspace* L^* *in the following way:* $\ker(\hat{\mathbf{Q}}) = \mathrm{L}^*$*. If only* (6) *and* (7) *hold, then* $\ker(\hat{\mathbf{Q}}) \supseteq \mathrm{L}^*$*.*

2.4.1 Weaker Alternatives of Conditions (6) and (7)

It is sufficient to guarantee exact recovery by requiring (9) and that for an arbitrarily chosen solution of (8), $\hat{\mathbf{Q}}_0$, the following two conditions are satisfied:

$$\min_{\mathbf{Q} \in \mathbb{H}, \mathbf{Q}\mathbf{P}_{\mathrm{L}^{*\perp}} = \mathbf{0}} \sum_{\mathbf{x} \in \mathcal{X}_1} \|\mathbf{Q}\mathbf{x}\| > \sqrt{2} \left\| \sum_{\mathbf{x} \in \mathcal{X}_0} \hat{\mathbf{Q}}_0 \mathbf{x}\mathbf{x}^T \mathbf{P}_{\mathrm{L}^{*\perp}} / \|\hat{\mathbf{Q}}_0 \mathbf{x}\| \right\| \tag{10}$$

and

$$\min_{\mathbf{Q} \in \mathbb{H}, \mathbf{Q}\mathbf{P}_{\mathrm{L}^{*\perp}} = \mathbf{0}} \sum_{\mathbf{x} \in \mathcal{X}_1} \|\mathbf{Q}\mathbf{x}\| > \sqrt{2} \left\| \sum_{\mathbf{x} \in \mathcal{X}_0} \hat{\mathbf{Q}}_0 \mathbf{x}\mathbf{x}^T \mathbf{P}_{\mathrm{L}^{*}} / \|\hat{\mathbf{Q}}_0 \mathbf{x}\| \right\|. \tag{11}$$

We note that condition (9) guarantees that $\hat{\mathbf{Q}}_0 \mathbf{x} \neq \mathbf{0}$ for all $\mathbf{x} \in \mathcal{X}_0$ and thus the RHS's of (10) and (11) are well-defined. We prove this statement in (7.3).

We note that conditions (10) and (11) can be verified when $\mathcal{X}_0$, $\mathcal{X}_1$ and L^* are known (unlike (6) and (7)), where $\hat{\mathbf{Q}}_0$ can be found by Algorithm 2. Furthermore, (10) and (11) are weaker than (6) and (7), though they are more technically involved and harder to motivate.

In order to demonstrate the near-tightness of (10) and (11), we formulate the following necessary conditions for the recovery of L^* as $\ker(\hat{\mathbf{Q}})$ (see the idea of their justification at the end of §7.3): For an arbitrarily chosen solution of (8), $\hat{\mathbf{Q}}_0$:

$$\min_{\mathbf{Q}\in\mathbb{H},\mathbf{Q}\mathbf{P}_{\mathrm{L}^{*\perp}}=\mathbf{0}} \sum_{\mathbf{x}\in\mathcal{X}_1} \|\mathbf{Q}\mathbf{x}\| \geq \| \sum_{\mathbf{x}\in\mathcal{X}_1} \hat{\mathbf{Q}}_0\mathbf{x}\mathbf{x}^T\mathbf{P}_{\mathrm{L}^{*\perp}}/\|\hat{\mathbf{Q}}_0\mathbf{x}\|\| \tag{12}$$

and

$$\sum_{\mathbf{x}\in\mathcal{X}_1} \|\mathbf{Q}(\tilde{\mathbf{P}}_{\mathrm{L}^*}\mathbf{x})\| \geq \sum_{\mathbf{x}\in\mathcal{X}_0} \left\langle \mathbf{Q}, \tilde{\mathbf{P}}^T_{\mathrm{L}^{*\perp}}\hat{\mathbf{Q}}_0\mathbf{x}\mathbf{x}^T\tilde{\mathbf{P}}_{\mathrm{L}^*}/\|\hat{\mathbf{Q}}_0\mathbf{x}\| \right\rangle_F \text{ for any } \mathbf{Q}\in\mathbb{R}^{(D-d)\times d}, \tag{13}$$

where for matrices $\mathbf{A}, \mathbf{B}\in\mathbb{R}^{k\times l}$: $\langle\mathbf{A},\mathbf{B}\rangle_F = \mathrm{tr}(A\,B^T)$ is the Frobenius dot product. Indeed, conditions (12) and (13) are close to conditions (10) and (11). In particular, (12) and (10) are only different by the constant factor $\sqrt{2}$, that is, (10) is practically tight.

2.5 Uniqueness of the Minimizer

We recall that Theorem 1 implies that if (6), (7) and (9) hold, then $\ker(\hat{\mathbf{Q}})$ is unique. Here we guarantee the uniqueness of $\hat{\mathbf{Q}}$ (which is required in §2.4.1, §2.7, §2.8 and §4.2) independently of the exact subspace recovery problem.

Theorem 2 *If the following condition holds:*

$$\{\mathcal{X}\cap \mathrm{L}_1\}\cup\{\mathcal{X}\cap \mathrm{L}_2\} \neq \mathcal{X} \text{ for all } (D-1)\text{-dimensional subspaces } \mathrm{L}_1, \mathrm{L}_2 \subset \mathbb{R}^D, \tag{14}$$

then $F(\mathbf{Q})$ is a strictly convex function on $\mathbb{H}$.

2.6 Exact Recovery under Probabilistic Models

We show that our conditions for exact recovery (or the main two of them) and our condition for uniqueness of the minimizer $\hat{\mathbf{Q}}$ hold with high probability under basic probabilistic models. Such a probabilistic theory is cleaner when the outliers are sampled from a spherically symmetric distribution as we carefully demonstrate in §2.6.1 (with two different models). The problem is that when the outliers are spherically symmetric then various non-robust algorithms (such as PCA) can asymptotically approach exact recovery and nearly recover the underlying subspace with sufficiently large sample. We thus also show in §2.6.2 how the theory in §2.6.1 can be slightly modified to establish exact recovery of the GMS algorithm in an asymmetric case, where PCA cannot even nearly recover the underlying subspace.

2.6.1 Cases with Spherically Symmetric Distributions of Outliers

First we assume a more general probabilistic model. We say that μ on $\mathbb{R}^D$ is an Outliers-Inliers Mixture (OIM) measure (w.r.t. the fixed subspace $\mathrm{L}^*\in\mathrm{G}(D,d)$) if $\mu = \alpha_0\mu_0 + \alpha_1\mu_1$, where $\alpha_0,\ \alpha_1 > 0,\ \alpha_0+\alpha_1 = 1$, μ_1 is a sub-Gaussian probability measure and μ_0 is a sub-Gaussian probability measure on $\mathbb{R}^D$ (representing outliers) that can be decomposed to

a product of two independent measures $\mu_0 = \mu_{0,\mathrm{L}^*} \times \mu_{0,\mathrm{L}^{*\perp}}$ such that the supports of μ_{0,L^*} and $\mu_{0,\mathrm{L}^{*\perp}}$ are L^* and $\mathrm{L}^{*\perp}$ respectively, and $\mu_{0,\mathrm{L}^{*\perp}}$ is spherically symmetric with respect to rotations within $\mathrm{L}^{*\perp}$.

To provide cleaner probabilistic estimates, we also invoke the needle-haystack model of Lerman et al. (2012). It assumes that both μ_0 and μ_1 are the Gaussian distributions: $\mu_0 = N(\mathbf{0}, \sigma_0^2 \mathbf{I}/D)$ and $\mu_1 = N(\mathbf{0}, \sigma_1^2 \mathbf{P}_{\mathrm{L}^*} \mathbf{P}_{\mathrm{L}^*}^T/d)$ (the factors $1/D$ and $1/d$ normalize the magnitude of outliers and inliers respectively so that their norms are comparable). While Lerman et al. (2012) assume a fixed number of outliers and inliers independently sampled from μ_0 and μ_1 respectively, here we independently sample from the mixture measure $\mu = \alpha_0\mu_0 + \alpha_1\mu_1$; we refer to μ as a needle-haystack mixture measure.

In order to prove exact recovery under any of these models, one needs to restrict the fraction of inliers per outliers (or equivalently, the ratio α_1/α_0). We refer to this ratio as SNR (signal to noise ratio) since we may view the inliers as the pure signal and the outliers as some sort of "noise". For the needle-haystack model we require the following SNR, which is similar to the one of Lerman et al. (2012):

$$\frac{\alpha_1}{\alpha_0} > 4\,\frac{\sigma_0}{\sigma_1}\,\frac{d}{\sqrt{(D-d)D}}. \tag{15}$$

We later explain how to get rid of the term σ_1/σ_0. For the OIM model we assume the following more general condition:

$$\alpha_1 \min_{\mathbf{Q}\in\mathbb{H},\mathbf{Q}\mathbf{P}_{\mathrm{L}^{*\perp}}=\mathbf{0}} \int \|\mathbf{Q}\mathbf{x}\|\,\mathrm{d}\mu_1(\mathbf{x}) > 2\sqrt{2}\frac{\alpha_0}{D-d}\int \|\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\|\,\mathrm{d}\mu_0(\mathbf{x}). \tag{16}$$

Under the needle-haystack model, this condition is a weaker version of (15). That is,

Lemma 3 *If μ is a needle-haystack mixture measure, then* (15) *implies* (16).

For i.i.d. samples from an OIM measure satisfying (16), we can establish our modified conditions of unique exact recovery (i.e., (10), (11) and (9)) with overwhelming probability in the following way (we also guarantee the uniqueness of the minimizer $\hat{\mathbf{Q}}$).

Theorem 4 *If $\mathcal{X}$ is an i.i.d. sample from an OIM measure μ satisfying* (16)*, then conditions* (10)*,* (11)*, and* (9) *hold with probability $1 - C\exp(-N/C)$, where C is a constant depending on μ and its parameters. Moreover,* (14) *holds with probability* 1 *if there are at least $2D-1$ outliers (i.e., the number of points in $\mathcal{X}\setminus \mathrm{L}^*$ is at least $2D-1$).*

Under the needle-haystack model, the SNR established by Theorem 4 is comparable to the best SNR among other convex exact recovery algorithms (this is later clarified in Table 1). However, the probabilistic estimate under which this SNR holds is rather loose and thus its underlying constant C is not specified. Indeed, the proof of Theorem 4 uses ϵ-nets and union-bounds arguments, which are often not useful for deriving tight probabilistic estimates (see, e.g., Mendelson 2003, page 18). One may thus view Theorem 4 as a near-asymptotic statement.

The statement of Theorem 4 does not contradict our previous observation that the number of outliers should be larger than at least $D-d$. Indeed, the constant C is sufficiently

large so that the corresponding probability is negative when the number of outliers is smaller than $D - d$.

In the next theorem we assume only a needle-haystack model and thus we can provide a stronger probabilistic estimate based on the concentration of measure phenomenon (our proof follows directly Lerman et al., 2012). However, the SNR is worse than the one in Theorem 4 by a factor of order $\sqrt{D-d}$. This is because we are unable to estimate $\hat{\mathbf{Q}}_0$ of (8) by concentration of measure. Similarly, in this theorem we do not estimate the probability of (9) (which also involves $\hat{\mathbf{Q}}_0$). Nevertheless, we observed in experiments that (9) holds with high probability for $N_0 = 2(D-d)$ and the probability seems to go to 1 as $N_0 = 2(D-d)$ and $D - d \to \infty$. Moreover, one of the algorithms proposed below (EGMS) does not require condition (9).

Theorem 5 *If $\mathcal{X}$ is an i.i.d. sample of size N from a needle-haystack mixture measure μ and if*

$$\frac{\alpha_1}{\alpha_0} > \frac{\sigma_0}{\sigma_1} \frac{\sqrt{2/\pi} - 1/4 - 1/10}{\sqrt{2/\pi} + 1/4 + 1/10} \sqrt{\frac{d^2}{D}} \tag{17}$$

and

$$N > 64 \max(2d/\alpha_1, 2d/\alpha_0, 2(D-d)/\alpha_0), \tag{18}$$

then (6) *and* (7) *hold with probability* $1 - e^{-\alpha_1^2 N/2} - 2e^{-\alpha_0^2 N/2} - e^{-\alpha_1 N/800} - e^{-\alpha_0 N/800}$.

In Table 1 we present the theoretical asymptotic SNRs for exact recovery of some recent algorithms. We assume the needle-haystack model with fixed d, D, α_0, α_1, σ_0 and σ_1 and $N \to \infty$. Let us clarify these results. We first remark that the pure SNR of the High-dimensional Robust PCA (HR-PCA) algorithm of Xu et al. (2010a) approaches infinity (see Remark 3 of Xu et al. 2010a). However, as we explained earlier the violation of exact recovery does not necessarily imply non-robustness of the estimator as it may nearly recover the subspace. Indeed, Xu et al. (2010a) show that if (for simplicity) $\sigma_0 = \sigma_1$ and the SNR is greater than 1, then the subspace estimated by HR-PCA is a good approximation in the following sense: there exists a constant $c > 0$ such that for the inliers set $\mathcal{X}_0$ and the estimated subspace L: $\sum_{\mathbf{x} \in \mathcal{X}_0} \|\mathbf{P}_{\mathrm{L}}\mathbf{x}\|_2^2 > c \sum_{\mathbf{x} \in \mathcal{X}_0} \|\mathbf{x}\|_2^2$ (see Remark 4 of Xu et al. (2010a)). We thus use the notation: SNR(HR-PCA) "$\gtrapprox$" 1 (see Table 1 with appropriate scales of σ_0 and σ_1). Xu et al. (2012) established the SNR for their Outlier Pursuit (OP) algorithm (equivalently the Low-Leverage Decomposition (LLD) of McCoy and Tropp 2011) in Theorem 1 of their work. Their analysis assumes a deterministic condition, but it is possible to show that this condition is asymptotically valid under the needle-haystack model. Lerman et al. (2012) established w.h.p. the SNR of the REAPER algorithm in Theorem 1 of their work (for simplicity of their expressions they assumed that $d \leq (D-1)/2$). Zhang (2012) established the SNR for Tyler's M-Estimator (TME) in Theorem 1 of his work. His result is deterministic, but it is easy to show that the deterministic condition holds with probability 1 under the needle-haystack model. Hardt and Moitra (2013) proposed randomized and deterministic robust recovery algorithms, RF (or RandomizedFind) and DRF (or DERandomizedFind) respectively, and proved that they obtained the same SNR as in Zhang (2012) under a similar (slightly weaker) combinatorial condition (they only guarantee polynomial time, where Zhang, 2012 specifies a complexity similar to that of

HR-PCA	LLD (OP)	$\hat{\mathrm{L}} := \ker(\hat{\mathbf{Q}})$	REAPER ($d \leq (D-1)/2$)	TME & D/RF
$\frac{\sigma_1\alpha_1}{\sigma_0\alpha_0}$ "$\gtrapprox$" 1	$\frac{\alpha_1}{\alpha_0} \geq \frac{121d}{9}$	$\frac{\alpha_1}{\alpha_0} > 4\frac{\sigma_0}{\sigma_1}\frac{d}{\sqrt{(D-d)D}}$	$\frac{\alpha_1}{\alpha_0} > \frac{\sigma_0}{\sigma_1}\left(C_1\frac{d}{D} - \frac{d}{C_2\alpha_1}\right)$	$\frac{\alpha_1}{\alpha_0} > \frac{d}{D-d}$

Table 1: Theoretical SNR (lowest bound on α_1/α_0) for exact recovery when $N \to \infty$

GMS). We remark that both Zhang (2012) and Hardt and Moitra (2013) appeared after the submission of this manuscript.

The asymptotic SNR of the minimization proposed in this paper is of the same order as that of the REAPER algorithm (which was established for $d \leq (D-1)/2$) and both of them are better than that of the HR-PCA algorithm. The asymptotic SNRs of OP, TME, RF and DRF are independent of σ_1 and σ_0. However, by normalizing all data points to the unit sphere, we may assume that $\sigma_1 = \sigma_0$ in all other algorithms and treat them equally (see Lerman et al., 2012). In this case, the SNR of OP is significantly worse than that of the minimization proposed in here, especially when $d \ll D$ (it is also worse than the weaker SNR specified in (17)). When $d \ll D$, the SNR of TME, RF and DRF is of the same order as the asymptotic SNR of our formulation. However, when d is very close to D, the SNR of our formulation is better than the SNR of TME by a factor of $\sqrt{D}$. We question whether a better asymptotic rate than the one of GMS and REAPER can be obtained by a convex algorithm for robust subspace recovery for the needle-haystack model. Hardt and Moitra (2013) showed that it is small set expansion hard for any algorithm to obtain better SNR than theirs for all scenarios satisfying their combinatorial condition.

We note though that there are non-convex methods for removing outliers with asymptotically zero SNRs. Such SNRs are valid only for the noiseless case and may be differently formulated for detecting the hidden low-dimensional structure among uniform outliers. For example, Arias-Castro et al. (2005) proved that the scan statistics may detect points sampled uniformly from a d-dimensional graph in $\mathbb{R}^D$ of an m-differentiable function among uniform outliers in a cube in $\mathbb{R}^D$ with SNR of order $O(N^{-m(D-d)/(d+m(D-d))})$. Arias-Castro et al. (2011) used higher order spectral clustering affinities to remove outliers and thus detect differentiable surfaces (or certain unions of such surfaces) among uniform outliers with similar SNR to that of the scan statistics. Soltanolkotabi and Candès (2012) removed outliers with "large dictionary coefficients" and showed that this detection works well for outliers uniform in S^{D-1}, inliers uniform in $S^{D-1} \cap \mathrm{L}^*$ and SNR at least $\frac{d}{D} \cdot ((\frac{\alpha_1 N-1}{d})^{\frac{cD}{d}-1} - 1)^{-1}$ (where α_1 is the fraction of inliers) as long as $N < e^{c\sqrt{D}}/D$. For fixed D and d and sufficiently large N, this SNR, which depends on N, can be arbitrarily small. Furthermore, Lerman and Zhang (2010) showed that the global minimizer of (2) (that we relax in this paper so that the minimization is convex) can in theory recover the subspace with asymptotically zero SNR. They also showed that the underlying subspace is a local minimum of (2) with SNR of order $\omega(1/\sqrt{N})$. However, these non-convex procedures do not have efficient or sufficiently fast implementations for subspace recovery. Furthermore, their impressive theoretical estimates often break down in the presence of noise. Indeed, in the noisy case their near-recovery is not better than the one stated for GMS in Theorem 6 (see, e.g., (16) and (17) of Arias-Castro et al. (2011) or Theorem 1.2 of Lerman and Zhang (2010)). On the other hand, in view of Coudron and Lerman (2012) we may obtain significantly better

asymptotic SNR for GMS when the noise is symmetrically distributed with respect to the underlying subspace.

2.6.2 A Special Case with Asymmetric Outliers

In the case of spherically symmetric outliers, PCA cannot exactly recover the underlying subspace, but it can asymptotically recover it (see, e.g., Lerman and Zhang, 2010). In particular, with sufficiently large sample with spherically symmetric outliers, PCA nearly recovers the underlying subspace. We thus slightly modify the two models of §2.6.1 so that the distribution of outliers is asymmetric and show that our combinatorial conditions for exact recovery still hold (with overwhelming probability). On the other hand, the subspace recovered by PCA, when sampling data from these models, is sufficiently far from the underlying subspace for any given sample size.

We first generalize Theorem 5 under a generalized needle-haystack model: Let $\mu = \alpha_0\mu_0 + \alpha_1\mu_1$, $\mu_0 = N(\mathbf{0}, \mathbf{\Sigma}_0/D)$, where $\mathbf{\Sigma}_0$ is an arbitrary positive definite matrix (not necessarily a scalar matrix as before), and as before $\mu_1 = N(\mathbf{0}, \sigma_1^2\mathbf{P}_{\mathrm{L}^*}\mathbf{P}_{\mathrm{L}^*}^T/d)$. We claim that Theorem 5 still holds in this case if we replace σ_0 in the RHS of (17) with $\sqrt{\lambda_{\max}(\mathbf{\Sigma}_0)}$, where $\lambda_{\max}(\mathbf{\Sigma}_0)$ denotes the largest eigenvalue of $\mathbf{\Sigma}_0$ (see justification in §7.6.1).

In order to generalize Theorem 4 for asymmetric outliers, we assume that the outlier component μ_0 of the OIM measure μ is a sub-Gaussian distribution with an arbitrary positive definite covariance matrix $\mathbf{\Sigma}_0$. Following Coudron and Lerman (2012), we define the expected version of F, F_I, and its oracle minimizer, $\hat{\mathbf{Q}}_I$, which is analogous to (8) (the subscript I indicates integral):

$$F_I(\mathbf{Q}) = \int \|\mathbf{Q}\mathbf{x}\| \, \mathrm{d}\mu(x) \tag{19}$$

and

$$\hat{\mathbf{Q}}_I = \underset{\mathbf{Q}\in\mathbb{H},\mathbf{Q}\mathbf{P}_{\mathrm{L}^*}=\mathbf{0}}{\arg\min} F_I(\mathbf{Q}). \tag{20}$$

We assume that $\hat{\mathbf{Q}}_I$ is the unique minimizer in (20) (we remark that the two-subspaces criterion in (25) for the projection of μ onto $L^{*\perp}$ implies this assumption). Under these assumptions Theorem 4 still holds if we multiply the RHS of (16) by the ratio between the largest eigenvalue of $\mathbf{P}_{\mathrm{L}^{*\perp}}\hat{\mathbf{Q}}_I\mathbf{P}_{\mathrm{L}^{*\perp}}$ and the $(D-d)$th eigenvalue of $\mathbf{P}_{\mathrm{L}^{*\perp}}\hat{\mathbf{Q}}_I\mathbf{P}_{\mathrm{L}^{*\perp}}$ (see justification in §7.5.1).

2.7 Near Subspace Recovery for Noisy Samples

We show that in the case of sufficiently small additive noise (i.e., the inliers do not lie exactly on the subspace L^* but close to it), the GMS algorithm nearly recovers the underlying subspace.

We use the following notation: $\|\mathbf{A}\|_F$ and $\|\mathbf{A}\|$ denote the Frobenius and spectral norms of $\mathbf{A} \in \mathbb{R}^{k\times l}$ respectively. Furthermore, $\mathbb{H}_1$ denotes the set of all positive semidefinite matrices in $\mathbb{H}$, that is, $\mathbb{H}_1 = \{\mathbf{Q} \in \mathbb{H} : \mathbf{Q} \succcurlyeq 0\}$. We also define the following two constants

$$\gamma_0 = \frac{1}{N} \min_{\mathbf{Q}\in\mathbb{H}_1,\|\mathbf{\Delta}\|_F=1,\mathrm{tr}(\mathbf{\Delta})=0} \sum_{i=1}^{N} \frac{\|\mathbf{\Delta}\mathbf{x}_i\|^2\|\mathbf{Q}\mathbf{x}_i\|^2 - (\mathbf{x}_i^T\mathbf{\Delta}\mathbf{Q}\mathbf{x}_i)^2}{\|\mathbf{Q}\mathbf{x}_i\|^3}, \tag{21}$$

and

$$\gamma_0' = \frac{1}{N} \min_{\mathbf{Q}\in\mathbb{H}_1, \|\mathbf{\Delta}\|=1, \mathrm{tr}(\mathbf{\Delta})=0} \sum_{i=1}^{N} \frac{\|\mathbf{\Delta}\mathbf{x}_i\|^2\|\mathbf{Q}\mathbf{x}_i\|^2 - (\mathbf{x}_i^T\mathbf{\Delta}\mathbf{Q}\mathbf{x}_i)^2}{\|\mathbf{Q}\mathbf{x}_i\|^3}. \tag{22}$$

The sum in the RHS's of (21) and (22) is the following second directional derivative: $\frac{\mathrm{d}^2}{\mathrm{d}t^2}F(\mathbf{Q}+t\mathbf{\Delta})$; when $\mathbf{Q}\mathbf{x}_i = 0$, its ith term can be set to 0. It is interesting to note that both (21) and (22) express the Restricted Strong Convexity (RSC) parameter γ_l of Agarwal et al. (2012b, Definition 1), where their notation translates into ours as follows: $\mathcal{L}_n(\mathbf{Q}) := F(\mathbf{Q})/N$, $\tau_l := 0$, $\Omega' := \mathbb{H}_1$ and $\theta - \theta' := \Delta$. The difference between γ_0 and γ_0' of (21) and (22) is due to the choice of either the Frobenius or the spectral norms respectively for measuring the size of $\theta - \theta'$.

Using this notation, we formulate our noise perturbation result as follows.

Theorem 6 *Assume that $\{\epsilon_i\}_{i=1}^N$ is a set of positive numbers, $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ and $\tilde{\mathcal{X}} = \{\tilde{\mathbf{x}}_i\}_{i=1}^N$ are two data sets such that $\|\tilde{\mathbf{x}}_i - \mathbf{x}_i\| \le \epsilon_i \ \ \forall 1 \le i \le N$ and $\mathcal{X}$ satisfies (14). Let $F_{\mathcal{X}}(\mathbf{Q})$ and $F_{\tilde{\mathcal{X}}}(\mathbf{Q})$ denote the corresponding versions of $F(\mathbf{Q})$ w.r.t. the sets $\mathcal{X}$ and $\tilde{\mathcal{X}}$ and let $\hat{\mathbf{Q}}$ and $\tilde{\mathbf{Q}}$ denote their respective minimizers. Then we have*

$$\|\tilde{\mathbf{Q}} - \hat{\mathbf{Q}}\|_F < \sqrt{2\sum_{i=1}^{N}\epsilon_i/(N\gamma_0)} \quad \textit{and} \quad \|\tilde{\mathbf{Q}} - \hat{\mathbf{Q}}\| < \sqrt{2\sum_{i=1}^{N}\epsilon_i/(N\gamma_0')}. \tag{23}$$

Moreover, if $\tilde{\mathrm{L}}$ and $\hat{\mathrm{L}}$ are the subspaces spanned by the bottom d eigenvectors of $\tilde{\mathbf{Q}}$ and $\hat{\mathbf{Q}}$ respectively and ν_{D-d} is the $(D-d)$th eigengap of $\hat{\mathbf{Q}}$, then

$$\|\mathbf{P}_{\hat{\mathrm{L}}} - \mathbf{P}_{\tilde{\mathrm{L}}}\|_F \le \frac{2\sqrt{2\sum_{i=1}^{N}\epsilon_i/(N\gamma_0)}}{\nu_{D-d}} \quad \textit{and} \quad \|\mathbf{P}_{\hat{\mathrm{L}}} - \mathbf{P}_{\tilde{\mathrm{L}}}\| \le \frac{2\sqrt{2\sum_{i=1}^{N}\epsilon_i/(N\gamma_0')}}{\nu_{D-d}}. \tag{24}$$

We note that if $\mathcal{X}$ and $\tilde{\mathcal{X}}$ satisfy the conditions of Theorem 6, then given the perturbed data set $\tilde{\mathcal{X}}$ and the dimension d, Theorem 6 guarantees that GMS nearly recovers L^*. More interestingly, the theorem also implies that we may properly estimate the dimension of the underlying subspace in this case (we explain this in details in §7.7.1). Such dimension estimation is demonstrated later in Figure 2.

Theorem 6 is a perturbation result in the spirit of the stability analysis by Candès et al. (2006) and Xu et al. (2012, Theorem 2). In order to observe that the statement of Theorem 6 is comparable to that of Theorem 2 of Xu et al. (2012), we note that asymptotically the bounds on the recovery errors in (23) and (24) depend only on the empirical mean of $\{\epsilon_i\}_{i=1}^N$ and do not grow with N. To clarify this point we formulate the following proposition.

Proposition 7 *If $\mathcal{X}$ is i.i.d. sampled from a bounded distribution μ and*

$$\mu(\mathrm{L}_1) + \mu(\mathrm{L}_2) < 1 \ \ \textit{for any two } D-1\textit{-dimensional subspaces } \mathrm{L}_1 \textit{ and } \mathrm{L}_2, \tag{25}$$

then there exist constants $c_0(\mu) > 0$ and $c_0'(\mu) > 0$ depending on μ such that

$$\liminf_{N\to\infty} \gamma_0(\mathcal{X}) \ge c_0(\mu) \quad \textit{and} \quad \liminf_{N\to\infty} \gamma_0'(\mathcal{X}) \ge c_0'(\mu) \ \textit{almost surely.} \tag{26}$$

If (25) is strengthened so that $\mu(\mathrm{L}_1) + \mu(\mathrm{L}_2)$ is sufficiently smaller than 1, then it can be noticed empirically that $c_0(\mu)$ and $c_0'(\mu)$ are sufficiently larger than zero.

Nevertheless, the stability theory of Candès et al. (2006), Xu et al. (2012) and this section is not optimal. Stronger stability results require nontrivial analysis and we leave it to a possible future work. We comment though on some of the deficiencies of our stability theory and their possible improvements.

We first note that the bounds in Theorem 6 are generally not optimal. Indeed, if $\epsilon_i = O(\epsilon)$ for all $1 \leq i \leq N$, then the error bounds in Theorem 6 are $O(\sqrt{\epsilon})$, whereas we empirically noticed that these error bounds are $O(\epsilon)$. In §7.7.2 we sketch a proof for this empirical observation when ϵ is sufficiently small and $\mathrm{rank}(\hat{\mathbf{Q}}) = D$.

The dependence of the error on D, which follows from the dependence of γ_0 and γ_0' on D, is a difficult problem and strongly depends on the underlying distribution of $\mathcal{X}$ and of the noise. For example, in the very special case where the set $\mathcal{X}$ is sampled from a subspace $L_0 \subset \mathbb{R}^D$ of dimension $D_0 < D$, and the noise distribution is such that $\tilde{\mathcal{X}}$ also lies in L_0, then practically we are performing GMS over $P_{L_0}(\mathcal{X})$ and $P_{L_0}(\tilde{\mathcal{X}})$, and the bound in (23) would depend on D_0 instead of D.

Coudron and Lerman (2012) suggested a stronger perturbation analysis and also remarked on the dependence of the error on D in a very special scenario.

2.8 Near Subspace Recovery for Regularized Minimization

For our practical algorithm it is advantageous to regularize the function F as follows (see Theorems 11 and 12 below):

$$F_\delta(\mathbf{Q}) := \sum_{i=1,\|\mathbf{Q}\mathbf{x}_i\|\geq\delta}^{N} \|\mathbf{Q}\mathbf{x}_i\| + \sum_{i=1,\|\mathbf{Q}\mathbf{x}_i\|<\delta}^{N} \left(\frac{\|\mathbf{Q}\mathbf{x}_i\|^2}{2\delta} + \frac{\delta}{2} \right).$$

We remark that other convex algorithms (Candès et al., 2011; Xu et al., 2012; McCoy and Tropp, 2011) also regularize their objective function by adding the term $\delta\|\mathbf{X} - \mathbf{L} - \mathbf{O}\|_F^2$. However, their proofs are not formulated for this regularization.

In order to address the regularization in our case and conclude that the GMS algorithm nearly recovers L^* for the regularized objective function, we adopt a similar perturbation procedure as in §2.7. We denote by $\hat{\mathbf{Q}}_\delta$ and $\hat{\mathbf{Q}}$ the minimizers of $F_\delta(\mathbf{Q})$ and $F(\mathbf{Q})$ in $\mathbb{H}$ respectively. Furthermore, let $\hat{\mathrm{L}}_\delta$ and $\hat{\mathrm{L}}$ denote the subspaces recovered by the bottom d eigenvectors of $\hat{\mathbf{Q}}_\delta$ and $\hat{\mathbf{Q}}$ respectively. Using the constants ν_{D-d} and γ_0 of Theorem 6, the difference between the two minimizers and subspaces can be controlled as follows.

Theorem 8 *If $\mathcal{X}$ is a data set satisfying* (14)*, then*

$$\|\hat{\mathbf{Q}}_\delta - \hat{\mathbf{Q}}\|_F < \sqrt{\delta/2\gamma_0}$$

and

$$\|\mathbf{P}_{\hat{\mathrm{L}}_\delta} - \mathbf{P}_{\hat{\mathrm{L}}}\|_F \leq \frac{2\sqrt{\delta/2\gamma_0}}{\nu_{D-d}}. \tag{27}$$

3. Understanding Our M Estimator: Interpretation and Formal Similarities with Other M Estimators

We highlight the formal similarity of our M-estimator with a common M-estimator and with Tyler's M-estimator in §3.1 and §3.2 respectively. We also show that in view of the standard assumptions on the algorithm for computing the common M-estimator, it may fail in exactly recovering the underlying subspace (see §3.1.1). At last, in §3.3 we interpret our M-estimator as a robust inverse covariance estimator.

3.1 Formal Similarity with the Common M-estimator for Robust Covariance Estimation

A well-known robust M-estimator for the $\mathbf{0}$-centered covariance matrix (Maronna, 1976; Huber and Ronchetti, 2009; Maronna et al., 2006) minimizes the following function over all $D \times D$ positive definite matrices (for some choices of a function ρ)

$$L(\mathbf{A}) = \sum_{i=1}^{N} \rho(\mathbf{x}_i^T \mathbf{A}^{-1} \mathbf{x}_i) - \frac{N}{2} \log(\det(\mathbf{A}^{-1})). \tag{28}$$

The image of the estimated covariance is clearly an estimator to the underlying subspace L^*.

If we set $\rho(x) = \sqrt{x}$ and $\mathbf{A}^{-1} = \mathbf{Q}^2$ then the objective function $L(\mathbf{A})$ in (28) is $\sum_{i=1}^{N} \|\mathbf{Q}\mathbf{x}_i\| - N \log(\det(\mathbf{Q}))$. This energy function is formally similar to our energy function. Indeed, using Lagrangian formulation, the minimizer $\hat{\mathbf{Q}}$ in (4) is also the minimizer of $\sum_{i=1}^{N} \|\mathbf{Q}\mathbf{x}_i\| - \lambda \operatorname{tr}(\mathbf{Q})$ among all $D \times D$ symmetric matrices (or equivalently nonnegative symmetric matrices) for some $\lambda > 0$ (the parameter λ only scales the minimizer and does not effect the recovered subspace). Therefore, the two objective functions differ by their second terms. In the common M-estimator (with $\rho(x) = \sqrt{x}$ and $\mathbf{A}^{-1} = \mathbf{Q}^2$) it is $\log(\det(\mathbf{Q}))$, or equivalently, $\operatorname{tr}(\log(\mathbf{Q}))$, where in our M-estimator, it is $\operatorname{tr}(\mathbf{Q})$.

3.1.1 Problems with Exact Recovery by the Common M-estimator

The common M-estimator is designed for robust covariance estimation, however, we show here that in general it cannot exactly recover the underlying subspace. To make this statement more precise we recall the following uniqueness and existence conditions for the minimizer of (28), which were established by Kent and Tyler (1991): 1) $u = 2\rho'$ is positive, continuous and non-increasing. 2) Condition M: $u(x)x$ is strictly increasing. 3) Condition D_0: For any linear subspace L: $|\mathcal{X} \cap \mathrm{L}|/N < 1 - (D - \dim(\mathrm{L}))/\lim_{x\to\infty} xu(x)$. The following Theorem 9 shows that the uniqueness and existence conditions of the common M-estimator are incompatible with exact recovery.

Theorem 9 *Assume that $d, D \in \mathbb{N}$, $d < D$, $\mathcal{X}$ is a data set in $\mathbb{R}^D$ and $\mathrm{L}^* \in \mathrm{G}(D, d)$ and let $\hat{\mathbf{A}}$ be the minimizer of (28). If conditions M and D_0 hold, then $\operatorname{Im}(\hat{\mathbf{A}}) \neq \mathrm{L}^*$.*

For symmetric outliers (as the ones of §2.6.1) the common M-estimator can still asymptotically achieve exact recovery (similarly to PCA). However, for many scenarios of asymmetric outliers, in particular, the one of §2.6.2, the subspace recovered by the common M-estimator is sufficiently far from the underlying subspace for any given sample size.

We remark that Tyler's M-estimator (Tyler, 1987) can still recover the subspace exactly. This estimator uses $\rho(x) = D\log(x)/2$ in (28) and adds an additional assumption $\mathrm{tr}(\mathbf{A}) = 1$. Zhang (2012) recently showed that this M-estimator satisfies $\mathrm{Im}(\hat{\mathbf{A}}) = \mathrm{L}^*$. However, it does not belong to the class of estimators of Kent and Tyler (1991) addressed by Theorem 9 (it requires that $\mathrm{tr}(\mathbf{A}) = 1$, otherwise it has multiple minimizers; it also does not satisfy condition M).

3.2 Formal Similarity with Tyler's M-Estimator

We show here that the algorithms for our estimator and Tyler's M-estimator (Tyler, 1987) are formally similar. Following Tyler (1987), we write the iterative algorithm for the Tyler's M-estimator for robust covariance estimation as follows:

$$\mathbf{\Sigma}_{n+1} = \sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\mathbf{x}_i^T \mathbf{\Sigma}_n^{-1} \mathbf{x}_i} \Big/ \mathrm{tr}\left(\sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\mathbf{x}_i^T \mathbf{\Sigma}_n^{-1} \mathbf{x}_i} \right). \tag{29}$$

The unregularized iterative algorithm for GMS is later described in (38). Let us formally substitute $\mathbf{\Sigma} = \mathbf{Q}^{-1}/\mathrm{tr}(\mathbf{Q}^{-1})$ in (38); in view of the later discussion of 3.3, $\mathbf{\Sigma}$ (if exists) can be interpreted as a robust estimator for the covariance matrix (whose top d eigenvectors span the estimated subspace). Then an unregularized version for GMS can be formally written as

$$\mathbf{\Sigma}_{n+1} = \sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\|\mathbf{\Sigma}_n^{-1} \mathbf{x}_i\|} \Big/ \mathrm{tr}\left(\sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\|\mathbf{\Sigma}_n^{-1} \mathbf{x}_i\|} \right). \tag{30}$$

Clearly, (30) is obtained from (29) by replacing $\mathbf{x}_i^T \mathbf{\Sigma}_n^{-1} \mathbf{x}_i$ with $\|\mathbf{\Sigma}_n^{-1} \mathbf{x}_i\| \equiv \sqrt{\mathbf{x}_i^T \mathbf{\Sigma}_n^{-2} \mathbf{x}_i}$.

3.3 Interpretation of $\hat{\mathbf{Q}}$ as Robust Inverse Covariance Estimator

The total least squares subspace approximation is practically the minimization over $\mathrm{L} \in \mathrm{G}(D, d)$ of the function

$$\sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{P}_{\mathrm{L}} \mathbf{x}_i\|^2 \equiv \sum_{i=1}^{N} \|\mathbf{P}_{\mathrm{L}^\perp} \mathbf{x}_i\|^2 . \tag{31}$$

Its solution is obtained by the span of the top d right vectors of the data matrix $\mathbf{X}$ (whose rows are the data points in $\mathcal{X}$), or equivalently, the top d eigenvectors of the covariance matrix $\mathbf{X}^T\mathbf{X}$. The convex relaxation used in (31) can be also applied to (31) to obtain the following convex minimization problem:

$$\hat{\mathbf{Q}}_2 := \operatorname*{arg\,min}_{\mathbf{Q} \in \mathbb{H}} \sum_{i=1}^{N} \|\mathbf{Q}\mathbf{x}_i\|^2. \tag{32}$$

The "relaxed" total least squares subspace is then obtained by the span of the bottom d eigenvectors of $\hat{\mathbf{Q}}$.

We show here that $\hat{\mathbf{Q}}_2$ coincides with a scaled version of the empirical inverse covariance matrix. This clearly imply that the "relaxed" total least squared subspace coincides with the original one (as the bottom eigenvectors of the inverse empirical covariance are the

top eigenvectors of the empirical covariance). We require though that the data is of full rank so that the empirical inverse covariance is well-defined. This requirement does not hold if the data points are contained within a lower-dimensional subspace, in particular, if their number is smaller than the dimension. We can easily avoid this restriction by initial projection of the data points onto the span of eigenvectors of the covariance matrix with nonzero eigenvalues. That is, by projecting the data onto the lowest-dimensional subspace containing it without losing any information.

Theorem 10 *If* $\mathbf{X}$ *is the data matrix,* $\hat{\mathbf{Q}}_2$ *is the minimizer of* (32) *and* $\operatorname{rank}(\mathbf{X}) = D$ *(equivalently the data points span* $\mathbb{R}^D$*), then*

$$\hat{\mathbf{Q}}_2 = (\mathbf{X}^T\mathbf{X})^{-1}/\operatorname{tr}((\mathbf{X}^T\mathbf{X})^{-1}). \tag{33}$$

We view (4) as a robust version of (32). Since we verified robustness of the subspace recovered by (4) and also showed that (32) yields the inverse covariance matrix, we sometimes refer to the solution of (4) as a robust inverse covariance matrix (though we have only verified robustness to subspace recovery). This idea helps us interpret our numerical procedure for minimizing (4), which we present in §4.

4. IRLS Algorithms for Minimizing (4)

We propose a fast algorithm for computing our M-estimator by using a straightforward iterative re-weighted least squares (IRLS) strategy. We first motivate this strategy in §4.1 (in particular, see (38) and (40)). We then establish its linear convergence in §4.2. At last, we describe its practical choices in §4.3 and summarize its complexity in §4.4.

4.1 Heuristic Proposal for Two IRLS Algorithms

The procedure for minimizing (4) formally follows from the simple fact that the directional derivative of F at $\hat{\mathbf{Q}}$ in any direction $\tilde{\mathbf{Q}} - \hat{\mathbf{Q}}$, where $\tilde{\mathbf{Q}} \in \mathbb{H}$, is 0, that is,

$$\left\langle F'(\hat{\mathbf{Q}})\Big|_{\mathbf{Q}=\hat{\mathbf{Q}}}, \tilde{\mathbf{Q}} - \hat{\mathbf{Q}} \right\rangle_F = 0 \quad \text{for any} \quad \tilde{\mathbf{Q}} \in \mathbb{H}. \tag{34}$$

We remark that since $\mathbb{H}$ is an affine subspace of matrices, (34) holds globally in $\mathbb{H}$ and not just locally around $\hat{\mathbf{Q}}$.

We formally differentiate (4) at $\hat{\mathbf{Q}}$ as follows (see more details in (44), which appears later):

$$F'(\mathbf{Q})\Big|_{\mathbf{Q}=\hat{\mathbf{Q}}} = \sum_{i=1}^{N} \frac{\hat{\mathbf{Q}}\mathbf{x}_i\mathbf{x}_i^T + \mathbf{x}_i\mathbf{x}_i^T\hat{\mathbf{Q}}}{2\|\hat{\mathbf{Q}}\mathbf{x}_i\|}. \tag{35}$$

Throughout the formal derivation we ignore the possibility of zero denominator in (35), that is, we assume that $\hat{\mathbf{Q}}\mathbf{x}_i \neq \mathbf{0}\ \forall\ 1 \leq i \leq N$; we later address this issue.

Since $F'(\hat{\mathbf{Q}})$ is symmetric and $\tilde{\mathbf{Q}} - \hat{\mathbf{Q}}$ can be any symmetric matrix with trace 0, it is easy to note that (34) implies that $F'(\hat{\mathbf{Q}})$ is a scalar matrix (e.g., multiply it by a basis of symmetric matrices with trace 0 whose members have exactly 2 nonzero matrix elements).

That is,

$$\sum_{i=1}^{N} \frac{\hat{\mathbf{Q}}\mathbf{x}_i\mathbf{x}_i^T + \mathbf{x}_i\mathbf{x}_i^T\hat{\mathbf{Q}}}{2\|\hat{\mathbf{Q}}\mathbf{x}_i\|} = c\mathbf{I} \tag{36}$$

for some $c \in \mathbb{R}$. This implies that

$$\hat{\mathbf{Q}} = c\left(\sum_{i=1}^{N} \frac{\mathbf{x}_i\mathbf{x}_i^T}{\|\hat{\mathbf{Q}}\mathbf{x}_i\|}\right)^{-1}. \tag{37}$$

Indeed, we can easily verify that (37) solves (36), furthermore, (36) is a Lyapunov equation whose solution is unique (see, e.g., page 1 of Bhatia and Drissi (2005)). Since $\mathrm{tr}(\hat{\mathbf{Q}}) = 1$, we obtain that

$$\hat{\mathbf{Q}} = \left(\sum_{i=1}^{N} \frac{\mathbf{x}_i\mathbf{x}_i^T}{\|\hat{\mathbf{Q}}\mathbf{x}_i\|}\right)^{-1} / \mathrm{tr}\left(\left(\sum_{i=1}^{N} \frac{\mathbf{x}_i\mathbf{x}_i^T}{\|\hat{\mathbf{Q}}\mathbf{x}_i\|}\right)^{-1}\right),$$

which suggests the following iterative estimate of $\hat{\mathbf{Q}}$:

$$\mathbf{Q}_{k+1} = \left(\sum_{i=1}^{N} \frac{\mathbf{x}_i\mathbf{x}_i^T}{\|\mathbf{Q}_k\mathbf{x}_i\|}\right)^{-1} / \mathrm{tr}\left(\left(\sum_{i=1}^{N} \frac{\mathbf{x}_i\mathbf{x}_i^T}{\|\mathbf{Q}_k\mathbf{x}_i\|}\right)^{-1}\right). \tag{38}$$

Formula (38) is undefined whenever $\mathbf{Q}_k\mathbf{x}_i = \mathbf{0}$ for some $k \in \mathbb{N}$ and $1 \le i \le N$. In theory, we address it as follows. Let $I(\mathbf{Q}) = \{1 \le i \le N : \mathbf{Q}\mathbf{x}_i = \mathbf{0}\}$, $\mathrm{L}(\mathbf{Q}) = \mathrm{Sp}\{\mathbf{x}_i\}_{i\in I(\mathbf{Q})}$ and

$$T(\mathbf{Q}) = \mathbf{P}_{\mathrm{L}(\mathbf{Q})^\perp}\left(\sum_{i\notin I(\mathbf{Q})} \frac{\mathbf{x}_i\mathbf{x}_i^T}{\|\mathbf{Q}\mathbf{x}_i\|}\right)^{-1}\mathbf{P}_{\mathrm{L}(\mathbf{Q})^\perp} / \mathrm{tr}\left(\mathbf{P}_{\mathrm{L}(\mathbf{Q})^\perp}\left(\sum_{i\notin I(\mathbf{Q})} \frac{\mathbf{x}_i\mathbf{x}_i^T}{\|\mathbf{Q}\mathbf{x}_i\|}\right)^{-1}\mathbf{P}_{\mathrm{L}(\mathbf{Q})^\perp}\right).$$

Using this notation, the iterative formula can be corrected as follows

$$\mathbf{Q}_{k+1} = T(\mathbf{Q}_k). \tag{39}$$

In practice, we can avoid data points satisfying $\|\mathbf{Q}_k\mathbf{x}_i\| \le \delta$ for a sufficiently small parameter δ (instead of $\|\mathbf{Q}_k\mathbf{x}_i\| = 0$). We follow a similar idea by replacing F with the regularized function F_δ for a regularized parameter δ. In this case, (39) obtains the following form:

$$\mathbf{Q}_{k+1} = \left(\sum_{i=1}^{N} \frac{\mathbf{x}_i\mathbf{x}_i^T}{\max(\|\mathbf{Q}_k\mathbf{x}_i\|, \delta)}\right)^{-1} / \mathrm{tr}\left(\left(\sum_{i=1}^{N} \frac{\mathbf{x}_i\mathbf{x}_i^T}{\max(\|\mathbf{Q}_k\mathbf{x}_i\|, \delta)}\right)^{-1}\right). \tag{40}$$

We note that the RHS of (39) is obtained as the limit of the RHS of (40) when δ approaches 0.

The two iterative formulas, that is, (39) and (40), give rise to IRLS algorithms. For simplicity of notation, we exemplify this idea with the formal expression in (38). It iteratively finds the solution to the following weighted (with weight $1/\|\mathbf{Q}_k\mathbf{x}_i\|$) least squares problem:

$$\arg\min_{\mathbf{Q}\in\mathbb{H}} \sum_{i=1}^{N} \frac{1}{\|\mathbf{Q}_k\mathbf{x}_i\|}\|\mathbf{Q}\mathbf{x}_i\|^2. \tag{41}$$

To show this, we note that (41) is a quadratic function and any formal directional derivative at $\mathbf{Q}_{k+1}$ is 0. Indeed,

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{Q}}\sum_{i=1}^{N}\frac{1}{\|\mathbf{Q}_k\mathbf{x}_i\|}\|\mathbf{Q}\mathbf{x}_i\|^2\Big|_{\mathbf{Q}=\mathbf{Q}_{k+1}}=\mathbf{Q}_{k+1}\left(\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\|\mathbf{Q}_k\mathbf{x}_i\|}\right)+\left(\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\|\mathbf{Q}_k\mathbf{x}_i\|}\right)\mathbf{Q}_{k+1}=c\mathbf{I}$$

for some $c\in\mathbb{R}$, and $\left\langle \mathbf{I},\tilde{\mathbf{Q}}-\mathbf{Q}_{k+1}\right\rangle_F=0$ for any $\tilde{\mathbf{Q}}\in\mathbb{H}$. Consequently, $\mathbf{Q}_{k+1}$ of (38) is the minimizer of (41).

Formula (40) (as well as (39)) provides another interpretation for $\hat{\mathbf{Q}}$ as robust inverse covariance (in addition to the one discussed in §3.3). Indeed, we note for example that the RHS of (40) is the scaled inverse of a weighted covariance matrix; the scaling enforces the trace of the inverse to be 1 and the weights of $\mathbf{x}_i\mathbf{x}_i^T$ are significantly larger when $\mathbf{x}_i$ is an inlier. In other words, the weights apply a shrinkage procedure for outliers. Indeed, since $\mathbf{Q}_k\mathbf{x}_i$ approaches $\hat{\mathbf{Q}}\mathbf{x}_i$ and the underlying subspace, which contain the inliers, is recovered by $\ker(\hat{\mathbf{Q}})$, for an inlier $\mathbf{x}_i$ the coefficient of $\mathbf{x}_i\mathbf{x}_i^T$ approaches $1/\delta$, which is a very large number (in practice we use $\delta=10^{-20}$). On the other hand, when $\mathbf{x}_i$ is sufficiently far from the underlying subspace, the coefficient of $\mathbf{x}_i\mathbf{x}_i^T$ is significantly smaller.

4.2 Theory: Convergence Analysis of the IRLS Algorithms

The following theorem analyzes the convergence of the sequence proposed by (39) to the minimizer of (4).

Theorem 11 *Let $\mathcal{X}=\{\mathbf{x}_i\}_{i=1}^N$ be a data set in $\mathbb{R}^D$ satisfying* (14)*, $\hat{\mathbf{Q}}$ the minimizer of* (4)*, $\mathbf{Q}_0$ an arbitrary symmetric matrix with $\mathrm{tr}(\mathbf{Q}_0)=1$ and $\{\mathbf{Q}_i\}_{i\in\mathbb{N}}$ the sequence obtained by iteratively applying* (39) *(while initializing it with $\mathbf{Q}_0$), then $\{\mathbf{Q}_i\}_{i\in\mathbb{N}}$ converges to a matrix $\tilde{\mathbf{Q}}\in\mathbb{H}$. If $\tilde{\mathbf{Q}}\mathbf{x}_i\neq\mathbf{0}$ for all $1\le i\le N$, then $\tilde{\mathbf{Q}}=\hat{\mathbf{Q}}$ and furthermore, $\{F(\mathbf{Q}_i)\}_{i\in\mathbb{N}}$ converges linearly to $F(\tilde{\mathbf{Q}})$ and $\{\mathbf{Q}_i\}_{i\in\mathbb{N}}$ converges r-linearly to $\tilde{\mathbf{Q}}$.*

The condition for the linear convergence to $\hat{\mathbf{Q}}$ in Theorem 11 (i.e., $\hat{\mathbf{Q}}\mathbf{x}_i\neq\mathbf{0}$ for all $1\le i\le N$) usually does not occur for noiseless data. This condition is common in IRLS algorithms whose objective functions are l_1-type and are not twice differentiable at $\mathbf{0}$. For example, Weiszfeld's Algorithm (Weiszfeld, 1937) may not converge to the geometric median but to one of the data points (Kuhn, 1973, §3.4). On the other hand, regularized IRLS algorithms often converge linearly to the minimizer of the regularized function. We demonstrate this principle in our case as follows.

Theorem 12 *Let $\mathcal{X}=\{\mathbf{x}_i\}_{i=1}^N$ be a data set in $\mathbb{R}^D$ satisfying* (14)*, $\mathbf{Q}_0$ an arbitrary symmetric matrix with $\mathrm{tr}(\mathbf{Q}_0)=1$ and $\{\mathbf{Q}_i\}_{i\in\mathbb{N}}$ the sequence obtained by iteratively applying* (40) *(while initializing it with $\mathbf{Q}_0$). Then, the sequence $\{F_\delta(\mathbf{Q}_i)\}_{i\in\mathbb{N}}$ converges linearly to the unique minimum of $F_\delta(\mathbf{Q})$, and $\{\mathbf{Q}_i\}_{i\in\mathbb{N}}$ converges r-linearly to the unique minimizer of $F_\delta(\mathbf{Q})$.*

The convergence rate of the iterative application of (40) depends on δ. Following Theorem 6.1 of Chan and Mulet (1999), this rate is at most

$$r(\delta) = \sqrt{\max_{\mathbf{\Delta}=\mathbf{\Delta}^T,\mathrm{tr}(\mathbf{\Delta})=0} \frac{\sum_{i=1,\|\mathbf{Q}_*\mathbf{x}_i\|>\delta}^{N} \frac{(\mathbf{x}_i^T\mathbf{\Delta}\mathbf{Q}_*\mathbf{x}_i)^2}{\|\mathbf{Q}_*\mathbf{x}_i\|^3}}{\sum_{i=1}^{N} \frac{\|\mathbf{\Delta}\mathbf{x}_i\|^2}{\max(\|\mathbf{Q}_*\mathbf{x}_i,\delta\|)}}}.$$

That is, $\|\mathbf{Q}_k - \hat{\mathbf{Q}}\| < C \cdot r(\delta)^k$ for some constant $C > 0$. If (14) holds, then $r(\delta) < 1$ for all $\delta > 0$ and $r(\delta)$ is a non-increasing function. Furthermore, if $\{\mathbf{x}_i \in \mathcal{X} : \|\hat{\mathbf{Q}}\mathbf{x}_i\| \neq 0\}$ satisfies assumption (14), then $\lim_{\delta\to 0} r(\delta) < 1$.

4.3 The Practical Choices for the IRLS Algorithm

Following the theoretical discussion in §4.2 we prefer using the regularized version of the IRLS algorithm. We fix the regularization parameter to be smaller than the rounding error, that is, $\delta = 10^{-20}$, so that the regularization is very close to the original problem (even without regularization the iterative process is stable, but may have few warnings on badly scaled or close to singular matrices). The idea of the algorithm is to iteratively apply (40) with an arbitrary initialization (symmetric with trace 1). We note that in theory $\{F_\delta(\mathbf{Q}_k)\}_{k\in\mathbb{N}}$ is non-increasing (see, e.g., the proof of Theorem 12). However, empirically the sequence decreases when it is within the rounding error to the minimizer. Therefore, we check $F_\delta(\mathbf{Q}_k)$ every four iterations and stop our algorithm when we detect an increase (we noticed empirically that checking every four iterations, instead of every iteration, improves the accuracy of the algorithm). Algorithm 2 summarizes our practical procedure for minimizing (4).

Algorithm 2 Practical and Regularized Minimization of (4)

Input: $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \subseteq \mathbb{R}^D$: data
Output: $\hat{\mathbf{Q}}$: a symmetric matrix in $\mathbb{R}^{D\times D}$ with $\mathrm{tr}(\hat{\mathbf{Q}}) = 1$.
Steps:
- $\delta = 10^{-20}$
- Arbitrarily initialize $\mathbf{Q}_0$ to be a symmetric matrix with $\mathrm{tr}(\mathbf{Q}_0) = 1$
- $k = -1$

repeat
- k=k+1
- $\mathbf{Q}_{k+1} = \left(\sum_{i=1}^{N} \frac{\mathbf{x}_i\mathbf{x}_i^T}{\max(\|\mathbf{Q}_k\mathbf{x}_i\|,\delta)}\right)^{-1} / \mathrm{tr}\left(\left(\sum_{i=1}^{N} \frac{\mathbf{x}_i\mathbf{x}_i^T}{\max(\|\mathbf{Q}_k\mathbf{x}_i\|,\delta)}\right)^{-1}\right)$.

until $F(\mathbf{Q}_{k+1}) > F(\mathbf{Q}_{k-3})$ and $\mathrm{mod}\,(k+1, 4) = 0$
- Output $\hat{\mathbf{Q}} := \mathbf{Q}_k$

4.4 Complexity of Algorithm 2

Each update of Algorithm 2 requires a complexity of order $O(N \cdot D^2)$, due to the sum of N $D \times D$ matrices. Therefore, for n_s iterations the total running time of Algorithm 2 is of order $O(n_s \cdot N \cdot D^2)$. In most of our numerical experiments n_s was less than 40. The storage of this algorithm is $O(N \times D)$, which amounts to storing $\mathcal{X}$. Thus, Algorithm 2 has

the same order of storage and complexity as PCA. In practice, it might be a bit slower due to a larger constant for the actual complexity.

5. Subspace Recovery in Practice

We view the GMS algorithm as a prototype for various subspace recovery algorithms. We discuss here modifications and extensions of this procedure in order to make it even more practical. Sections 5.1 and 5.2 discuss the cases where d is unknown and known respectively; in particular, §5.2.1 proposes the EGMS algorithm when d is known. At last, §5.3 concludes with the computational complexity of the GMS and EGMS algorithms.

5.1 Subspace Recovery without Knowledge of d

In theory, the subspace recovery described here can work without knowing the dimension d. In the noiseless case, one may use (5) to estimate the subspace as guaranteed by Theorem 1. In the case of small noise one can estimate d from the eigenvalues of $\hat{\mathbf{Q}}$ and then apply the GMS algorithm. This strategy is theoretically justified by Theorems 1 and 6 as well as the discussion following (81). The problem is that condition (9) for guaranteeing exact recovery by GMS is restrictive; in particular, it requires the number of outliers to be larger than at least $D - d$ (according to our numerical experiments it is safe to use the lower bound $1.5\,(D-d)$). For practitioners, this is a failure mode of GMS, especially when the dimension of the data set is large (for example, $D > N$).

While this seems to be a strong restriction, we remark that the problem of exact subspace recovery without knowledge of the intrinsic dimension is rather hard and some assumptions on data sets or some knowledge of data parameters would be expected. Other algorithms for this problem, such as Chandrasekaran et al. (2011), Candès et al. (2011), Xu et al. (2010b) and McCoy and Tropp (2011), require estimates of unknown regularization parameters (which often depend on various properties of the data, in particular, the unknown intrinsic dimension) or strong assumptions on the underlying distribution of the outliers or corrupted elements.

We first note that if only conditions (6) and (7) hold, then Theorem 1 still guarantees that the GMS algorithm outputs a subspace containing the underlying subspace. Using some information on the data one may recover the underlying subspace from the outputted subspace containing it, even when dealing with the failure mode.

In the rest of this section we describe several practical solutions for dealing with the failure mode, in particular, with small number of outliers. We later demonstrate them numerically in §6.2 for artificial data and in §6.7 and §6.8 for real data.

Our first practical solution is to reduce the ambient dimension of the data. When the reduction is not too aggressive, it can be performed via PCA. In §5.2.1 we also propose a robust dimensionality reduction which can be used instead. There are two problems with this strategy. First of all, the reduced dimension is another parameter that requires tuning. Second of all, some information may be lost by the dimensionality reduction and thus exact recovery of the underlying subspace is generally impossible.

A second practical solution is to add artificial outliers. The number of added outliers should not be too large (otherwise (6) and (7) will be violated), but they should sufficiently permeate through $\mathbb{R}^D$ so that (9) holds. In practice, the number of outliers can be $2D$,

since empirically (9) holds with high probability when $N_0 = 2(D-d)$. To overcome the possible impact of outliers with arbitrarily large magnitude, we project the data with artificial outliers onto the sphere (following Lerman et al. 2012). Furthermore, if the original data matrix does not have full rank (in particular if $N < D$) we reduce the dimension of the data (by PCA) to be the rank of the data matrix. This dimensionality reduction clearly does not result in any loss of information. We refer to the whole process of initial "lossless dimensionality reduction" (if necessary), addition of $2D$ artificial Gaussian outliers, normalization onto the sphere and application of GMS (with optional estimation of d by the eigenvalues of $\hat{\mathbf{Q}}$) as the GMS2 algorithm. We believe that it is the best practical solution to avoid condition (9) when d is unknown.

A third solution is to regularize our M estimator, that is, to minimize the following objective function with the regularization parameter λ:

$$\hat{\mathbf{Q}} = \underset{\mathrm{tr}(\mathbf{Q})=1, \mathbf{Q}=\mathbf{Q}^T}{\arg\min} \sum_{i=1}^{N} \|\mathbf{Q}\mathbf{x}_i\| + \lambda \|\mathbf{Q}\|_F^2. \tag{42}$$

The IRLS algorithm then becomes

$$\mathbf{Q}_{k+1} = \left(\sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\max(\|\mathbf{Q}_k \mathbf{x}_i\|, \delta)} + 2\lambda \mathbf{I} \right)^{-1} / \mathrm{tr} \left(\left(\sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\max(\|\mathbf{Q}_k \mathbf{x}_i\|, \delta)} \right)^{-1} + 2\lambda \mathbf{I} \right).$$

We note that if $\lambda = 0$ and there are only few outliers, then in the noiseless case $\dim(\ker(\hat{\mathbf{Q}})) > d$ and in the small noise case the number of significantly small eigenvalues is bigger than d. On the other hand when $\lambda \to \infty$, $\hat{\mathbf{Q}} \to \mathbf{I}/D$, whose kernel is degenerate (similarly, it has no significantly small eigenvalues). Therefore, there exists an appropriate λ for which $\dim(\ker(\hat{\mathbf{Q}}))$ (or the number of significantly small eigenvalues of $\hat{\mathbf{Q}}$) is d. This formulation transforms the estimation of d into estimation of λ. This strategy is in line with other common regularized solutions to this problem (see, e.g., Chandrasekaran et al. 2011; Candès et al. 2011; Xu et al. 2010b; McCoy and Tropp 2011), however, we find it undesirable to estimate a regularization parameter that is hard to interpret in terms of the data.

5.2 Subspace Recovery with Knowledge of d

Knowledge of the intrinsic dimension d can help improve the performance of GMS or suggest completely new variants (especially as GMS always finds a subspace containing the underlying subspace). For example, knowledge of d can be used to carefully estimate the parameter λ of (42), for example, by finding λ yielding exactly a d-dimensional subspace via a bisection procedure.

Lerman et al. (2012) modified the strategy described in here by requiring an additional constraint on the maximal eigenvalue of $\mathbf{Q}$ in (28): $\lambda_{\max}(\mathbf{Q}) \leq \frac{1}{D-d}$ (where $\lambda_{\max}(\mathbf{Q})$ is the largest eigenvalue of $\mathbf{Q}$). This approach has theoretical guarantees, but it comes with the price of additional SVD in each iteration, which makes the algorithm slightly more expensive. Besides, in practice (i.e., noisy setting) this approach requires tuning the upper bound on $\lambda_{\max}(\mathbf{Q})$. Indeed, the solution $\mathbf{Q}'$ to their minimization problem (with $\lambda_{\max}(\mathbf{Q}') \leq 1/(D-d)$ and $\mathrm{tr}(\mathbf{Q}') = 1$) satisfies that $\dim(\ker(\mathbf{Q}')$ is at most d and equals d when $\mathbf{Q}'$ is a

scaled projector operator. They proved that $\dim(\ker(\mathbf{Q}') = d$ for the setting of pure inliers (lying exactly on a subspace) under some conditions avoiding the three types of enemies. However, in practice (especially in noisy cases) the actual subspace often has dimension smaller than d and thus the bound on $\lambda_{\max}(\mathbf{Q})$ has to be tuned as an additional parameter. In some cases, one may take $\lambda_{\max}(\mathbf{Q}) > \frac{1}{D-d}$ and find the subspace according to the bottom d eigenvectors. In other cases, a bisection method on the bound of $\lambda_{\max}(\mathbf{Q})$ provide more accurate results (see related discussion in Lerman et al. (2012, §6.1.6)).

5.2.1 The EGMS Algorithm

We formulate in Algorithm 3 the Extended Geometric Median Subspace (EGMS) algorithm for subspace recovery with known intrinsic dimension.

Algorithm 3 The Extended Geometric Median Subspace Algorithm

Input: $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N \subseteq \mathbb{R}^D$: data, d: dimension of L^*, an algorithm for minimizing (4)
Output: $\hat{\mathrm{L}}$: a d-dimensional linear subspace in $\mathbb{R}^D$.
Steps:
• $\hat{\mathrm{L}} = \mathbb{R}^D$
repeat
• $\hat{\mathbf{Q}} = \arg\min_{\mathbf{Q}\in\mathbb{H},\mathbf{Q}\mathbf{P}_{\hat{\mathrm{L}}^\perp}=\mathbf{0}} F(\mathbf{Q})$
• $\mathbf{u} =$ the top eigenvector of $\hat{\mathbf{Q}}$
• $\hat{\mathrm{L}} = \hat{\mathrm{L}} \cap \mathrm{Sp}(\mathbf{u}^\perp)$
until $\dim(\hat{\mathrm{L}}) = d$

We justify this basic procedure in the noiseless case without requiring (9) as follows.

Theorem 13 *Assume that $d, D \in \mathbb{N}$, $d < D$, $\mathcal{X}$ is a data set in $\mathbb{R}^D$ and $\mathrm{L}^* \in \mathrm{G}(D, d)$. If only conditions* (6) *and* (7) *hold, then the EGMS Algorithm exactly recovers L^*.*

In §6.5 we show how the vectors obtained by EGMS at each iteration can be used to form robust principal components (in reverse order), even when $\hat{\mathbf{Q}}$ is degenerate.

5.3 Computational Complexity of GMS and EGMS

The computational complexity of GMS is of the same order as that of Algorithm 2, that is, $O(n_s \cdot N \cdot D^2)$ (where n_s is the number of required iterations for Algorithm 2). Indeed, after obtaining $\hat{\mathbf{Q}}$, computing L^* by its smallest d eigenvectors takes an order of $O(d \cdot D^2)$ operations.

EGMS on the other hand repeats Algorithm 2 $D - d$ times; therefore it adds an order of $O((D-d) \cdot n_s \cdot N \cdot D^2)$ operations, where n_s denotes the total number of iterations for Algorithm 2. In implementation, we can speed up the EGMS algorithm by excluding the span of some of the top eigenvectors of $\hat{\mathbf{Q}}$ from $\hat{\mathrm{L}}$ (instead of excluding only the top eigenvector in the third step of Algorithm 3). We demonstrate this modified procedure on artificial setting in §6.2.

6. Numerical Experiments

We compare our proposed estimator to other algorithms, while using both synthetic and real data. We also demonstrate the effectiveness of some of our practical proposals. In §6.1 we describe a model for generating synthetic data. Using this model, we respectively demonstrate in §6.2-§6.4 the effectiveness of the following strategies: the practical solutions of §5.1 and §5.2, our estimation of the subspace dimension, and our regularization (more precisely, its effect on the recovery error). In §6.5 we demonstrate the use of our M estimator for robust estimation of eigenvectors of the covariance (or the inverse covariance) matrix. At last, actual comparisons are demonstrated in §6.6-§6.8 for synthetic data, face data and video surveillance data respectively.

6.1 Model for Synthetic Data

In §6.2-§6.4 and §6.6 we generate data from the following model. We randomly choose $\mathrm{L}^* \in \mathrm{G}(D,d)$, sample N_1 inliers from the d-dimensional Multivariate Normal distribution $N(\mathbf{0}, \mathbf{I}_{d\times d})$ on L^* and add N_0 outliers sampled from a uniform distribution on $[0,1]^D$. The outliers are strongly asymmetric around the subspace to make the subspace recovery problem more difficult (Lerman and Zhang, 2010). In some experiments below additional Gaussian noise is considered. When referring to this synthetic data we only need to specify its parameters N_1, N_0, D, d and possibly the standard deviation for the additive noise. For any subspace recovery algorithm (or heuristics), we denote by $\tilde{\mathrm{L}}$ its output (i.e., the estimator for L^*) and measure the corresponding recovery error by $e_{\tilde{\mathrm{L}}} = \|\mathbf{P}_{\tilde{\mathrm{L}}} - \mathbf{P}_{\mathrm{L}^*}\|_F$.

6.2 Demonstration of Practical Solutions of §5.1 and §5.2

We present two different artificial cases, where in one of them condition (9) holds and in the other one it does not hold and test the practical solutions of §5.1 and §5.2 in the second case.

The two cases are the following instances of the synthetic model of §6.1: (a) (N_1, N_0, D, d) = (100 , 100, 100, 20) and (b) $(N_1, N_0, D, d) = (100, 20, 100, 20)$. The GMS algorithm estimates the underlying subspace L^* given $d = 20$ with recovery errors 2.1×10^{-10} and 3.4 in cases (a) and (b) respectively. In case (a) there are sufficiently many outliers (with respect to $D - d$) and the GMS algorithm is successful. We later show in §6.3 that the underlying dimension ($d = 20$) can be easily estimated by the eigenvalues of $\hat{\mathbf{Q}}$. In case (b) $N_0 = 0.25 * (D - d)$, therefore, condition (9) is violated and the GMS algorithm completely fails.

We demonstrate the success of the practical solutions of §5.1 and §5.2 in case (b). We assume that the dimension d is known, though in §6.3 we estimate d correctly for the non-regularized solutions of §5.1. Therefore, these solutions can be also applied without knowing the dimension. If we reduce the dimension of the data set in case (b) from $D = 100$ to $D = 35$ (via PCA; though one can also use EGMS), then GMS (with $d = 20$) achieves a recovery error of 0.23, which indicates that GMS almost recovers the subspace correctly. We remark though that if we reduce the dimension to, for example, $D = 55$, then the GMS algorithm will still fail. We also note that the recovery error is not as attractive as the

ones below; this observation probably indicates that some information was lost during the dimension reduction.

The GMS2 algorithm with $d = 20$ recovers the underlying subspace in case (b) with error 1.2×10^{-10}. This is the method we advocated for when possibly not knowing the intrinsic dimension.

The regularized minimization of (42) with $\lambda = 100$ works well for case (b). In fact, it recovers the subspace as $\ker \hat{\mathbf{Q}}$ (without using its underlying dimension) with error 3.3×10^{-13}. The only issue is how to determine the value of λ. We claimed in §5.2 that if d is known, then λ can be carefully estimated by the bisection method. This is true for this example, in fact, we initially chose λ this way.

We remark that the REAPER algorithm of Lerman et al. (2012) did not perform well for this particular data, though in general it is a very successful solution. The recovery error of the direct REAPER algorithm was 3.725 (and 3.394 for S-REAPER) and the error for its modified version via bisection (relaxing the bound on the largest eigenvalue so that $\dim(\ker(\hat{\mathbf{Q}})) = 20$) was 3.734 (and 3.175 for S-REAPER).

At last we demonstrate the performance of EGMS and its faster heuristic with $d = 20$. The recovery error of the original EGMS for case (b) is only 0.095. We suggested in §5.3 a faster heuristic for EGMS, which can be reformulated as follows: In the third step of Algorithm 3, we replace $\mathbf{u}$ (the top eigenvector of $\hat{\mathbf{Q}}$) with $\mathbf{U}$, the subspace spanned by several top eigenvectors. In the noiseless case, we could let $\mathbf{U}$ be the span of the nonzero eigenvectors of $\hat{\mathbf{Q}}$. This modification of EGMS (for the noiseless case) required only two repetitions of Algorithm 2 and its recovery error was 2.2×10^{-13}. In real data sets with noise we need to determine the number of top eigenvectors spanning $\mathbf{U}$, which makes this modification of EGMS less automatic.

6.3 Demonstration of Dimension Estimation

We test dimension estimation by eigenvalues of $\hat{\mathbf{Q}}$ for cases (a) and (b) of §6.2. The eigenvalues of $\hat{\mathbf{Q}}$ obtained by Algorithm 2 for the two cases are shown in Figure 2. In case (a), the largest logarithmic eigengap (i.e., the largest gap in logarithms of eigenvalues) occurs at 80, so we can correctly estimate that $d = D - 80 = 20$ (the eigenvalues are not zero since Algorithm 2 uses the δ-regularized objective function). However, in case (b) the largest eigengap occurs at 60 and thus mistakenly predicts $d = 40$.

As we discussed in §6.2, the dimension estimation fails here since condition (9) is not satisfied. However, we have verified that if we try any of the solutions proposed in §5.1 then we can correctly recover that $d = 20$ by the logarithmic eigengap. For example, in Figure 2 we demonstrate the logarithms of eigenvalues of $\hat{\mathbf{Q}}$ in case (b) after dimensionality reduction (via PCA) onto dimension $D = 35$ and it is clear that the largest gap is at $d = 20$ (or $D - d = 80$). We obtained similar graphs when using $2D$ artificial outliers (more precisely, the GMS2 algorithm without the final application of the GMS algorithm) or the regularization of (42) with $\lambda = 100$.

6.4 The Effect of the Regularization Parameter δ

We assume a synthetic data set sampled according to the model of §6.1 with $(N_1, N_0, D, d) = (250, 250, 100, 10)$. We use the GMS algorithm with $d = 10$ and different values of the

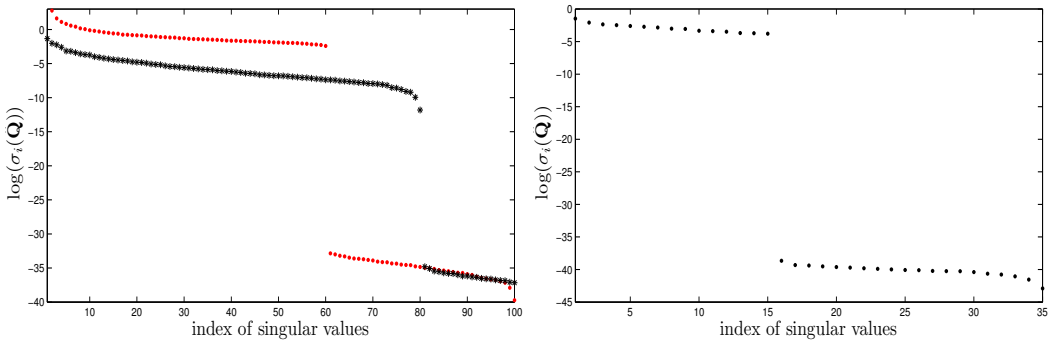


Figure 2: Dimension estimation: In the left figure, the starred points and the dotted point represent log-scaled eigenvalues of the output of Algorithm 2 for cases (a) and (b) respectively (see §6.3). The right figure corresponds to case (b) with dimension reduced to 35.

regularization parameter δ and record the recovery error in Figure 3. For $10^{-14} \leq \delta \leq 10^{-2}$, $\log(\text{error}) - \log(\delta)$ is constant. We thus empirically obtain that the error is of order $O(\delta)$ in this range. On the other hand, (27) only obtained an order of $O(\sqrt{\delta})$. It is possible that methods similar to those of Coudron and Lerman (2012) can obtain sharper error bounds. We also expect that for δ sufficiently small (here smaller than 10^{-14}), the rounding error becomes dominant. On the other hand, perturbation results are often not valid for sufficiently large δ (here this is the case for $\delta > 10^{-2}$).

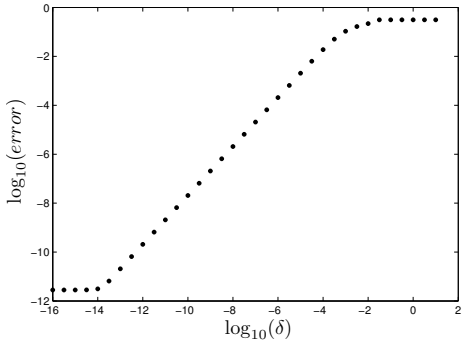


Figure 3: The recovery errors and the regularization parameters δ

6.5 Information Obtained from Eigenvectors

Throughout the paper we emphasized the subspace recovery problem, but did not discuss at all the information that can be inferred from the eigenvectors of our robust PCA strategy. Since in standard PCA these vectors have significant importance, we exemplify the information obtained from our robust PCA and compare it to that obtained from PCA and some other robust PCA algorithms.

We create a sample from a mixture of two Gaussian distributions with the same mean and same eigenvalues of the covariance matrices, but different eigenvectors of the covariance matrices. The mixture percentages are 25% and 75%. We expect the eigenvectors of any good robust PCA algorithm (robust to outliers as perceived in this paper) to be close to that of the covariance of the main component (with 75%).

More precisely, we sample 300 points from $N(\mathbf{0}, \mathbf{\Sigma}_1)$, where $\mathbf{\Sigma}_1$ is a 10×10 diagonal matrix with elements $1, 2^{-1}, 2^{-2}, \cdots, 2^{-9}$ and 100 points from $N(\mathbf{0}, \mathbf{\Sigma}_2)$, where $\mathbf{\Sigma}_2 = \mathbf{U}\mathbf{\Sigma}_1\mathbf{U}^T$, where $\mathbf{U}$ is randomly chosen from the set of all orthogonal matrices in $\mathbb{R}^{10\times 10}$. The goal is to estimate the eigenvectors of $\mathbf{\Sigma}_1$ (i.e., the standard basis vectors in $\mathbb{R}^{10}$) in the presence of 25% "outliers". Unlike the subspace recovery problem, where we can expect to exactly recover a linear structure among many outliers, here the covariance structure is more complex and we cannot exactly recover it with 25% outliers.

We estimated the eigenvectors of $\mathbf{\Sigma}_1$ by the the eigenvectors of $\hat{\mathbf{Q}}$ of Algorithm 2 in reverse order (recall that $\hat{\mathbf{Q}}$ is a scaled and robust version of the inverse covariance). We refer to this procedure as "EVs (eigenvectors) of $\hat{\mathbf{Q}}^{-1}$". We also estimated these eigenvectors by standard PCA, LLD (McCoy and Tropp, 2011) with $\lambda = 0.8\sqrt{D/N}$ and PCP (Candès et al., 2011) with $\lambda = 1/\sqrt{\max(D, N)}$. We repeated the random simulation (with different samples for the random orthogonal matrix $\mathbf{U}$) 100 times and reported in Table 2 the average angles between the estimated and actual top two eigenvectors of $\mathbf{\Sigma}_1$ according to the different methods. We note that the "EVs of $\hat{\mathbf{Q}}^{-1}$" outperforms PCA, LLD (or OP) and PCP in terms of estimation of the top two eigenvectors of $\mathbf{\Sigma}_1$. We remark though that PCP does not suit for robust estimation of the empirical covariance and thus the comparison is unfair for PCP.

	EVs of $\hat{\mathbf{Q}}^{-1}$	LLD	PCP	PCA
Eigenvector 1	3.0°	5.5°	45.7°	14.8°
Eigenvector 2	3.0°	5.5°	47.4°	40.3°

Table 2: Angles (in degrees) between the estimated and actual top two eigenvectors of $\mathbf{\Sigma}_1$.

When the covariance matrix $\mathbf{\Sigma}_1$ (and consequently also $\mathbf{\Sigma}_2$) is degenerate, $\hat{\mathbf{Q}}$ might be singular and therefore $\hat{\mathbf{Q}}$ cannot be directly used to robustly estimate eigenvectors of the covariance matrix. For this case, EGMS (Algorithm 3) can be used, where the vector $\mathbf{u}$ obtained in the ith iteration of Algorithm 3 can be considered as the $(D - i + 1)$st robust eigenvector (that is, we reverse the order again). To test the performance of this method, we modify $\mathbf{\Sigma}_1$ in the above model as follows: $\mathbf{\Sigma}_1$=diag$(1, 0.5, 0.25, 0, 0, \cdots, 0)$. We repeated the random simulations of this modified model 100 times and reported in Table 2 the average angles between the estimated and actual top two eigenvectors of $\mathbf{\Sigma}_1$ according to the different methods. Here LLD did slightly better than EGMS and they both outperformed PCA (and PCP).

	EGMS	LLD	PCP	PCA
Eigenvector 1	5.2°	3.4°	42.6°	8.2°
Eigenvector 2	5.2°	3.4°	47.3°	16.1°

Table 3: Angles (in degrees) between the estimated and actual top two eigenvectors of $\mathbf{\Sigma}_1$.

6.6 Detailed Comparison with Other Algorithms for Synthetic Data

Using the synthetic data of §6.1, we compared the GMS algorithm with the following algorithms: MDR (Mean Absolute Deviation Rounding) of McCoy and Tropp (2011), LLD (Low-Leverage Decomposition) of McCoy and Tropp (2011), OP (Outlier Pursuit) of Xu et al. (2010b), PCP (Principal Component Pursuit) of Candès et al. (2011), MKF (Median K-flats with $K = 1$) of Zhang et al. (2009), HR-PCA (High-dimensional Robust PCA) of Xu et al. (2010a), a common M-estimator (Huber and Ronchetti, 2009, see, e.g.,) and R_1-PCA of Ding et al. (2006). The codes of OP and HR-PCA were obtained from `http://guppy.mpe.nus.edu.sg/~mpexuh`, the code of MKF from `http://www.math.umn.edu/~zhang620/mkf`, the code of PCP from `http://perception.csl.illinois.edu/matrix-rank/sample_code.html` with the Accelerated Proximal Gradient and full SVD version, the codes of MDR and LLD from `http://www.acm.caltech.edu/~mccoy/code/` and the codes of the common M-estimator, R_1-PCA and GMS will appear in a supplemental webpage. We also record the output of standard PCA, where we recover the subspace by the span of the top d eigenvectors. We ran the experiments on a computer with Intel Core 2 CPU at 2.66GHz and 2 GB memory.

We remark that since the basic GMS algorithm already performed very well on these artificial instances, we did not test its extensions and modifications described in §5 (e.g., GMS2 and EGMS).

For all of our experiments with synthetic data, we could correctly estimate d by the largest logarithmic eigengap of the output of Algorithm 2. Nevertheless, we used the knowledge of d for all algorithms for the sake of fair comparison.

For LLD, OP and PCP we estimated L^* by the span of the top d eigenvectors of the low-rank matrix. Similarly, for the common M-estimator we used the span of the top d eigenvectors of the estimated covariance $\mathbf{A}$. For the HR-PCA algorithm we also used the true percentage of outliers (50% in our experiments). For LLD, OP and PCP we set the mixture parameter λ as $0.8\sqrt{D/N}, 0.8\sqrt{D/N}, 1/\sqrt{\max(D,N)}$ respectively (following the suggestions of McCoy and Tropp (2011) for LLD/OP and Candès et al. (2011) for PCP). These choices of parameters are also used in experiments with real data sets in §6.7 and §6.8.

For the common M-estimator, we used $u(x) = 2\max(\ln(x)/x, 10^{30})$ and the algorithm discussed by Kent and Tyler (1991). Considering the conditions in §3.1.1, we also tried other functions: $u(x) = \max(x^{-0.5}, 10^{30})$ had a significantly larger recovery error and $u(x) = \max(x^{-0.9}, 10^{30})$ resulted in a similar recovery error as $\max(\ln(x)/x, 10^{30})$ but a double running time.

We used the syntectic data with different values of (N_1, N_0, D, d). In some instances we also add noise from the Gaussian distribution $N(0, \eta^2 \mathbf{I})$ with $\eta = 0.1$ or 0.01. We repeated each experiment 20 times (due to the random generation of data). We record in Table 4 the mean running time, the mean recovery error and their standard deviations.

We remark that PCP is designed for uniformly corrupted coordinates of data, instead of corrupted data points (i.e., outliers), therefore, the comparison with PCP is somewhat unfair for this kind of data. On the other hand, the applications in §6.7 and §6.8 are tailored for the PCP model (though the other algorithms still apply successfully to them).

From Table 4 we can see that GMS is the fastest robust algorithm. Indeed, its running time is comparable to that of PCA. We note that this is due to its linear convergence rate (usually it converges in less than 40 iterations). The common M-estimator is the closest algorithm in terms of running time to GMS, since it also has the linear convergence rate. In contrast, PCP, OP and LLD need a longer running time since their convergence rates are much slower. Overall, GMS performs best in terms of exact recovery. The PCP, OP and LLD algorithms cannot approach exact recovery even by tuning the parameter λ. For example, in the case where $(N_1, N_0, D, d) = (125, 125, 10, 5)$ with $\eta = 0$, we checked a geometric sequence of 101 λ values from 0.01 to 1, and the smallest recovery errors for LLD, OP and PCP are 0.17, 0.16 and 0.22 respectively. The common M-estimator performed very well for many cases (sometimes slightly better than GMS), but its performance deteriorates as the density of outliers increases (e.g., poor performance for the case where $(N_1, N_0, D, d) = (125, 125, 10, 5)$). Indeed, Theorem 9 indicates problems with the exact recovery of the common M-estimator.

At last, we note that the empirical recovery error of the GMS algorithm for noisy data sets is in the order of $\sqrt{\eta}$, where η is the size of noise.

6.7 Yale Face data

Following Candès et al. (2011), we apply our algorithm to face images. It has been shown that face images from the same person lie in a low-dimensional linear subspace of dimension at most 9 (Basri and Jacobs, 2003). However, cast shadows, specular reflections and saturations could possibly distort this low-rank modeling. Therefore, one can use a good robust PCA algorithm to remove these errors if one has many images from the same face.

We used the images of the first two persons in the extended Yale face database B (Lee et al., 2005), where each of them has 65 images of size 192×168 under different illumination conditions. Therefore we represent each person by 65 vectors of length 32256. Following Basri and Jacobs (2003) we applied GMS, GMS2 and EGMS with $d = 9$ and we also reduced the 65×32256 matrix to 65×65 (in fact, we only reduced the representation of the column space) by rejecting left vectors with zero singular values. We also applied the GMS algorithm after initial dimensionality reduction (via PCA) to $D = 20$. The running times of EGMS and GMS (without dimensionality reduction) are 13 and 0.16 seconds respectively on average for each face (we used the same computer as in §6.6). On the other hand, the running times of PCP and LLD are 193 and 2.7 seconds respectively. Moreover, OP ran out of memory. The recovered images are shown in Figure 4, where the shadow of the nose and the parallel lines were removed best by EGMS. The GMS algorithm without dimension reduction did not perform well, due to the difficulty explained in §5 and demonstrated in

(N_1, N_0, D, d)		GMS	MDR	LLD	OP	PCP	HR-PCA	MKF	PCA	M-est.	R_1-PCA
	e	6e-11	0.275	1.277	0.880	0.605	0.210	0.054	0.193	0.102	0.121
$(125, 125, 10, 5)$	$std.e$	4e-11	0.052	0.344	0.561	0.106	0.049	0.030	0.050	0.037	0.048
$\eta = 0$	$t(s)$	0.008	0.371	0.052	0.300	0.056	0.378	0.514	0.001	0.035	0.020
	$std.t$	0.002	0.120	0.005	0.054	0.002	0.001	0.262	8e-06	4e-04	0.014
	e	0.011	0.292	1.260	1.061	0.567	0.233	0.069	0.213	0.115	0.139
$(125, 125, 10, 5)$	$std.e$	0.004	0.063	0.316	0.491	0.127	0.075	0.036	0.073	0.054	0.073
$\eta = 0.01$	$t(s)$	0.008	0.340	0.053	0.287	0.056	0.380	0.722	0.001	0.035	0.052
	$std.t$	0.001	0.075	0.007	0.033	0.001	0.009	0.364	1e-05	4e-04	0.069
	e	0.076	0.264	1.352	0.719	0.549	0.200	0.099	0.185	0.122	0.128
$(125, 125, 10, 5)$	$std.e$	0.023	0.035	0.161	0.522	0.102	0.051	0.033	0.048	0.041	0.050
$\eta = 0.1$	$t(s)$	0.007	0.332	0.055	0.301	0.056	0.378	0.614	0.001	0.035	0.032
	$std.t$	0.001	0.083	0.004	0.044	0.001	0.001	0.349	7e-06	4e-04	0.037
	e	2e-11	0.652	0.258	0.256	0.261	0.350	0.175	0.350	1e-12	0.307
$(125, 125, 50, 5)$	$std.e$	3e-11	0.042	0.030	0.032	0.033	0.023	0.028	0.025	5e-12	0.029
$\eta = 0$	$t(s)$	0.015	0.420	0.780	1.180	3.164	0.503	0.719	0.006	0.204	0.020
	$std.t$	0.001	0.128	0.978	0.047	0.008	0.055	0.356	9e-05	0.001	0.011
	e	0.061	0.655	0.274	0.271	0.273	0.355	0.196	0.359	0.007	0.321
$(125, 125, 50, 5)$	$std.e$	0.009	0.027	0.039	0.038	0.040	0.038	0.038	0.033	0.001	0.038
$\eta = 0.01$	$t(s)$	0.023	0.401	4.155	1.506	0.499	0.653	0.656	0.006	0.191	0.028
	$std.t$	0.002	0.079	0.065	0.197	0.006	0.044	0.377	8e-05	0.001	0.022
	e	0.252	0.658	0.292	0.290	0.296	0.358	0.264	0.363	0.106	0.326
$(125, 125, 50, 5)$	$std.e$	0.027	0.033	0.032	0.032	0.033	0.027	0.031	0.032	0.014	0.032
$\eta = 0.1$	$t(s)$	0.021	0.363	0.923	1.726	0.501	0.638	0.641	0.006	0.191	0.025
	$std.t$	0.001	0.063	0.033	0.470	0.009	0.051	0.240	1e-04	0.001	0.012
	e	3e-12	0.880	0.214	0.214	0.215	0.332	0.161	0.330	2e-12	0.259
$(250, 250, 100, 10)$	$std.e$	2e-12	0.018	0.019	0.019	0.019	0.014	0.024	0.012	9e-12	0.016
$\eta = 0$	$t(s)$	0.062	1.902	3.143	7.740	2.882	1.780	1.509	0.039	0.819	1.344
	$std.t$	0.006	0.354	4.300	0.038	0.014	0.041	1.041	3e-04	0.023	0.708
	e	0.077	0.885	0.217	0.216	0.219	0.334	0.164	0.335	0.009	0.263
$(250, 250, 100, 10)$	$std.e$	0.006	0.031	0.019	0.018	0.020	0.019	0.019	0.017	3e-04	0.018
$\eta = 0.01$	$t(s)$	0.084	1.907	21.768	11.319	2.923	1.785	1.412	0.039	0.400	1.086
	$std.t$	0.010	0.266	0.261	0.291	0.014	0.041	0.988	3e-04	0.002	0.738
	e	0.225	0.888	0.238	0.237	0.262	0.342	0.231	0.345	0.136	0.276
$(250, 250, 100, 10)$	$std.e$	0.016	0.020	0.019	0.019	0.019	0.019	0.018	0.015	0.010	0.019
$\eta = 0.1$	$t(s)$	0.076	1.917	4.430	16.649	2.876	1.781	1.555	0.039	0.413	1.135
	$std.t$	0.007	0.299	0.069	1.184	0.014	0.025	0.756	4e-04	0.011	0.817
	e	4e-11	1.246	0.162	0.164	0.167	0.381	0.136	0.381	3e-13	0.239
$(500, 500, 200, 20)$	$std.e$	1e-10	0.018	0.011	0.011	0.011	0.010	0.009	0.008	6e-14	0.009
$\eta = 0$	$t(s)$	0.464	23.332	16.778	89.090	16.604	8.602	5.557	0.347	6.517	15.300
	$std.t$	0.024	2.991	0.878	1.836	0.100	0.216	4.810	0.009	0.126	3.509
	e	0.082	1.247	0.160	0.162	0.166	0.374	0.139	0.378	0.012	0.236
$(500, 500, 200, 20)$	$std.e$	0.003	0.018	0.007	0.007	0.008	0.011	0.010	0.006	2e-04	0.007
$\eta = 0.01$	$t(s)$	0.592	23.214	128.51	122.61	16.823	8.541	6.134	0.354	2.361	15.165
	$std.t$	0.060	3.679	1.155	6.500	0.036	0.219	4.318	0.019	0.064	3.485
	e	0.203	1.262	0.204	0.204	0.250	0.391	0.275	0.398	0.166	0.270
$(500, 500, 200, 20)$	$std.e$	0.007	0.012	0.007	0.007	0.007	0.012	0.272	0.009	0.005	0.008
$\eta = 0.1$	$t(s)$	0.563	24.112	24.312	202.22	16.473	8.552	8.745	0.348	2.192	15.150
	$std.t$	0.061	2.362	0.226	8.362	0.050	0.155	3.408	0.010	0.064	3.420

Table 4: Mean running times, recovery errors and their standard deviations for synthetic data.

§6.2. The GMS2 algorithm turns out to work well, except for the second image of face 2. However, other algorithms such as PCP and GMS with dimension reduction ($D = 20$) performed even worse on this image and LLD did not remove any shadow at all; the only good algorithm for this image is EGMS.

Figure 4: Recovering faces: (a) given images, (b)-(f) the recovered images by EGMS, GMS without dimension reduction, GMS2, GMS with dimension reduced to 20, PCP and LLD respectively

6.8 Video Surveillance

For background subtraction in surveillance videos (Li et al., 2004), we consider the following two videos used by Candès et al. (2011): "Lobby in an office building with switching on / off lights" and "Shopping center" from `http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html`. In the first video, the resolution is 160×128 and we used 1546 frames from 'SwitchLight1000.bmp' to 'SwitchLight2545.bmp'. In the second video, the resolution is 320×256 and we use 1000 frames from 'ShoppingMall1001.bmp' to 'ShoppingMall2000.bmp'. Therefore, the data matrices are of size 1546×20480 and 1001×81920. We used a computer with Intel Core 2 Quad Q6600 2.4GHz and 8 GB memory due to the large size of these data.

We applied GMS, GMS2 and EGMS with $d = 3$ and with initial dimensionality reduction to 200 to reduce running time. For this data we are unaware of a standard choice of d; though we noticed empirically that the outputs of our algorithms as well as other algorithms are very stable to changes in d within the range $2 \leq d \leq 5$. We obtain the foreground by the orthogonal projection to the recovered 3-dimensional subspace. Figure 5 demonstrates foregrounds detected by EGMS, GMS, GMS2, PCP and LLD, where PCP and LLD used $\lambda = 1/\sqrt{\max(D, N)}, 0.8\sqrt{D/N}$. We remark that OP ran out of memory. Using truth labels provided in the data, we also form ROC curves for GMS, GMS2, EGMS and PCP in Figure 6 (LLD is not included since it performed poorly for any value of λ we tried). We note that PCP performs better than both GMS and EGMS in the 'Shoppingmall' video, whereas the latter algorithms perform better than PCP in the 'SwitchLight' video. Furthermore, GMS is significantly faster than EGMS and PCP. Indeed, the running times (on average) of GMS, EGMS and PCP are 91.2, 1018.8 and 1209.4 seconds respectively.

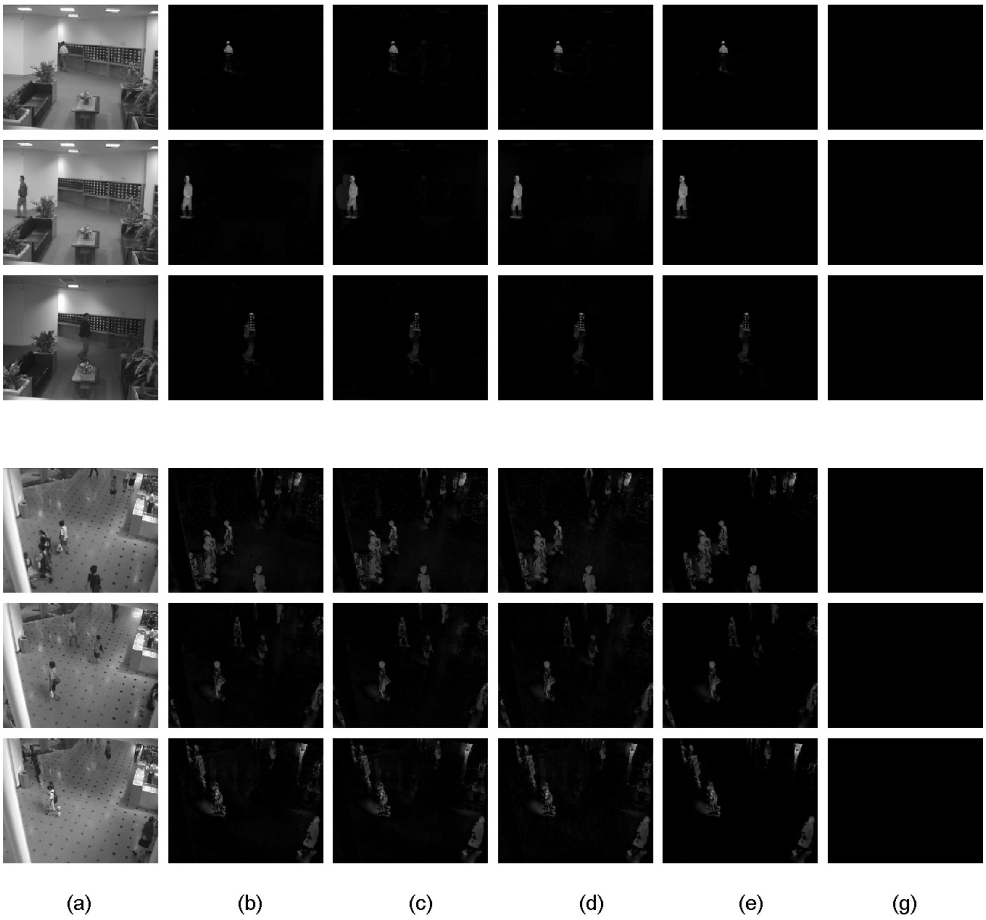


Figure 5: Video surveillance: (a) the given frames (b)-(e) the detected foreground by EGMS, GMS, GMS2, PCP, LLD respectively

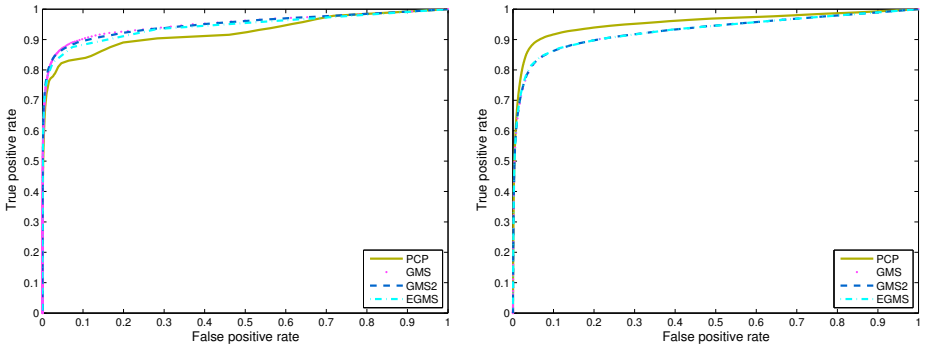


Figure 6: ROC curves for EGMS, GMS, GMS2 and PCP in the 'SwitchLight' video (the left figure) and the 'Shoppingmall' video (the right figure)

7. Proofs of Theorems

We present the technical proofs of the theoretical statements of this paper according to their order of appearance.

7.1 Proof of Theorem 1

We will prove that if conditions (6) and (7) hold, then the set of all minimizers satisfying (4) coincides with the set of all minimizers satisfying (8). This clearly implies that if conditions (6) and (7) hold, then any minimizer $\hat{\mathbf{Q}}$ of (4) satisfies $\ker(\hat{\mathbf{Q}}) \supseteq \mathrm{L}^*$ (indeed, this condition is equivalent with the condition $\mathbf{Q}\mathbf{P}_{\mathrm{L}^*} = \mathbf{0}$, which appears in the formulation of (8)). If condition (9) also holds, then $\ker(\hat{\mathbf{Q}}) = \mathrm{L}^*$ and the theorem is concluded.

We assume that conditions (6) and (7) hold and arbitrarily fix a minimizer $\hat{\mathbf{Q}}_0$ of the oracle problem (8). We claim that in order to establish the equivalence of the sets of solutions of (4) and (8), it is sufficient to prove that

$$F(\hat{\mathbf{Q}}_0 + \mathbf{\Delta}) - F(\hat{\mathbf{Q}}_0) > 0 \text{ for any symmetric } \mathbf{\Delta} \text{ with } \mathrm{tr}(\mathbf{\Delta}) = 0 \text{ and } \mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*} \neq \mathbf{0}. \quad (43)$$

Indeed, we first note that (43) implies that $\hat{\mathbf{Q}}_0$ is also a minimizer of (4). This observation follows from combining (43) with the following equation:

$$F(\hat{\mathbf{Q}}_0 + \mathbf{\Delta}) - F(\hat{\mathbf{Q}}_0) \geq 0 \text{ for any symmetric } \mathbf{\Delta} \text{ with } \mathrm{tr}(\mathbf{\Delta}) = 0 \text{ and } \mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*} = \mathbf{0},$$

which is an immediate consequence of the definition of (8). To conclude the equivalence, we assume on the contrary that there exists $\hat{\mathbf{Q}}_0$, which is a minimizer of (8) but not a minimizer of (4). We denote by $\hat{\mathbf{Q}}_0'$ a minimizer of (8), which is also a minimizer of (4) and let $\mathbf{\Delta} := \hat{\mathbf{Q}}_0' - \hat{\mathbf{Q}}_0$. Then by the definitions of $\hat{\mathbf{Q}}_0$, $\hat{\mathbf{Q}}_0'$ and $\mathbf{\Delta}$: $\mathrm{tr}(\mathbf{\Delta}) = 0$, $\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*} \neq \mathbf{0}$ and $F(\hat{\mathbf{Q}}_0') = F(\hat{\mathbf{Q}}_0)$. This contradicts (43) and thus concludes the proof.

In order to conclude (43) (and thus the theorem) we first differentiate $\|\mathbf{Q}\mathbf{x}\|$ at $\mathbf{Q} = \mathbf{Q}_0$ when $\mathbf{x} \in \ker(\mathbf{Q}_0)^{\perp}$ as follows:

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{Q}}\|\mathbf{Q}\mathbf{x}\|\Big|_{\mathbf{Q}=\mathbf{Q}_0} = \frac{\mathrm{d}}{\mathrm{d}\mathbf{Q}}\sqrt{\|\mathbf{Q}\mathbf{x}\|^2}\Big|_{\mathbf{Q}=\mathbf{Q}_0} = \frac{\mathrm{d}}{\mathrm{d}\mathbf{Q}}\frac{\mathbf{Q}\mathbf{x}\mathbf{x}^T\mathbf{Q}^T}{2\|\mathbf{Q}_0\mathbf{x}\|}\Big|_{\mathbf{Q}=\mathbf{Q}_0} = \frac{\mathbf{Q}_0\mathbf{x}\mathbf{x}^T + \mathbf{x}\mathbf{x}^T\mathbf{Q}_0}{2\|\mathbf{Q}_0\mathbf{x}\|}. \quad (44)$$

We note that for any $\mathbf{x} \in \mathbb{R}^D \setminus \{\mathbf{0}\}$ satisfying $\hat{\mathbf{Q}}_0\mathbf{x} \neq \mathbf{0}$ and $\mathbf{\Delta} \in \mathbb{R}^{D\times D}$ symmetric:

$$\|(\hat{\mathbf{Q}}_0 + \mathbf{\Delta})\mathbf{x}\| - \|\hat{\mathbf{Q}}_0\mathbf{x}\| \geq \left\langle \mathbf{\Delta}, (\hat{\mathbf{Q}}_0\mathbf{x}\mathbf{x}^T + \mathbf{x}\mathbf{x}^T\hat{\mathbf{Q}}_0)/2\|\hat{\mathbf{Q}}_0\mathbf{x}\| \right\rangle_F = \left\langle \mathbf{\Delta}, \hat{\mathbf{Q}}_0\mathbf{x}\mathbf{x}^T/\|\hat{\mathbf{Q}}_0\mathbf{x}\| \right\rangle_F. \tag{45}$$

Indeed, the first equality follows from (44) and the convexity of $\|\mathbf{Q}\mathbf{x}\|$ in $\mathbf{Q}$ and the second equality follows from the symmetry of $\mathbf{\Delta}$ and $\hat{\mathbf{Q}}_0$ as well as the definition of the Frobenius dot product.

If on the other hand $\hat{\mathbf{Q}}_0\mathbf{x} = \mathbf{0}$, then clearly

$$\|(\hat{\mathbf{Q}}_0 + \mathbf{\Delta})\mathbf{x}\| - \|\hat{\mathbf{Q}}_0\mathbf{x}\| = \|\mathbf{\Delta}\mathbf{x}\|. \tag{46}$$

For simplicity of our presentation, we use (46) only for $\mathbf{x} \in \mathcal{X}_1$ (where obviously $\hat{\mathbf{Q}}_0\mathbf{x} = \mathbf{0}$ since $\hat{\mathbf{Q}}_0\mathbf{P}_{\mathrm{L}^*} = 0$). On the other hand, we use (45) for all $\mathbf{x} \in \mathcal{X}_0$. One can easily check that if $\mathbf{x} \in \mathcal{X}_0$ and $\hat{\mathbf{Q}}_0\mathbf{x} = \mathbf{0}$, then replacing (45) with (46) does not change the analysis below. Using these observations we note that

$$F(\hat{\mathbf{Q}}_0 + \mathbf{\Delta}) - F(\hat{\mathbf{Q}}_0) \geq \sum_{\mathbf{x}\in\mathcal{X}_1} \|\mathbf{\Delta}\mathbf{x}\| + \sum_{\mathbf{x}\in\mathcal{X}_0} \left\langle \mathbf{\Delta}, \hat{\mathbf{Q}}_0\mathbf{x}\mathbf{x}^T/\|\hat{\mathbf{Q}}_0\mathbf{x}\| \right\rangle_F. \tag{47}$$

We assume first that $\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*} = \mathbf{0}$. In this case, $\hat{\mathbf{Q}}_0 + \mathbf{\Delta} \in \mathbb{H}$ and $(\hat{\mathbf{Q}}_0 + \mathbf{\Delta})\mathbf{P}_{\mathrm{L}^*} = \mathbf{0}$. Since $\hat{\mathbf{Q}}_0$ is the minimizer of (8), we obtain the following identity (which is analogous to (34)):

$$\sum_{\mathbf{x}\in\mathcal{X}_0} \left\langle \mathbf{\Delta}, \frac{\hat{\mathbf{Q}}_0\mathbf{x}\mathbf{x}^T}{\|\hat{\mathbf{Q}}_0\mathbf{x}\|} \right\rangle_F \geq 0 \quad \forall\ \mathbf{\Delta} \in \mathbb{R}^{D\times D} \text{ s.t. } \mathrm{tr}(\mathbf{\Delta}) = 0\ , \mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*} = \mathbf{0}. \tag{48}$$

We will prove (43) by showing that the RHS of (47) is positive for any symmetric $\mathbf{\Delta}$ with $\mathrm{tr}(\mathbf{\Delta}) = 0$ and $\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*} \neq \mathbf{0}$. Using (47) and the facts that $\mathcal{X}_1 \subset \mathrm{L}^*$ and $\hat{\mathbf{Q}}_0 = \mathbf{P}_{\mathrm{L}^{*\perp}}\hat{\mathbf{Q}}_0$ (since $\mathbf{P}_{\mathrm{L}^*}\hat{\mathbf{Q}}_0 = \hat{\mathbf{Q}}_0\mathbf{P}_{\mathrm{L}^*} = \mathbf{0}$), we establish the following inequality:

$$\begin{aligned}
&F(\hat{\mathbf{Q}}_0 + \mathbf{\Delta}) - F(\hat{\mathbf{Q}}_0) \geq \sum_{\mathbf{x}\in\mathcal{X}_1} \|\mathbf{\Delta}\mathbf{x}\| + \sum_{\mathbf{x}\in\mathcal{X}_0} \left\langle \mathbf{\Delta}, \hat{\mathbf{Q}}_0\mathbf{x}\mathbf{x}^T/\|\hat{\mathbf{Q}}_0\mathbf{x}\| \right\rangle_F \\
&= \sum_{\mathbf{x}\in\mathcal{X}_1} \|\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*}\mathbf{x}\| + \sum_{\mathbf{x}\in\mathcal{X}_0} \left\langle (\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*} + \mathbf{\Delta}\mathbf{P}_{\mathrm{L}^{*\perp}}), \mathbf{P}_{\mathrm{L}^{*\perp}}\hat{\mathbf{Q}}_0\mathbf{x}\mathbf{x}^T/\|\hat{\mathbf{Q}}_0\mathbf{x}\| \right\rangle_F \\
&\geq \sum_{\mathbf{x}\in\mathcal{X}_1} \left(\|\mathbf{P}_{\mathrm{L}^*}\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*}\mathbf{x}\| + \|\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*}\mathbf{x}\|\right)/\sqrt{2} \\
&+ \sum_{\mathbf{x}\in\mathcal{X}_0} \left\langle (\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*} + \mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^{*\perp}}), \hat{\mathbf{Q}}_0\mathbf{x}\mathbf{x}^T/\|\hat{\mathbf{Q}}_0\mathbf{x}\| \right\rangle_F.
\end{aligned} \tag{49}$$

For ease of notation we denote $\mathbf{\Delta}_0 = \mathrm{tr}(\mathbf{P}_{\mathrm{L}^*}\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*})\mathbf{v}_0\mathbf{v}_0^T$, where $\mathbf{v}_0$ is the minimizer of the RHS of (6). Combining the following two facts: $\mathrm{tr}(\mathbf{\Delta}_0) - \mathrm{tr}(\mathbf{P}_{\mathrm{L}^*}\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*}) = 0$ and $\mathrm{tr}(\mathbf{P}_{\mathrm{L}^*}\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*}) + \mathrm{tr}(\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^{*\perp}}) = \mathrm{tr}(\mathbf{\Delta}) = 0$, we obtain that

$$\mathrm{tr}(\mathbf{\Delta}_0 + \mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^{*\perp}}) = 0.$$

Further application of (48) implies that

$$\sum_{\mathbf{x}\in\mathcal{X}_0} \left\langle \mathbf{\Delta}_0 + \mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^{*\perp}}, \hat{\mathbf{Q}}_0\mathbf{x}\mathbf{x}^T/\|\hat{\mathbf{Q}}_0\mathbf{x}\| \right\rangle_F \geq 0. \tag{50}$$

We note that

$$\begin{aligned}&\left\langle \mathbf{P}_{\mathrm{L}^{*\perp}}\boldsymbol{\Delta}\mathbf{P}_{\mathrm{L}^{*\perp}}, \hat{\mathbf{Q}}_0\mathbf{x}\mathbf{x}^T/\|\hat{\mathbf{Q}}_0\mathbf{x}\|\right\rangle_F = \left\langle \mathbf{P}_{\mathrm{L}^{*\perp}}\boldsymbol{\Delta}\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{P}_{\mathrm{L}^{*\perp}}, \hat{\mathbf{Q}}_0\mathbf{x}\mathbf{x}^T/\|\hat{\mathbf{Q}}_0\mathbf{x}\|\right\rangle_F\\ =&\left\langle \mathbf{P}_{\mathrm{L}^{*\perp}}\boldsymbol{\Delta}\mathbf{P}_{\mathrm{L}^{*\perp}}, \hat{\mathbf{Q}}_0\mathbf{x}\mathbf{x}^T\mathbf{P}_{\mathrm{L}^{*\perp}}/\|\hat{\mathbf{Q}}_0\mathbf{x}\|\right\rangle_F. \end{aligned} \tag{51}$$

Combining (50) and (51) we conclude that

$$\begin{aligned}&-\sum_{\mathbf{x}\in\mathcal{X}_0}\left\langle \mathbf{P}_{\mathrm{L}^{*\perp}}\boldsymbol{\Delta}\mathbf{P}_{\mathrm{L}^{*\perp}}, \hat{\mathbf{Q}}_0\mathbf{x}\mathbf{x}^T/\|\hat{\mathbf{Q}}_0\mathbf{x}\|\right\rangle_F \le \sum_{\mathbf{x}\in\mathcal{X}_0}\left\langle \boldsymbol{\Delta}_0, \hat{\mathbf{Q}}_0\mathbf{x}\mathbf{x}^T\mathbf{P}_{\mathrm{L}^{*\perp}}/\|\hat{\mathbf{Q}}_0\mathbf{x}\|\right\rangle_F\\ =&\sum_{\mathbf{x}\in\mathcal{X}_0}\operatorname{tr}(\mathbf{P}_{\mathrm{L}^*}\boldsymbol{\Delta}\mathbf{P}_{\mathrm{L}^*})(\mathbf{v}_0^T\hat{\mathbf{Q}}_0\mathbf{x}/\|\hat{\mathbf{Q}}_0\mathbf{x}\|)(\mathbf{v}_0^T\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}) \le |\operatorname{tr}(\mathbf{P}_{\mathrm{L}^*}\boldsymbol{\Delta}\mathbf{P}_{\mathrm{L}^*})|\sum_{\mathbf{x}\in\mathcal{X}_0}|\mathbf{v}_0^T\mathbf{x}|. \end{aligned} \tag{52}$$

We apply (52) and then use (6) with $\mathbf{Q}=\mathbf{P}_{\mathrm{L}^*}\boldsymbol{\Delta}\mathbf{P}_{\mathrm{L}^*}/\operatorname{tr}(\mathbf{P}_{\mathrm{L}^*}\boldsymbol{\Delta}\mathbf{P}_{\mathrm{L}^*})$ to obtain the inequality:

$$\begin{aligned}&\sum_{\mathbf{x}\in\mathcal{X}_1}\|\mathbf{P}_{\mathrm{L}^*}\boldsymbol{\Delta}\mathbf{P}_{\mathrm{L}^*}\mathbf{x}\|/\sqrt{2}+\sum_{\mathbf{x}\in\mathcal{X}_0}\left\langle \mathbf{P}_{\mathrm{L}^{*\perp}}\boldsymbol{\Delta}\mathbf{P}_{\mathrm{L}^{*\perp}}, \hat{\mathbf{Q}}_0\mathbf{x}\mathbf{x}^T/\|\hat{\mathbf{Q}}_0\mathbf{x}\|\right\rangle_F\\ \ge&\sum_{\mathbf{x}\in\mathcal{X}_1}\|\mathbf{P}_{\mathrm{L}^*}\boldsymbol{\Delta}\mathbf{P}_{\mathrm{L}^*}\mathbf{x}\|/\sqrt{2}-|\operatorname{tr}(\mathbf{P}_{\mathrm{L}^*}\boldsymbol{\Delta}\mathbf{P}_{\mathrm{L}^*})|\sum_{\mathbf{x}\in\mathcal{X}_0}|\mathbf{v}_0^T\mathbf{x}|>0. \end{aligned} \tag{53}$$

We define $\mathbb{H}_1=\{\mathbf{Q}\in\mathbb{H}:\ \mathbf{Q}\mathbf{P}_{\mathrm{L}^{*\perp}}=\mathbf{0}\}$ and claim that (7) leads to the following inequality:

$$\sum_{\mathbf{x}\in\mathcal{X}_1}\|\mathbf{Q}(\mathbf{P}_{\mathrm{L}^*}\mathbf{x})\|>\sqrt{2}\sum_{\mathbf{x}\in\mathcal{X}_0}\|\mathbf{Q}(\mathbf{P}_{\mathrm{L}^*}\mathbf{x})\|\quad\forall\mathbf{Q}\in\mathbb{H}_1. \tag{54}$$

Indeed, since the RHS of (54) is a convex function of $\mathbf{Q}$, its maximum is achieved at the set of all extreme points of $\mathbb{H}_1$, which is $\{\mathbf{Q}\in\mathbb{R}^{D\times D}:\mathbf{Q}=\mathbf{v}\mathbf{v}^T,\text{where }\mathbf{v}\in\mathrm{L}^*,\|\mathbf{v}\|=1\}$. Therefore the maximum of the RHS of (54) is the RHS of (7). Since the minimum of the LHS of (54) is also the LHS of (7), (54) is proved.

We also claim that (54) can be extended from $\mathbb{H}_1$ to all $\mathbf{Q}\in\mathbb{R}^{D\times D}$ such that $\mathbf{Q}\mathbf{P}_{\mathrm{L}^{*\perp}}=\mathbf{0}$. Indeed, for any $\mathbf{Q}\in\mathbb{R}^{D\times D}$ satisfying $\mathbf{Q}\mathbf{P}_{\mathrm{L}^{*\perp}}=\mathbf{0}$ and having the SVD decomposition $\mathbf{Q}=\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, we can assign the following matrix $\mathbf{Q}'=\mathbf{Q}'(\mathbf{Q})\in\mathbb{H}_1$: $\mathbf{Q}':=\mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^T/\operatorname{tr}(\mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^T)$. It is not hard to note that the inequality in (54) holds for $\mathbf{Q}$ if and only if it holds for $\mathbf{Q}'$.

By first applying Cauchy's inequality, then using the defining property of projections and at last applying (54) with $\mathbf{Q}=\mathbf{P}_{\mathrm{L}^{*\perp}}\boldsymbol{\Delta}\mathbf{P}_{\mathrm{L}^*}$ (while using its latter extension beyond $\mathbb{H}_1$), we obtain the inequality:

$$\begin{aligned}&\sum_{\mathbf{x}\in\mathcal{X}_1}\|\mathbf{P}_{\mathrm{L}^{*\perp}}\boldsymbol{\Delta}\mathbf{P}_{\mathrm{L}^*}\mathbf{x}\|/\sqrt{2}+\sum_{\mathbf{x}\in\mathcal{X}_0}\left\langle \mathbf{P}_{\mathrm{L}^{*\perp}}\boldsymbol{\Delta}\mathbf{P}_{\mathrm{L}^*}, \hat{\mathbf{Q}}_0\mathbf{x}\mathbf{x}^T/\|\hat{\mathbf{Q}}_0\mathbf{x}\|\right\rangle_F\\ \ge&\sum_{\mathbf{x}\in\mathcal{X}_1}\|\mathbf{P}_{\mathrm{L}^{*\perp}}\boldsymbol{\Delta}\mathbf{P}_{\mathrm{L}^*}\mathbf{x}\|/\sqrt{2}-\sum_{\mathbf{x}\in\mathcal{X}_0}\|\mathbf{P}_{\mathrm{L}^{*\perp}}\boldsymbol{\Delta}\mathbf{P}_{\mathrm{L}^*}\mathbf{x}\|\\ =&\sum_{\mathbf{x}\in\mathcal{X}_1}\|\mathbf{P}_{\mathrm{L}^{*\perp}}\boldsymbol{\Delta}\mathbf{P}_{\mathrm{L}^*}(\mathbf{P}_{\mathrm{L}^*}\mathbf{x})\|/\sqrt{2}-\sum_{\mathbf{x}\in\mathcal{X}_0}\|\mathbf{P}_{\mathrm{L}^{*\perp}}\boldsymbol{\Delta}\mathbf{P}_{\mathrm{L}^*}(\mathbf{P}_{\mathrm{L}^*}\mathbf{x})\|>0. \end{aligned} \tag{55}$$

Finally, we combine (53) and (55) and conclude that the RHS of (49) is nonnegative and consequently (43) holds.

7.2 Proof of Theorem 2

Assume on the contrary that F is not strictly convex, in particular, there exists $0 < t_0 < 1$ such that

$$t_0 \cdot F(\mathbf{Q}_1) + (1-t_0) \cdot F(\mathbf{Q}_2) = F(t_0 \cdot \mathbf{Q}_1 + (1-t_0) \cdot \mathbf{Q}_2) \;\text{ for }\; \mathbf{Q}_1 \neq \mathbf{Q}_2,$$

or equivalently,

$$t_0 \cdot \sum_{i=1}^{N} \|\mathbf{Q}_1 \mathbf{x}_i\| + (1-t_0) \cdot \sum_{i=1}^{N} \|\mathbf{Q}_2 \mathbf{x}_i\| = \sum_{i=1}^{N} \|(t_0 \cdot \mathbf{Q}_1 + (1-t_0) \cdot \mathbf{Q}_2)\mathbf{x}_i\|. \tag{56}$$

Combining (56) with the fact that $\|\mathbf{Q}_1 \mathbf{x}_i\| + \|\mathbf{Q}_2 \mathbf{x}_i\| \geq \|(\mathbf{Q}_1 + \mathbf{Q}_2)\mathbf{x}_i\|$, we obtain that $t_0 \cdot \|\mathbf{Q}_1 \mathbf{x}_i\| + (1-t_0) \cdot \|\mathbf{Q}_2 \mathbf{x}_i\| = \|(t_0 \cdot \mathbf{Q}_1 + (1-t_0) \cdot \mathbf{Q}_2)\mathbf{x}_i\|$ for any $1 \leq i \leq N$ and therefore there exists a sequence $\{c_i\}_{i=1}^{N} \subset \mathbb{R}$ such that

$$\mathbf{Q}_2 \mathbf{x}_i = \mathbf{0} \;\text{ or }\; \mathbf{Q}_1 \mathbf{x}_i = c_i \, \mathbf{Q}_2 \mathbf{x}_i \;\text{ for all } 1 \leq i \leq N. \tag{57}$$

We conclude Theorem 2 by considering two different cases. We first assume that $\ker(\mathbf{Q}_1) = \ker(\mathbf{Q}_2)$. We denote

$$\tilde{\mathbf{Q}}_1 = \mathbf{P}_{\ker(\mathbf{Q}_1)^\perp} \mathbf{Q}_1 \mathbf{P}_{\ker(\mathbf{Q}_1)^\perp} \text{ and } \tilde{\mathbf{Q}}_2 = \mathbf{P}_{\ker(\mathbf{Q}_1)^\perp} \mathbf{Q}_2 \mathbf{P}_{\ker(\mathbf{Q}_1)^\perp}.$$

It follows from (57) that

$$\tilde{\mathbf{Q}}_1(\mathbf{P}_{\ker(\mathbf{Q}_1)^\perp} \mathbf{x}_i) = c_i \, \tilde{\mathbf{Q}}_2(\mathbf{P}_{\ker(\mathbf{Q}_1)^\perp} \mathbf{x}_i)$$

and consequently that $\mathbf{P}_{\ker(\mathbf{Q}_1)^\perp} \mathbf{x}_i$ lies in one of the eigenspaces of $\tilde{\mathbf{Q}}_1^{-1} \tilde{\mathbf{Q}}_2$. We claim that $\tilde{\mathbf{Q}}_1^{-1} \tilde{\mathbf{Q}}_2$ is a scalar matrix. Indeed, if on the contrary $\tilde{\mathbf{Q}}_1^{-1} \tilde{\mathbf{Q}}_2$ is not a scalar matrix, then $\{\mathbf{P}_{\ker(\mathbf{Q}_1)^\perp} \mathbf{x}_i\}_{i=1}^{N}$ lies in a union of several eigenspaces with dimensions summing to $\dim(\mathbf{P}_{\ker(\mathbf{Q}_1)^\perp})$ and this contradicts (14). In view of this property of $\tilde{\mathbf{Q}}_1^{-1} \tilde{\mathbf{Q}}_2$ and the fact that $\mathrm{tr}(\tilde{\mathbf{Q}}_1) = \mathrm{tr}(\hat{\mathbf{Q}}_1) = 1$ we have that $\tilde{\mathbf{Q}}_1 = \tilde{\mathbf{Q}}_2$ and $\mathbf{Q}_1 = \mathbf{Q}_2$, which contradicts our current assumption.

Next, assume that $\ker(\mathbf{Q}_1) \neq \ker(\mathbf{Q}_2)$. We will first show that if $1 \leq i \leq N$ is arbitrarily fixed, then $\mathbf{x}_i \in \ker(\mathbf{Q}_2) \cup \ker(\mathbf{P}_{\ker(\mathbf{Q}_1)} \mathbf{Q}_2)$. Indeed, if $\mathbf{x}_i \notin \ker(\mathbf{Q}_2)$, then using (57) we have $\mathbf{Q}_1 \mathbf{x}_i = c_i \, \mathbf{Q}_2 \mathbf{x}_i$. This implies that $c_i \mathbf{P}_{\ker(\mathbf{Q}_1)} \mathbf{Q}_2 \mathbf{x}_i = \mathbf{P}_{\ker(\mathbf{Q}_1)} \mathbf{Q}_1 \mathbf{x}_i = \mathbf{0}$ and thus $\mathbf{x}_i \in \ker(\mathbf{P}_{\ker(\mathbf{Q}_1)} \mathbf{Q}_2)$. That is, $\mathcal{X}$ is contained in the union of the 2 subspaces $\ker(\mathbf{Q}_2)$ and $\ker(\mathbf{P}_{\ker(\mathbf{Q}_1)} \mathbf{Q}_2)$. The dimensions of both spaces are less than D. This obvious for $\ker(\mathbf{Q}_2)$, since $\mathrm{tr}(\mathbf{Q}_2) = 1$. For $\ker(\mathbf{P}_{\ker(\mathbf{Q}_1)} \mathbf{Q}_2)$ it follows from the fact that $\ker(\mathbf{Q}_1) \neq \ker(\mathbf{Q}_2)$ and thus $\mathbf{P}_{\ker(\mathbf{Q}_1)} \mathbf{Q}_2 \neq \mathbf{0}$. We thus obtained a contradiction to (14).

7.3 Verification of (10) and (11) as Sufficient Conditions and (12) and (13) as Necessary Ones

We revisit the proof of Theorem 1 and first show that (10) and (11) can replace (6) and (7) in the first part of Theorem 1. We only deal with the first part of Theorem 1, which assumes that (9) holds, since (9) guarantees that (10) and (11) are well-defined (see the discussion in §2.4.1).

To show that (11) can replace (7), we prove the inequality in (55) using (11) as follows. Assuming that the SVD of $\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*}$ is $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, then $\mathbf{Q}' := \mathbf{V}\mathbf{\Sigma}\mathbf{V}^T/\operatorname{tr}(\mathbf{\Sigma})$ satisfies $\mathbf{Q}' \in \mathbb{H}, \mathbf{Q}'\mathbf{P}_{\mathrm{L}^{*\perp}} = \mathbf{0}$ and $\|\mathbf{Q}'\mathbf{x}\| = \|\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*}\mathbf{x}\|/\operatorname{tr}(\mathbf{\Sigma}) = \|\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*}\mathbf{x}\|/\|\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*}\|_*$. Using this fact, we obtain that

$$\sum_{\mathbf{x}\in\mathcal{X}_1}\|\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*}\mathbf{x}\| \geq \|\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*}\|_* \min_{\mathbf{Q}'\in\mathbb{H},\mathbf{Q}'\mathbf{P}_{\mathrm{L}^{*\perp}}=\mathbf{0}} \sum_{\mathbf{x}\in\mathcal{X}_1}\|\mathbf{Q}'\mathbf{x}\|. \tag{58}$$

We also note that

$$\begin{aligned}\sum_{\mathbf{x}\in\mathcal{X}_0}\left\langle \mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*}, \hat{\mathbf{Q}}_0\mathbf{x}\mathbf{x}^T/\|\hat{\mathbf{Q}}_0\mathbf{x}\|\right\rangle_F &= \sum_{\mathbf{x}\in\mathcal{X}_0}\left\langle \mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*}, \hat{\mathbf{Q}}_0\mathbf{x}\mathbf{x}^T\mathbf{P}_{\mathrm{L}^*}/\|\hat{\mathbf{Q}}_0\mathbf{x}\|\right\rangle_F \\ \geq -\|\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{\Delta}\mathbf{P}_{\mathrm{L}^*}\|_* &\left\|\sum_{\mathbf{x}\in\mathcal{X}_0}\hat{\mathbf{Q}}_0\mathbf{x}\mathbf{x}^T\mathbf{P}_{\mathrm{L}^*}/\|\hat{\mathbf{Q}}_0\mathbf{x}\|\right\|. \end{aligned}\tag{59}$$

Therefore (55) follows from (11), (58) and (59). Similarly, one can show that (10) may replace (6).

One can also verify that (12) and (13) are necessary conditions for exact recovery by revisiting the proof of Theorem 1 and reversing inequalities.

7.4 Proof of Lemma 3

We first note by symmetry that the minimizer of the LHS of (16) for the needle-haystack model is $\mathbf{Q} = \mathbf{P}_{\mathrm{L}^*}/d$. We can thus rewrite (16) in this case as $\alpha_1 \mathbb{E}\, r_1/d > 2\sqrt{2}\alpha_0 \mathbb{E}\, r_0/(D-d)$, where the "radii" r_1 and r_0 are the norms of the normal distributions with covariances $\sigma_1^2 d^{-1}\mathbf{P}_{\mathrm{L}^*}$ and $\sigma_0^2 D^{-1}\mathbf{P}_{\mathrm{L}^{*\perp}}$ respectively. Let $\tilde{r}_1$ and $\tilde{r}_2$ be the χ-distributed random variables with d and $D-d$ degrees of freedoms, then (16) obtains the form

$$\frac{\alpha_1\sigma_1}{d\sqrt{d}}\mathbb{E}\,\tilde{r}_1 > \frac{2\sqrt{2}\alpha_0\sigma_0}{(D-d)\sqrt{D}}\mathbb{E}\,\tilde{r}_0.$$

Applying (B.7) of Lerman et al. (2012), $\mathbb{E}\,\tilde{r}_1 \geq \sqrt{d/2}$ and $\mathbb{E}\,\tilde{r}_0 \leq \sqrt{D-d}$. Therefore (16) follows from (15).

7.5 Proof of Theorem 4

For simplicity of the proof we first assume that the supports of μ_0 and μ_1 are contained in a ball centered at the origin of radius M.

We start with the proof of (9) "in expectation" and then extend it to hold with high probability. We use the notation $F_I(\mathbf{Q})$ and $\hat{\mathbf{Q}}_I$ defined in (19) and (20) respectively. The spherical symmetry of $\mu_{0,\mathrm{L}^{*\perp}}$ implies that

$$\hat{\mathbf{Q}}_I = \frac{1}{D-d}\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{P}_{\mathrm{L}^{*\perp}}^T \tag{60}$$

is the unique minimizer of (20). To see this formally, we first note that $\mu_{0,\mathrm{L}^{*\perp}}$ satisfies the two-subspaces criterion of Coudron and Lerman (2012) for any $0 < \gamma \leq 1$ (this criterion

generalizes (14) of this paper to continuous measures) and thus by Theorem 2.1 of Coudron and Lerman (2012) (whose proof follows directly the one of Theorem 2 here) the solution of this minimization must be unique. On the other hand, any application of an arbitrary rotation of L^* (within $\mathbb{R}^D$) to the minimizer expressed in the RHS of (20) should also be a minimizer of the RHS of (20). We note that $\frac{1}{D-d}\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{P}_{\mathrm{L}^{*\perp}}^T$ is the only element in the domain of this minimization that is preserved under any rotation of L^*. Therefore, due to uniqueness, this can be the only solution of this minimization problem.

Let

$$\mathbb{H}_2 = \{\mathbf{Q} \in \mathbb{H} : \ \mathbf{Q}\mathbf{P}_{\mathrm{L}^*} = \mathbf{0}, \ \mathbf{Q} \succeq \mathbf{0} \text{ and } \mathrm{cond}(\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{Q}\mathbf{P}_{\mathrm{L}^{*\perp}}) \geq 2\}, \tag{61}$$

where $\mathbf{Q} \succeq \mathbf{0}$ denotes the positive semidefiniteness of $\mathbf{Q}$ and $\mathrm{cond}(\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{Q}\mathbf{P}_{\mathrm{L}^{*\perp}})$ denotes the condition number of this matrix, that is, the ratio between the largest and lowest eigenvalues of $\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{Q}\mathbf{P}_{\mathrm{L}^{*\perp}}$, or equivalently, the ratio between the top eigenvalue and the $(D-d)$th eigenvalue of $\mathbf{Q}$. Since $\hat{\mathbf{Q}}_I$ is the unique minimizer of (20) and $\hat{\mathbf{Q}}_I \notin \mathbb{H}_2$, then

$$c_1 := \min_{\mathbf{Q}\in\mathbb{H}_2} (F_I(\mathbf{Q}) - F_I(\hat{\mathbf{Q}}_I)) > 0. \tag{62}$$

We note that if $\mathbf{x}$ is a random variable sampled from μ and $\mathbf{Q} \in \mathbb{H}$ (so that $\|\mathbf{Q}\| \leq \|\mathbf{Q}\|_* = 1$), then $\|\mathbf{Q}\mathbf{x}\| \leq M$. Applying this fact, (62) and Hoeffding's inequality, we conclude that for any fixed $\mathbf{Q} \in \mathbb{H}_2$

$$F(\mathbf{Q}) - F(\hat{\mathbf{Q}}_I) > c_1 N/2 \ \text{ w.p. } 1 - \exp(-c_1^2 N/2M^2). \tag{63}$$

We also observe that

$$F(\mathbf{Q}_1) - F(\mathbf{Q}_2) \leq \|\mathbf{Q}_1 - \mathbf{Q}_2\| \sum_{i=1}^{N} \|\mathbf{x}\| \leq \|\mathbf{Q}_1 - \mathbf{Q}_2\| N\, M\,. \tag{64}$$

Combining (63) and (64), we obtain that for all $\mathbf{Q}$ in a ball of radius $r_1 := c_1/2M$ centered around a fixed element in $\mathbb{H}_2$: $F(\mathbf{Q}) - F(\hat{\mathbf{Q}}_I) > 0$ w.p. $1 - \exp(-c_1^2 N/2M^2)$.

We thus cover the compact space $\mathbb{H}_2$ by an r_1-net. Denoting the corresponding covering number by $N(\mathbb{H}_2, r_1)$ and using the above observation we note that w.p. $1 - N(\mathbb{H}_2, r_1)\exp(-c_1^2 N/2M^2)$

$$F(\mathbf{Q}) - F(\hat{\mathbf{Q}}_I) > 0 \quad \text{for all } \mathbf{Q} \in \mathbb{H}_2. \tag{65}$$

The definition of $\hat{\mathbf{Q}}_0$ (that is, (8)) implies that $F(\hat{\mathbf{Q}}_0) \leq F(\hat{\mathbf{Q}}_I)$. Combining this observation with (65), we conclude that w.h.p. $\hat{\mathbf{Q}}_0 \notin \mathbb{H}_2$. We also claim that $\hat{\mathbf{Q}}_0 \succeq \mathbf{0}$ (see, e.g., the proof of Lemma 14, which appears later). Since $\hat{\mathbf{Q}}_0 \notin \mathbb{H}_2$ and $\hat{\mathbf{Q}}_0 \succeq \mathbf{0}$, $\hat{\mathbf{Q}}_0$ satisfies the following property w.h.p.:

$$\mathrm{cond}(\mathbf{P}_{\mathrm{L}^{*\perp}}^T \hat{\mathbf{Q}}_0 \mathbf{P}_{\mathrm{L}^{*\perp}}^T) < 2. \tag{66}$$

Consequently, (9) holds w.h.p. (more precisely, w.p. $1 - N(\mathbb{H}_2, c_1/2M)\exp(-c_1^2 N/2M^2)$).

Next, we verify (10) w.h.p. as follows. Since $\hat{\mathbf{Q}}_0$ is symmetric and $\hat{\mathbf{Q}}_0\mathbf{P}_{\mathrm{L}^*} = \mathbf{0}$ (see (8)), then

$$\hat{\mathbf{Q}}_0 = \mathbf{P}_{\mathrm{L}^{*\perp}}\hat{\mathbf{Q}}_0\mathbf{P}_{\mathrm{L}^{*\perp}}. \tag{67}$$

Applying (67), basic inequalities of operators' norms and (66), we bound the RHS of (10) from above as follows:

$$\begin{aligned}
&\sqrt{2}\left\|\sum_{\mathbf{x}\in\mathcal{X}_0}\hat{\mathbf{Q}}_0\mathbf{x}\mathbf{x}^T\mathbf{P}_{\mathrm{L}^{*\perp}}/\|\hat{\mathbf{Q}}_0\mathbf{x}\|\right\| = \sqrt{2}\left\|\mathbf{P}_{\mathrm{L}^{*\perp}}\hat{\mathbf{Q}}_0\mathbf{P}_{\mathrm{L}^{*\perp}}\cdot\sum_{\mathbf{x}\in\mathcal{X}_0}\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\mathbf{x}^T\mathbf{P}_{\mathrm{L}^{*\perp}}/\|\hat{\mathbf{Q}}_0\mathbf{x}\|\right\| \\
\le&\sqrt{2}\cdot\left\|\mathbf{P}_{\mathrm{L}^{*\perp}}\hat{\mathbf{Q}}_0\mathbf{P}_{\mathrm{L}^{*\perp}}\right\|\cdot\left\|\sum_{\mathbf{x}\in\mathcal{X}_0}\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\mathbf{x}^T\mathbf{P}_{\mathrm{L}^{*\perp}}/\|\hat{\mathbf{Q}}_0\mathbf{x}\|\right\| \qquad (68)\\
\le&\sqrt{2}\cdot\lambda_{\max}(\mathbf{P}_{\mathrm{L}^{*\perp}}\hat{\mathbf{Q}}_0\mathbf{P}_{\mathrm{L}^{*\perp}})\cdot\|\sum_{\mathbf{x}\in\mathcal{X}_0}\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\mathbf{x}^T\mathbf{P}_{\mathrm{L}^{*\perp}}/\|\lambda_{\min}(\mathbf{P}_{\mathrm{L}^{*\perp}}\hat{\mathbf{Q}}_0\mathbf{P}_{\mathrm{L}^{*\perp}})\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\|\| \\
<&\sqrt{8}\left\|\sum_{\mathbf{x}\in\mathcal{X}_0}\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\mathbf{x}^T\mathbf{P}_{\mathrm{L}^{*\perp}}/\|\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\|\right\| = \max_{\mathbf{u}\in S^{D-1}\cap\mathrm{L}^{*\perp}}\sqrt{8}\mathbf{u}^T(\sum_{\mathbf{x}\in\mathcal{X}_0}\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\mathbf{x}^T\mathbf{P}_{\mathrm{L}^{*\perp}}/\|\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\|)\mathbf{u}.
\end{aligned}$$

Therefore to prove (10), we only need to prove that with high probability

$$\min_{\mathbf{Q}\in\mathbb{H},\mathbf{Q}\mathbf{P}_{\mathrm{L}^{*\perp}}=\mathbf{0}}\sum_{\mathbf{x}\in\mathcal{X}_1}\|\mathbf{Q}\mathbf{x}\| > \max_{\mathbf{u}\in S^{D-1}\cap\mathrm{L}^{*\perp}}\sqrt{8}\mathbf{u}^T(\sum_{\mathbf{x}\in\mathcal{X}_0}\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\mathbf{x}^T\mathbf{P}_{\mathrm{L}^{*\perp}}/\|\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\|)\mathbf{u}. \qquad (69)$$

We will prove that the LHS and RHS of (69) concentrates w.h.p. around the LHS and RHS of (16) respectively and consequently verify (69) w.h.p. Let ϵ_1 be the difference between the RHS and LHS of (69). Theorem 1 of Coudron and Lerman (2012) implies that the LHS of (69) is within distance $\epsilon_1/4$ to the RHS of (16) with probability $1-C\exp(-N/C)$ (where C is a constant depending on ϵ_1, μ and its parameters).

The concentration of the RHS of (16) can be concluded as follows. The spherical symmetry of $\mu_{0,\mathrm{L}^{*\perp}}$ implies that the expectation (w.r.t. μ_0) of $\sum_{\mathbf{x}\in\mathcal{X}_0}\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\mathbf{x}^T\mathbf{P}_{\mathrm{L}^{*\perp}}/\|\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\|$ is a scalar matrix within $\mathrm{L}^{*\perp}$, that is, it equals $\rho_\mu\,\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\mathbf{x}^T\mathbf{P}_{\mathrm{L}^{*\perp}}/\|\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\|$ for some $\rho_\mu\in\mathbb{R}$. We observe that

$$\mathbb{E}_{\mu_0}\operatorname{tr}(\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\mathbf{x}^T\mathbf{P}_{\mathrm{L}^{*\perp}}/\|\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\|) = \mathbb{E}_{\mu_0}\|\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\|$$

and thus conclude that $\rho_\mu = \mathbb{E}_{\mu_0}\|\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\|/(D-d)$. Therefore, for any $\mathbf{u}\in S^{D-1}\cap\mathrm{L}^{*\perp}$

$$\mathbb{E}_{\mu_0}\mathbf{u}^T(\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\mathbf{x}^T\mathbf{P}_{\mathrm{L}^{*\perp}}/\|\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\|)\mathbf{u} = \mathbb{E}_{\mu_0}\|\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\|/(D-d) = \int\|\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\|\,\mathrm{d}\mu_0(\mathbf{x})/(D-d). \qquad (70)$$

We thus conclude from (70) and Hoeffding's inequality that for any fixed $\mathbf{u}\in S^{D-1}\cap\mathrm{L}^{*\perp}$ the function $\sqrt{8}\mathbf{u}^T(\sum_{\mathbf{x}\in\mathcal{X}_0}\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\mathbf{x}^T\mathbf{P}_{\mathrm{L}^{*\perp}}/\|\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\|)\mathbf{u}$ is within distance $\epsilon_1/4$ to the RHS of (16) with probability $1-C\exp(-N/C)$ (where C is a constant depending on ϵ_1, μ and its parameters). Furthermore, applying ϵ-nets and covering (i.e., union bounds) arguments with regards to $S^{D-1}\cap\mathrm{L}^{*\perp}$, we obtain that for all $\mathbf{u}\in S^{D-1}\cap\mathrm{L}^{*\perp}$, $\sqrt{8}\mathbf{u}^T(\sum_{\mathbf{x}\in\mathcal{X}_0}\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\mathbf{x}^T\mathbf{P}_{\mathrm{L}^{*\perp}}/\|\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{x}\|)\mathbf{u}$ is within distance $\epsilon_1/2$ to the RHS of (16) with probability $1-C\exp(-N/C)$ (where C is a constant depending on ϵ_1, μ and its parameters). In particular, the RHS of (69) is within distance $\epsilon_1/2$ to the RHS of (16) with the same probability. We thus conclude (69) with probability $1-C'\exp(-N/C')$.

Similarly we can also prove (11), noting that the expectation (w.r.t. μ_0) of $\hat{\mathbf{Q}}_0\mathbf{x}\mathbf{x}^T\mathbf{P}_{\mathrm{L}^*}/\|\hat{\mathbf{Q}}_0\mathbf{x}\|$ is $\mathbf{0}$, since $\hat{\mathbf{Q}}_0\mathbf{x}/\|\hat{\mathbf{Q}}_0\mathbf{x}\|$ and $\mathbf{x}^T\mathbf{P}_{\mathrm{L}^*}$ are independent when $\mathbf{x}$ is restricted to lie in the complement of L^* (that is, $\mathbf{x}\in\mathcal{X}_0$).

If we remove the assumption of bounded supports (with radius M), then we need to replace Hoeffding's inequality with the Hoeffding-type inequality for sub-Gaussian measures of Proposition 5.10 of Vershynin (2012), where in this proposition $a_i = 1$ for all $1 \leq i \leq n$.

We emphasize that our probabilistic estimates are rather loose and can be interpreted as near-asymptotic; we thus did not fully specify their constants. We clarify this point for the probability estimate we have for (9), that is, $1 - N(\mathbb{H}_2, c_1/2M)\exp(-c_1^2 N/2M^2)$. Its constant $N(\mathbb{H}_2, r_1)$ can be bounded from above by the covering number $N(\mathbb{H}_0, r_1)$ of the larger set $\mathbb{H}_0 = \{\mathbf{Q} \in \mathbb{R}^{D\times D} : |\mathbf{Q}_{i,i}| \leq 1\}$, which is bounded from above by $(8/r_1)^{D(D-1)/2}$ (see, e.g., Lemma 5.2 of Vershynin, 2012). This is clearly a very loose estimate that cannot reveal interesting information, such as, the right dependence of N on D and d in order to obtain a sufficiently small probability.

At last, we explain why (14) holds with probability 1 if there are at least $2D-1$ outliers. We denote the set of outliers by $\{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_{N_0}\}$, where $N_0 \geq 2D-1$, and assume on the contrary that (14) holds with probability smaller than 1. Then, there exists a sequence $\{i_j\}_{j=1}^{D-1} \subset \{1, 2, 3, \cdots, N_0\}$ such that the subspace spanned by the $D-1$ points $\mathbf{y}_{i_1}, \mathbf{y}_{i_2}, \cdots, \mathbf{y}_{i_{D-1}}$ contains another outlier with positive probability. However, this is not true for haystack model and thus our claim is proved.

7.5.1 Proof of the Extension of Theorem 4 to the Asymmetric Case

We recall our assumptions that μ_0 is a sub-Gaussian distribution with covariance $\mathbf{\Sigma}_0$ and that $\hat{\mathbf{Q}}_I$ is unique. We follow the proof of Theorem 4 in §7.5 with the following changes. First of all, we replace the requirement

$$\mathrm{cond}(\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{Q}\mathbf{P}_{\mathrm{L}^{*\perp}}) \geq 2. \tag{71}$$

in (61) with the following one:

$$\mathrm{cond}(\mathbf{P}_{\mathrm{L}^{*\perp}}\mathbf{Q}\mathbf{P}_{\mathrm{L}^{*\perp}}) \geq 2 \cdot \mathrm{cond}(\mathbf{P}_{\mathrm{L}^{*\perp}}\hat{\mathbf{Q}}_I\mathbf{P}_{\mathrm{L}^{*\perp}}). \tag{72}$$

We note that (71) follows from (72) in the symmetric case. Indeed, in this case the expression of $\hat{\mathbf{Q}}_I$ in (60) implies that the RHS of (72) is 2. Similarly, instead of (66) we prove that

$$\mathrm{cond}(\mathbf{P}^T_{\mathrm{L}^{*\perp}}\hat{\mathbf{Q}}_0\mathbf{P}^T_{\mathrm{L}^{*\perp}}) < 2 \cdot \mathrm{cond}(\mathbf{P}_{\mathrm{L}^{*\perp}}\hat{\mathbf{Q}}_I\mathbf{P}_{\mathrm{L}^{*\perp}}).$$

Second of all, in the third inequality of (68) the term

$$\sqrt{2}\,\lambda_{\max}(\mathbf{P}_{\mathrm{L}^{*\perp}}\hat{\mathbf{Q}}_0\mathbf{P}_{\mathrm{L}^{*\perp}})/\lambda_{\max}(\mathbf{P}_{\mathrm{L}^{*\perp}}\hat{\mathbf{Q}}_0\mathbf{P}_{\mathrm{L}^{*\perp}})$$

needs to be bounded above by $\sqrt{8}\,\mathrm{cond}(\mathbf{P}_{\mathrm{L}^{*\perp}}\hat{\mathbf{Q}}_I\mathbf{P}_{\mathrm{L}^{*\perp}})$, instead of $\sqrt{8}$. We can thus conclude the revised theorem, in particular, the last modification in the proof clarifies why we need to multiply the RHS of (16) by $\mathrm{cond}(\mathbf{P}_{\mathrm{L}^{*\perp}}\hat{\mathbf{Q}}_I\mathbf{P}_{\mathrm{L}^{*\perp}})$, which is the ratio between the largest eigenvalue of $\mathbf{P}_{\mathrm{L}^{*\perp}}\hat{\mathbf{Q}}_I\mathbf{P}_{\mathrm{L}^{*\perp}}$ and the $(D-d)$th eigenvalue of $\mathbf{P}_{\mathrm{L}^{*\perp}}\hat{\mathbf{Q}}_I\mathbf{P}_{\mathrm{L}^{*\perp}}$.

7.6 Proof of Theorem 5

This proof follows ideas of Lerman et al. (2012). We bound from below the LHS of (7) by applying (A.15) of Lerman et al. (2012) as follows

$$\min_{\mathbf{Q}\in\mathbb{H},\mathbf{Q}\mathbf{P}_{\mathrm{L}^{*\perp}}=\mathbf{0}} \sum_{\mathbf{x}\in\mathcal{X}_1} \|\mathbf{Q}\mathbf{x}\| \geq \frac{1}{\sqrt{d}} \min_{\mathbf{v}\in\mathrm{L}^*,\|\mathbf{v}\|=1} \sum_{\mathbf{x}\in\mathcal{X}_1} |\mathbf{v}^T\mathbf{x}|. \tag{73}$$

We denote the number of inliers sampled from μ_1 by N_1 and the number of outliers sampled from μ_0 by $N_0(=N-N_1)$. We bound from below w.h.p. the RHS of (73) by applying Lemma B.2 of Lerman et al. (2012) in the following way:

$$\frac{1}{\sqrt{d}} \min_{\mathbf{v}\in \mathrm{L}^*, \|\mathbf{v}\|=1} \sum_{\mathbf{x}\in\mathcal{X}_1} |\mathbf{v}^T\mathbf{x}| \geq \frac{\sigma_1}{d}\Big(\sqrt{2/\pi}N_1 - 2\sqrt{N_1 d} - t\sqrt{N_1}\Big) \text{ w.p. } 1-e^{-t^2/2}. \tag{74}$$

By following the proof of Lemma B.2 of Lerman et al. (2012) we bound from above w.h.p. the RHS of (7) as follows

$$\max_{\mathbf{v}\in \mathrm{L}^*, \|\mathbf{v}\|=1} \sum_{\mathbf{x}\in\mathcal{X}_0} |\mathbf{v}^T\mathbf{x}| \leq \frac{\sigma_0}{\sqrt{D}}\Big(\sqrt{2/\pi}N_0 + 2\sqrt{N_0 d} + t\sqrt{N_0}\Big) \text{ w.p. } 1-e^{-t^2/2}. \tag{75}$$

We need to show w.h.p. that the RHS of (75) is strictly less than the RHS of (74). We note that Hoeffding's inequality implies that

$$N_1 > \alpha_1 N/2 \text{ w.p. } 1-e^{-\alpha_1^2 N/2} \text{ and } |N_0 - \alpha_0 N| < \alpha_0 N/2 \text{ w.p. } 1-2e^{-\alpha_0^2 N/2}. \tag{76}$$

Furthermore, (18) and (76) imply that

$$d < N_1/4 \text{ w.p. } 1-e^{-\alpha_1^2 N/2} \text{ and } d < N_0/4 \text{ w.p. } 1-e^{-\alpha_0^2 N/2}. \tag{77}$$

Substituting $t = \sqrt{N_1}/10$ ($> \sqrt{\alpha_1 N}/20$ w.p. $1-e^{-\alpha_1^2 N/2}$) in (74) and $t = \sqrt{N_0}/10$ ($> \sqrt{\alpha_0 N}/20$ w.p. $1-2e^{-\alpha_0^2 N/2}$) in (75) and combining (17) and (73)-(77), we obtain that (7) holds w.p. $1-e^{-\alpha_1^2 N/2} - 2e^{-\alpha_0^2 N/2} - e^{-\alpha_1 N/800} - e^{-\alpha_0 N/800}$. We can similarly obtain that (6) holds with the same probability.

7.6.1 Proof of the Extension of Theorem 5 to the Asymmetric Case

We assume the generalized needle-haystack model of §2.6.2. The proof of Theorem 5 in §7.6 immediately extends to this model, where σ_0 in the RHS of (75) needs to be replaced with $\sqrt{\lambda_{\max}(\mathbf{\Sigma}_0)}$ (recall that $\lambda_{\max}(\mathbf{\Sigma}_0)$ denotes the largest eigenvalue of $\mathbf{\Sigma}_0$). Consequently, Theorem 5 still holds in this case when replacing σ_0 in the RHS of (17) with $\sqrt{\lambda_{\max}(\mathbf{\Sigma}_0)}$.

7.7 Proof of Theorem 6

We first establish the following lemma.

Lemma 14 *The minimizer of $F(\mathbf{Q})$, $\hat{\mathbf{Q}}$, is a semi-definite positive matrix.*

Proof We assume that $\hat{\mathbf{Q}}$ has some negative eigenvalues and show that this assumption contradicts the defining property of $\hat{\mathbf{Q}}$, that is, being the minimizer of $F(\mathbf{Q})$. We denote the eigenvalue decomposition of $\hat{\mathbf{Q}}$ by $\hat{\mathbf{Q}} = \mathbf{V}_{\hat{\mathbf{Q}}}\mathbf{\Sigma}_{\hat{\mathbf{Q}}}\mathbf{V}_{\hat{\mathbf{Q}}}^T$ and define $\mathbf{\Sigma}_{\hat{\mathbf{Q}}}^+ = \max(\mathbf{\Sigma}_{\hat{\mathbf{Q}}}, 0)$ and $\hat{\mathbf{Q}}^+ = \mathbf{V}_{\hat{\mathbf{Q}}}\mathbf{\Sigma}_{\hat{\mathbf{Q}}}^+\mathbf{V}_{\hat{\mathbf{Q}}}^T / \operatorname{tr}(\mathbf{\Sigma}_{\hat{\mathbf{Q}}}^+) \in \mathbb{H}$. Then $\operatorname{tr}(\mathbf{\Sigma}_{\hat{\mathbf{Q}}}^+) > \operatorname{tr}(\mathbf{\Sigma}_{\hat{\mathbf{Q}}}) = \operatorname{tr}(\hat{\mathbf{Q}}) = 1$ and for any $\mathbf{x}\in\mathbb{R}^D$ we have

$$\|\hat{\mathbf{Q}}^+\mathbf{x}\| < \operatorname{tr}(\mathbf{\Sigma}_{\hat{\mathbf{Q}}}^+)\|\hat{\mathbf{Q}}^+\mathbf{x}\| = \|\mathbf{\Sigma}_{\hat{\mathbf{Q}}}^+(\mathbf{V}_{\hat{\mathbf{Q}}}^T\mathbf{x})\| \leq \|\mathbf{\Sigma}_{\hat{\mathbf{Q}}}(\mathbf{V}_{\hat{\mathbf{Q}}}^T\mathbf{x})\| = \|\hat{\mathbf{Q}}\mathbf{x}\|.$$

Summing it over all $\mathbf{x}\in\mathcal{X}$, we conclude the contradiction $F(\hat{\mathbf{Q}}^+) < F(\hat{\mathbf{Q}})$.

■

In order to prove Theorem 6 we first notice that by definition and the connection of γ_0, γ_0 with second derivative of $F(\mathbf{Q})$

$$F_{\mathcal{X}}(\tilde{\mathbf{Q}}) - F_{\mathcal{X}}(\tilde{\mathbf{Q}}) \geq N\gamma_0 \|\tilde{\mathbf{Q}} - \hat{\mathbf{Q}}\|_F^2, \tag{78}$$

and

$$F_{\mathcal{X}}(\tilde{\mathbf{Q}}) - F_{\mathcal{X}}(\tilde{\mathbf{Q}}) \geq N\gamma_0' \|\tilde{\mathbf{Q}} - \hat{\mathbf{Q}}\|^2. \tag{79}$$

Next, we observe that

$$|F_{\mathcal{X}}(\hat{\mathbf{Q}}) - F_{\tilde{\mathcal{X}}}(\hat{\mathbf{Q}})| \leq \sum_{i=1}^{N} \left| \|\hat{\mathbf{Q}}\tilde{\mathbf{x}}_i\| - \|\hat{\mathbf{Q}}\mathbf{x}_i\| \right| \leq \sum_{i=1}^{N} \|\hat{\mathbf{Q}}(\tilde{\mathbf{x}}_i - \mathbf{x}_i)\| \leq \sum_{i=1}^{N} \|\tilde{\mathbf{x}}_i - \mathbf{x}_i\| \leq \sum_{i=1}^{N} \epsilon_i$$

and similarly $|F_{\mathcal{X}}(\tilde{\mathbf{Q}}) - F_{\tilde{\mathcal{X}}}(\tilde{\mathbf{Q}})| \leq \sum_{i=1}^{N} \epsilon_i$. Therefore,

$$\begin{aligned} F_{\mathcal{X}}(\tilde{\mathbf{Q}}) - F_{\mathcal{X}}(\hat{\mathbf{Q}}) &= (F_{\tilde{\mathcal{X}}}(\tilde{\mathbf{Q}}) - F_{\tilde{\mathcal{X}}}(\hat{\mathbf{Q}})) + (F_{\mathcal{X}}(\tilde{\mathbf{Q}}) - F_{\tilde{\mathcal{X}}}(\tilde{\mathbf{Q}})) + (F_{\tilde{\mathcal{X}}}(\hat{\mathbf{Q}}) \\ -F_{\mathcal{X}}(\hat{\mathbf{Q}})) &\leq 0 + |F_{\mathcal{X}}(\tilde{\mathbf{Q}}) - F_{\tilde{\mathcal{X}}}(\tilde{\mathbf{Q}})| + |F_{\tilde{\mathcal{X}}}(\hat{\mathbf{Q}}) - F_{\mathcal{X}}(\hat{\mathbf{Q}})| \leq 2\sum_{i=1}^{N} \epsilon_i. \end{aligned} \tag{80}$$

Therefore (23) follows from (78), (79) and (80). Applying the Davis-Kahan perturbation Theorem (Davis and Kahan, 1970) to (23), we conclude (24).

7.7.1 Implication of Theorem 6 to Dimension Estimation

Theorem 6 implies that we may properly estimate the dimension of the underlying subspace for low-dimensional data with sufficiently small perturbation. We make this statement more precise by assuming the setting of Theorem 6 and further assuming that $\hat{\mathbf{Q}}$ is a low-rank matrix with $\ker(\hat{\mathbf{Q}}) = \mathrm{L}^*$. We note that the $(D-d+1)$st eigenvalue of $\hat{\mathbf{Q}}$ is 0. Thus applying the following eigenvalue stability inequality (Tao, 2012, (1.63)):

$$|\lambda_i(\mathbf{A}+\mathbf{B}) - \lambda_i(\mathbf{A})| \leq \|\mathbf{B}\|, \tag{81}$$

we obtain that the $(D-d+1)$st eigenvalue of $\tilde{\mathbf{Q}}$ is smaller than $\sqrt{2\sum_{i=1}^{N} \epsilon_i/\gamma_0}$, and the $(D-d)$th eigengap of $\tilde{\mathbf{Q}}$ is larger than $\nu_{D-d} - 2\sqrt{2\sum_{i=1}^{N} \epsilon_i/\gamma_0}$ (recall that ν_{D-d} is the $(D-d)$th eigengap of $\hat{\mathbf{Q}}$). This means that when the noise is small and the conditions of Theorem 1 hold, then we can estimate the dimension of the underlying subspace for $\tilde{\mathcal{X}}$ from the number of small eigenvalues.

7.7.2 Improved Bounds in a Restricted Setting

We assume that $\epsilon_i = O(\epsilon)$ for all $1 \leq i \leq N$, where ϵ is sufficiently small, and further assume that $\mathrm{rank}(\hat{\mathbf{Q}}) = D$. We show that in this special case the norm of $\hat{\mathbf{Q}} - \tilde{\mathbf{Q}}$ is of order $O(\epsilon)$ instead of order $O(\sqrt{\epsilon})$ that is specified in Theorem 6.

We note that since $\hat{\mathbf{Q}}$ is of full rank, then the first and second directional derivative of F are well-defined in a sufficiently small neighborhood around $\hat{\mathbf{Q}}$. Therefore, if $\mathbf{\Delta} \in \mathbb{R}^{D\times D}$ and $\|\mathbf{\Delta}\|$ is sufficiently small then

$$F'_{\mathcal{X}}(\hat{\mathbf{Q}}) - F'_{\mathcal{X}}(\hat{\mathbf{Q}} + \mathbf{\Delta}) = O(\|\mathbf{\Delta}\|). \tag{82}$$

Furthermore, we note by basic calculations that

$$F'_{\mathcal{X}}(\mathbf{Q}) - F'_{\tilde{\mathcal{X}}}(\mathbf{Q}) = O(\epsilon). \tag{83}$$

Combining (83) with the following facts: $F'_{\mathcal{X}}(\hat{\mathbf{Q}}) = 0$ and $F'_{\tilde{\mathcal{X}}}(\tilde{\mathbf{Q}}) = 0$, we obtain that

$$F'_{\mathcal{X}}(\hat{\mathbf{Q}}) - F'_{\mathcal{X}}(\tilde{\mathbf{Q}}) = F'_{\tilde{\mathcal{X}}}(\tilde{\mathbf{Q}}) - F'_{\mathcal{X}}(\tilde{\mathbf{Q}}) = O(\epsilon). \tag{84}$$

At last, the combination of (82) and (84) implies that $\|\hat{\mathbf{Q}} - \tilde{\mathbf{Q}}\| = O(\epsilon)$. Clearly, the spectral norm of $\hat{\mathbf{Q}} - \tilde{\mathbf{Q}}$ can be replaced with any other norm, in particular, the Frobenius norm.

7.8 Proof of Proposition 7

We recall the function F_I, which was defined in (19), and the notation ${F_I}_{,1}''(\mathbf{Q}, \mathbf{\Delta})$ should be clear, where now F_I replaces F.

The law of large numbers implies that ${F_1}''(\mathbf{Q}, \mathbf{\Delta})/N \to {F_I}''(\mathbf{Q}, \mathbf{\Delta})$ almost surely for any $\mathbf{\Delta}$ and $\mathbf{Q}$ (see also related bounds in Coudron and Lerman 2012). Since $\mathbf{Q}$ and $\mathbf{\Delta}$ lie in compact space, we conclude (26) for γ_0 and c_0; the proof is identical for γ'_0 and c'_0.

7.9 Proof of Theorem 8

The theorem follows from the observation that $0 \le F(\mathbf{Q}) - F_\delta(\mathbf{Q}) \le N\delta/2$ for all $\mathbf{Q} \in \mathbb{H}$ and the proof of Theorem 6.

7.10 Proof of Theorem 9

It is sufficient to verify that

$$\text{If } \tilde{\mathbf{A}} \in \mathbb{R}^{D\times D} \text{ with } \mathrm{Im}(\tilde{\mathbf{A}}) = \mathrm{L}^*, \text{ then } L(\tilde{\mathbf{A}} + \eta\mathbf{I}) \to \infty \text{ as } \eta \to 0. \tag{85}$$

Indeed, since $L(\mathbf{A})$ is a continuous function, (85) implies that $L(\tilde{\mathbf{A}})$ is undefined (or infinite) and therefore $\tilde{\mathbf{A}}$ is not the minimizer of (28) as stated in Theorem 9.

We fix $a_1 < \lim_{x\to\infty} xu(x)$ and note that Condition D_0 (w.r.t. L^*) implies that

$$|\mathcal{X}_0|/N > (D-d)/a_1. \tag{86}$$

Condition M implies that there exists x_1 such that for any $x > x_1$: $xu(x) \ge a_1$ and therefore (recalling that $u = \rho'$) $\rho(x) \ge a_1 \ln(x - x_1)/2 + u(x_1)/2$. Thus for any $\mathbf{x}_i \in \mathcal{X}_0$, we have

$$\rho(\mathbf{x}_i^T(\tilde{\mathbf{A}} + \eta\mathbf{I})^{-1}\mathbf{x}_i) \ge a_1 \ln(1/\eta - x_1)/2 + C_i \text{ for some constant } C_i \equiv C_i(\mathbf{x}_i, \tilde{\mathbf{A}}) \tag{87}$$

and

$$\frac{N}{2}\log(\det(\mathbf{A})) \le NC_0 + (D-d)/2\ln(\eta) \text{ for some } C_0 \equiv C_0(\tilde{\mathbf{A}}). \tag{88}$$

Equation (85) thus follows from (86)-(88) and the theorem is concluded.

7.11 Proof of Theorem 10

The derivative of the energy function in the RHS of (32) is $\mathbf{Q}\mathbf{X}^T\mathbf{X} + \mathbf{X}^T\mathbf{X}\mathbf{Q}$. Using the argument establishing (36) and the fact that $\hat{\mathbf{Q}}_2$ is the minimizer of (32), we conclude that $\mathbf{Q}\mathbf{X}^T\mathbf{X} + \mathbf{X}^T\mathbf{X}\mathbf{Q}$ is a scalar matrix. We then conclude (33) by using the argument establishing (37) as well as the following two facts: $\mathrm{tr}(\hat{\mathbf{Q}}_2) = 1$ and $\mathbf{X}$ is full rank (so the inverse of $\mathbf{X}^T\mathbf{X}$ exists).

7.12 Proof of Theorem 11

We frequently use here some of the notation introduced in §4.1, in particular, $I(\mathbf{Q})$, $\mathrm{L}(\mathbf{Q})$ and $T(\mathbf{Q})$. We will first prove that $F(\mathbf{Q}_k) \geq F(\mathbf{Q}_{k+1})$ for all $k \geq 1$. For this purpose, we use the convex quadratic function:

$$G(\mathbf{Q}, \mathbf{Q}^*) = \frac{1}{2} \sum_{\substack{i=1 \\ i \notin I(\mathbf{Q}^*)}}^{N} \left(\|\mathbf{Q}\mathbf{x}_i\|^2 / \|\mathbf{Q}^*\mathbf{x}_i\| + \|\mathbf{Q}^*\mathbf{x}_i\| \right).$$

Following the same derivation of (44) and (36), we obtain that

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{Q}} G(\mathbf{Q}, \mathbf{Q}_k)\big|_{\mathbf{Q}=\mathbf{Q}_{k+1}} = \left(\mathbf{Q}_{k+1} \left(\sum_{\substack{i=1 \\ i \notin I(\mathbf{Q}_k)}}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\|\mathbf{Q}_k \mathbf{x}_i\|} \right) + \left(\sum_{\substack{i=1 \\ i \notin I(\mathbf{Q}_k)}}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\|\mathbf{Q}_k \mathbf{x}_i\|} \right) \mathbf{Q}_{k+1} \right) /2.$$

We let $\mathbf{A}_k = \sum_{i=1, i \notin I(\mathbf{Q}_k)}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\|\mathbf{Q}_k \mathbf{x}_i\|}$, $c_k = \mathbf{P}_{\mathrm{L}(\mathbf{Q}_k)^\perp} \mathbf{A}_k^{-1} \mathbf{P}_{\mathrm{L}(\mathbf{Q}_k)^\perp}$ and for any symmetric $\mathbf{\Delta} \in \mathbb{R}^{D \times D}$ with $\mathrm{tr}(\mathbf{\Delta}) = 0$ and $\mathbf{P}_{\mathrm{L}(\mathbf{Q}_k)} \mathbf{\Delta} = \mathbf{0}$ we let $\mathbf{\Delta}_0 = \tilde{\mathbf{P}}^T_{\mathrm{L}(\mathbf{Q}_k)^\perp} \mathbf{\Delta} \tilde{\mathbf{P}}_{\mathrm{L}(\mathbf{Q}_k)^\perp}$. We note that

$$\begin{aligned} \mathrm{tr}(\mathbf{\Delta}_0) &= \langle \mathbf{\Delta}_0, \mathbf{I} \rangle_F = \left\langle \tilde{\mathbf{P}}^T_{\mathrm{L}(\mathbf{Q}_k)^\perp} \mathbf{\Delta} \tilde{\mathbf{P}}_{\mathrm{L}(\mathbf{Q}_k)^\perp}, \mathbf{I} \right\rangle_F = \left\langle \mathbf{\Delta}, \tilde{\mathbf{P}}_{\mathrm{L}(\mathbf{Q}_k)^\perp} \tilde{\mathbf{P}}^T_{\mathrm{L}(\mathbf{Q}_k)^\perp} \right\rangle_F \\ &= \left\langle \mathbf{\Delta}, \mathbf{I} - \mathbf{P}_{\mathrm{L}(\mathbf{Q}_k)} \right\rangle_F = \langle \mathbf{\Delta}, \mathbf{I} \rangle_F - \left\langle \mathbf{\Delta}, \mathbf{P}_{\mathrm{L}(\mathbf{Q}_k)} \right\rangle_F = \langle \mathbf{\Delta}, \mathbf{I} \rangle_F = \mathrm{tr}(\mathbf{\Delta}) = 0. \end{aligned}$$

Consequently, we establish that the derivative of $G(\mathbf{Q}, \mathbf{Q}_k)$ at $\mathbf{Q}_{k+1}$ in the direction $\mathbf{\Delta}$ is zero as follows.

$$\begin{aligned} &\langle (\mathbf{Q}_{k+1}\mathbf{A}_k + \mathbf{A}_k\mathbf{Q}_{k+1})/2, \mathbf{\Delta} \rangle_F = \langle \mathbf{Q}_{k+1}\mathbf{A}_k, \mathbf{\Delta} \rangle_F = c_k \left\langle \mathbf{P}_{\mathrm{L}(\mathbf{Q}_k)^\perp} \mathbf{A}_k^{-1} \mathbf{P}_{\mathrm{L}(\mathbf{Q}_k)^\perp} \mathbf{A}_k, \mathbf{\Delta} \right\rangle_F \\ =& c_k \left\langle \mathbf{P}_{\mathrm{L}(\mathbf{Q}_k)^\perp} \mathbf{A}_k^{-1} \mathbf{P}_{\mathrm{L}(\mathbf{Q}_k)^\perp} \mathbf{A}_k, \tilde{\mathbf{P}}_{\mathrm{L}(\mathbf{Q}_k)^\perp} \mathbf{\Delta}_0 \tilde{\mathbf{P}}^T_{\mathrm{L}(\mathbf{Q}_k)^\perp} \right\rangle_F \\ =& c_k \left\langle (\tilde{\mathbf{P}}^T_{\mathrm{L}(\mathbf{Q}_k)^\perp} \mathbf{A}_k^{-1} \tilde{\mathbf{P}}_{\mathrm{L}(\mathbf{Q}_k)^\perp})(\tilde{\mathbf{P}}^T_{\mathrm{L}(\mathbf{Q}_k)^\perp} \mathbf{A}_k \tilde{\mathbf{P}}_{\mathrm{L}(\mathbf{Q}_k)^\perp}), \mathbf{\Delta}_0 \right\rangle_F = c_k \langle \mathbf{I}, \mathbf{\Delta}_0 \rangle_F = 0. \end{aligned}$$

This and the strict convexity of $G(\mathbf{Q}, \mathbf{Q}_k)$ (which follows from $\mathrm{Sp}(\{\mathbf{x}_i\}_{i \notin I(\mathbf{Q}_k)}) = \mathbb{R}^D$ using (14)) imply that $\mathbf{Q}_{k+1}$ is the unique minimizer of $G(\mathbf{Q}, \mathbf{Q}_k)$ among all $\mathbf{Q} \in \mathbb{H}$ such that $\mathbf{P}_{\mathrm{L}(\mathbf{Q}_k)}\mathbf{Q} = \mathbf{0}$.

Combining this with the following two facts: $\mathbf{Q}_{k+1}\mathbf{x}_i = 0$ for any $i \in I(\mathbf{Q}_k)$ and $G(\mathbf{Q}_k, \mathbf{Q}_k) = F(\mathbf{Q}_k)$, we conclude that

$$F(\mathbf{Q}_{k+1}) = \sum_{i \notin I(\mathbf{Q}_k)} \|\mathbf{Q}_{k+1}\mathbf{x}_i\| = \sum_{i \notin I(\mathbf{Q}_k)} \frac{\|\mathbf{Q}_{k+1}\mathbf{x}_i\|\|\mathbf{Q}_k\mathbf{x}_i\|}{\|\mathbf{Q}_k\mathbf{x}_i\|}$$
$$\leq \sum_{i \notin I(\mathbf{Q}_k)} \frac{\|\mathbf{Q}_{k+1}\mathbf{x}_i\|^2 + \|\mathbf{Q}_k\mathbf{x}_i\|^2}{2\|\mathbf{Q}_k\mathbf{x}_i\|} = G(\mathbf{Q}_{k+1}, \mathbf{Q}_k) \leq G(\mathbf{Q}_k, \mathbf{Q}_k) = F(\mathbf{Q}_k). \tag{89}$$

Since F is positive, $F(\mathbf{Q}_k)$ converges and

$$F(\mathbf{Q}_k) - F(\mathbf{Q}_{k+1}) \to 0 \text{ as } k \to \infty. \tag{90}$$

Applying (89) we also have that

$$F(\mathbf{Q}_k) - F(\mathbf{Q}_{k+1}) \geq G(\mathbf{Q}_k, \mathbf{Q}_k) - G(\mathbf{Q}_{k+1}, \mathbf{Q}_k) = \frac{1}{2} \sum_{i \notin I(\mathbf{Q}_k)} \|(\mathbf{Q}_k - \mathbf{Q}_{k+1})\mathbf{x}_i\|^2 / \|\mathbf{Q}_k\mathbf{x}_i\|. \tag{91}$$

We note that if $\mathbf{Q}_k \neq \mathbf{Q}_{k+1}$, then $\mathrm{Sp}(\{\mathbf{x}_i\}_{i \notin I(\mathbf{Q}_k)}) = \mathbb{R}^D \supset \ker(\mathbf{Q}_k - \mathbf{Q}_{k+1})$ and $1/\|\mathbf{Q}_k\mathbf{x}_i\| \geq 1/\max_i \|\mathbf{x}_i\|$. Combining this observation with (90) and (91) we obtain that

$$\|\mathbf{Q}_k - \mathbf{Q}_{k+1}\|_2 \to 0 \text{ as } k \to \infty. \tag{92}$$

Since for all $k \in \mathbb{N}$, $\mathbf{Q}_k$ is nonnegative (this follows from its defining formula (39)) and $\mathrm{tr}(\mathbf{Q}_k) = 1$, the sequence $\{\mathbf{Q}_k\}_{k\in\mathbb{N}}$ lies in a compact space (of nonnegative matrices) and it thus has a converging subsequence. Assume a subsequence of $\{\mathbf{Q}_k\}_{k\in\mathbb{N}}$, which converges to $\tilde{\mathbf{Q}}$. We claim the following property of $\tilde{\mathbf{Q}}$:

$$\tilde{\mathbf{Q}} = \operatorname*{arg\,min}_{\mathbf{Q} \in \mathbb{H}_0} F(\mathbf{Q}), \text{ where } \mathbb{H}_0 := \{\mathbf{Q} \in \mathbb{H} : \ker \mathbf{Q} \supseteq \mathrm{L}(\tilde{\mathbf{Q}})\}. \tag{93}$$

In order to prove (93), we note that (89) and the convergence of the subsequence imply that $F(\tilde{\mathbf{Q}}) = F(T(\tilde{\mathbf{Q}}))$. Combining this with (89) (though replacing $\mathbf{Q}_k$ and $\mathbf{Q}_{k+1}$ in (89) with $\tilde{\mathbf{Q}}$ and $T(\tilde{\mathbf{Q}})$ respectively) we get that $G(T(\tilde{\mathbf{Q}}), \tilde{\mathbf{Q}}) = G(\tilde{\mathbf{Q}}, \tilde{\mathbf{Q}})$. We conclude that $T(\tilde{\mathbf{Q}}) = \tilde{\mathbf{Q}}$ from this observation and the following three facts: 1) $\mathbf{Q} = \tilde{\mathbf{Q}}$ is the unique minimizer of $G(\mathbf{Q}, \tilde{\mathbf{Q}})$ among all $\mathbf{Q} \in \mathbb{H}$, 2) $\mathbf{P}_{\mathrm{L}(\tilde{\mathbf{Q}})}\tilde{\mathbf{Q}} = \mathbf{0}$, 3) $\mathbf{Q} = T(\tilde{\mathbf{Q}})$ is the unique minimizer of $G(\mathbf{Q}, \tilde{\mathbf{Q}})$ among all $\mathbf{Q} \in \mathbb{H}$ such that $\mathbf{P}_{\mathrm{L}(\tilde{\mathbf{Q}})}\mathbf{Q} = \mathbf{0}$ (we remark that $F(\mathbf{Q})$ is strictly convex in $\mathbb{H}$ and consequently also in $\mathbb{H}_0$ by Theorem 2). Therefore, for any symmetric $\mathbf{\Delta} \in \mathbb{R}^{D\times D}$ with $\mathrm{tr}(\mathbf{\Delta}) = 0$ and $\mathbf{P}_{\mathrm{L}(\tilde{\mathbf{Q}})}\mathbf{\Delta} = \mathbf{0}$, the directional derivative at $\tilde{\mathbf{Q}}$ is 0:

$$0 = \left\langle \mathbf{\Delta}, \frac{\mathrm{d}}{\mathrm{d}\mathbf{Q}} G(\mathbf{Q}, \tilde{\mathbf{Q}})\big|_{\mathbf{Q}=\tilde{\mathbf{Q}}} \right\rangle_F = \left\langle \mathbf{\Delta}, \tilde{\mathbf{Q}} \sum_{i \notin I(\tilde{\mathbf{Q}})} \frac{\mathbf{x}_i\mathbf{x}_i^T}{\|\tilde{\mathbf{Q}}\mathbf{x}_i\|} \right\rangle_F. \tag{94}$$

We note that (94) is the corresponding directional derivative of $F(\mathbf{Q})$ when restricted to $\mathbf{Q} \in \mathbb{H}_0$ and we thus conclude (93).

Next, we will prove that $\{\mathbf{Q}_k\}_{k\in\mathbb{N}}$ converge to $\tilde{\mathbf{Q}}$ by proving that there are only finite choices for $\tilde{\mathbf{Q}}$. In view of (93) and the strict convexity of $F(\mathbf{Q})$ in $\mathbb{H}_0$, any limit $\tilde{\mathbf{Q}}$ (of

a subsequence as above) is uniquely determined by $I(\tilde{\mathbf{Q}})$. Since the number of choices for $I(\tilde{\mathbf{Q}})$ is finite (independently of $\tilde{\mathbf{Q}}$), the number of choices for $\tilde{\mathbf{Q}}$ is finite. That is, $\mathcal{Y} := \{\mathbf{Q} \in \mathbb{H} : F(\mathbf{Q}) = F(T(\mathbf{Q}))\}$ is a finite set. Combining this with (92) and the convergence analysis of the sequence $\{\mathbf{Q}_k\}_{k\in\mathbb{N}}$ (see Ostrowski, 1966, Theorem 28.1), we conclude that $\{\mathbf{Q}_k\}_{k\in\mathbb{N}}$ converges to $\tilde{\mathbf{Q}}$.

At last, we assume that $\tilde{\mathbf{Q}}\mathbf{x}_i \neq \mathbf{0}$ for all $1 \leq i \leq N$. We note that $I(\tilde{\mathbf{Q}}) = \emptyset$ and thus $\tilde{\mathbf{Q}} = \hat{\mathbf{Q}}$ by (93). The proof for the rate of convergence follows the analysis of generalized Weiszfeld's method by Chan and Mulet (1999) (in particular see §6 of that work). We practically need to verify Hypotheses 4.1 and 4.2 (see §4 of that work) and replace the functions F and G in that work by $F(\mathbf{Q})$ and

$$\tilde{G}(\mathbf{Q}, \mathbf{Q}^*) = \sum_{i=1}^{N} \left(\|\mathbf{Q}\mathbf{x}_i\|^2 / \|\mathbf{Q}^*\mathbf{x}_i\| + \|\mathbf{Q}^*\mathbf{x}_i\| \right)$$

respectively. We note that the functions $\tilde{G}$ and G (defined earlier in this work) coincide in the following way: $\tilde{G}(\mathbf{Q}, \mathbf{Q}_k) = G(\mathbf{Q}, \mathbf{Q}_k)$ for any $k \in \mathbb{N}$ (this follows from the fact that $\mathbf{Q}_k\mathbf{x}_i \neq \mathbf{0}$ for all $k \in \mathbb{N}$ and $1 \leq i \leq N$; indeed, otherwise for some i, $\mathbf{Q}_j\mathbf{x}_i = \mathbf{0}$ for $j \geq k$ by (39) and this leads to the contradiction $\hat{\mathbf{Q}}\mathbf{x}_i = \mathbf{0}$). We remark that even though Chan and Mulet (1999) consider vector-valued functions, their proof generalizes to matrix-valued functions as here. Furthermore, we can replace the global properties of Hypotheses 4.1 and 4.2 of Chan and Mulet (1999) by the local properties in $B(\hat{\mathbf{Q}}, \delta_0)$ for any $\delta_0 > 0$, since the convergence of $\mathbf{Q}_k$ implies the existence of $K_0 > 0$ such that $\mathbf{Q}_k \in B(\hat{\mathbf{Q}}, \delta_0)$ for all $k > K_0$. In particular, there is no need to check condition 2 in Hypothesis 4.1. Condition 1 in Hypothesis 4.1 holds since $F(\mathbf{Q})$ is twice differentiable in $B(\hat{\mathbf{Q}}, \delta_0)$ (which follows from the assumption on the limit $\tilde{\mathbf{Q}} \equiv \hat{\mathbf{Q}}$ and the continuity of the derivative). Conditions 1-3 in Hypothesis 4.2 are verified by the fact that C of Hypothesis 4.2 satisfies $C(\mathbf{Q}^*) = \sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^T / \|\mathbf{Q}^*\mathbf{x}_i\|$ and $\mathbf{Q}^*\mathbf{x}_i \neq \mathbf{0}$ when $\mathbf{Q}^* \in B(\hat{\mathbf{Q}}, \delta_0)$. Condition 3 in Hypothesis 3.1 and condition 4 in Hypothesis 4.2 are easy to check.

7.13 Proof of Theorem 12

The proof follows from the second part of the proof of Theorem 11, while using instead of $\tilde{G}(\mathbf{Q}, \mathbf{Q}^*)$ the function

$$G_\delta(\mathbf{Q}, \mathbf{Q}^*) = \frac{1}{2} \sum_{i=1, \|\mathbf{Q}^*\mathbf{x}_i\| \geq \delta}^{N} \left(\|\mathbf{Q}\mathbf{x}_i\|^2 / \|\mathbf{Q}^*\mathbf{x}_i\| + \|\mathbf{Q}^*\mathbf{x}_i\| \right) + \sum_{i=1, \|\mathbf{Q}^*\mathbf{x}_i\| < \delta}^{N} (\|\mathbf{Q}\mathbf{x}_i\|^2 / 2\delta + \delta/2).$$

7.14 Proof of Theorem 13

We note that the minimization of $F(\mathbf{Q})$ over all $\mathbf{Q} \in \mathbb{H}$ such that $\mathbf{Q}\mathbf{P}_{\hat{\mathrm{L}}^\perp} = \mathbf{0}$ in Algorithm 3 can be performed at each iteration with respect to the projected data: $\tilde{\mathbf{P}}_{\hat{\mathrm{L}}}(\mathcal{X}) = \{\tilde{\mathbf{P}}_{\hat{\mathrm{L}}}\mathbf{x}_1, \tilde{\mathbf{P}}_{\hat{\mathrm{L}}}\mathbf{x}_2, \cdots, \tilde{\mathbf{P}}_{\hat{\mathrm{L}}}\mathbf{x}_N\}$.

We note that conditions (6) and (7) hold for $\tilde{\mathbf{P}}_{\hat{\mathrm{L}}}(\mathcal{X})$ with any $\hat{\mathrm{L}} \supseteq \mathrm{L}^*$. Therefore, Theorem 1 implies that $\mathbf{u} \perp \mathrm{L}^*$ and $\hat{\mathrm{L}} \supseteq \mathrm{L}^*$ in each iteration. Since $\dim(\hat{\mathrm{L}})$ decreases by one in each iteration, $\dim(\hat{\mathrm{L}}) = d$ in $D - d$ iterations and thus $\hat{\mathrm{L}} = \mathrm{L}^*$.

8. Conclusion

We proposed an M-estimator for the problems of exact and near subspace recovery. Substantial theory has been developed to quantify the recovery obtained by this estimator as well as its numerical approximation. Numerical experiments demonstrated state-of-the-art speed and accuracy for our corresponding implementation on both synthetic and real data sets.

This work broadens the perspective of two recent ground-breaking theoretical works for subspace recovery by Candès et al. (2011) and Xu et al. (2012). We hope that it will motivate additional approaches to this problem.

There are many interesting open problems that stem from our work. We believe that by modifying or extending the framework described in here, one can even yield better results in various scenarios. For example, we have discussed in §1.2 the modification by Lerman et al. (2012) suggesting tighter convex relaxation of orthogonal projectors when d is known. We also discussed in §1.2 adaptation by Wang and Singer (2013) of the basic ideas in here to the different synchronization problem. Another direction was recently followed up by Coudron and Lerman (2012), where they established exact asymptotic subspace recovery under specific sampling assumptions, which may allow relatively large magnitude of noise. It is interesting to follow this direction and establish exact recovery when using in theory a sequence of IRLS regularization parameters $\{\delta_i\}_{i\in\mathbb{N}}$ approaching zero (in analogy to the work of Daubechies et al. 2010).

An interesting generalization that was not pursued so far is robust data modeling by multiple subspaces or by locally-linear structures. It is also interesting to know whether one can adapt the current framework so that it can detect linear structure in the presence of both sparse elementwise corruption (as in Candès et al. 2011) and the type of outliers addressed in here.

Acknowledgments

This work was supported by NSF grants DMS-09-15064 and DMS-09-56072, GL was also partially supported by the IMA (during 2010-2012). Arthur Szlam has inspired our extended research on robust l_1-type subspace recovery. We thank John Wright for referring us to Xu et al. (2010b) shortly after it appeared online and for some guidance with the real data sets. GL thanks Emmanuel Candès for inviting him to visit Stanford university in May 2010 and for his constructive criticism on the lack of a theoretically guaranteed algorithm for the l_1 subspace recovery of Lerman and Zhang (2010).

Supp. webpage: `http://www.math.umn.edu/~lerman/gms`.

References

A. Agarwal, S. Negahban, and M. J. Wainwright. Noisy matrix decomposition via convex relaxation: optimal rates in high dimensions. *Ann. Statist.*, 40(2):1171–1197, 2012a. ISSN 0090-5364. doi: 10.1214/12-AOS1000.

A. Agarwal, S. Negahban, and M. J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40(5):2452–2482, 2012b.

L. P. Ammann. Robust singular value decompositions: A new approach to projection pursuit. *Journal of the American Statistical Association*, 88(422):pp. 505–514, 1993. ISSN 01621459.

E. Arias-Castro, D. L. Donoho, X. Huo, and C. A. Tovey. Connect the dots: how many random points can a regular curve pass through? *Adv. in Appl. Probab.*, 37(3):571–603, 2005.

E. Arias-Castro, G. Chen, and G. Lerman. Spectral clustering based on local linear approximations. *Electron. J. Statist.*, 5:1537–1587, 2011.

O. Arslan. Convergence behavior of an iterative reweighting algorithm to compute multivariate M-estimates for location and scatter. *Journal of Statistical Planning and Inference*, 118(1-2):115 – 128, 2004. ISSN 0378-3758. doi: 10.1016/S0378-3758(02)00402-0.

A. Bargiela and J. K. Hartley. Orthogonal linear regression algorithm based on augmented matrix formulation. *Comput. Oper. Res.*, 20:829–836, October 1993. ISSN 0305-0548. doi: 10.1016/0305-0548(93)90104-Q.

R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, February 2003.

R. Basri, T. Hassner, and L. Zelnik-Manor. Approximate nearest subspace search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):266–278, 2011. ISSN 0162-8828. doi: http://doi.ieeecomputersociety.org/10.1109/TPAMI.2010.110.

R. Bhatia and D. Drissi. Generalized Lyapunov equations and positive definite functions. *SIAM J. Matrix Anal. Appl.*, 27(1):103–114, May 2005. ISSN 0895-4798. doi: 10.1137/040608970.

P. Bradley and O. Mangasarian. k-plane clustering. *J. Global optim.*, 16(1):23–32, 2000.

S. C. Brubaker. Robust PCA and clustering in noisy mixtures. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '09, pages 1078–1087, Philadelphia, PA, USA, 2009. Society for Industrial and Applied Mathematics.

E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006. doi: 10.1002/cpa.20124.

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

T. F. Chan and P. Mulet. On the convergence of the lagged diffusivity fixed point method in total variation image restoration. *SIAM J. Numer. Anal.*, 36:354–367, 1999. ISSN 0036-1429. doi: 10.1137/S0036142997327075.

V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.*, 21(2):572–596, 2011. ISSN 1052-6234. doi: 10.1137/090761793.

T.-J. Chin, J. Yu, and D. Suter. Accelerated hypothesis generation for multistructure data via preference analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(4):625–638, April 2012. ISSN 0162-8828. doi: 10.1109/TPAMI.2011.169.

A. K. Cline. Rate of convergence of Lawson's algorithm. *Mathematics of Computation*, 26 (117):pp. 167–176, 1972. ISSN 00255718.

M. Coudron and G. Lerman. On the sample complexity of robust PCA. In *NIPS*, pages 3230–3238, 2012.

C. Croux and G. Haesbroeck. Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies. *Biometrika*, 87: 603–618, 2000.

C. Croux, P. Filzmoser, and M. Oliveira. Algorithms for projectionc pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 87(2):218–225, 2007.

A. d'Aspremont, L. El Ghaoui, M. Jordan, and G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3):434–448, 2007. doi: 10.1137/050645506.

I. Daubechies, R. DeVore, M. Fornasier, and C. S. Gunturk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63:1–38, 2010. doi: 10.1002/cpa.20303.

G. David and S. Semmes. Singular integrals and rectifiable sets in $\mathbb{R}^n$: au-delà des graphes Lipschitziens. *Astérisque*, 193:1–145, 1991.

P. L. Davies. Asymptotic behaviour of s-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 15(3):pp. 1269–1292, 1987. ISSN 00905364.

C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM J. on Numerical Analysis*, 7:1–46, 1970.

S. J. Devlin, R. Gnandesikan, and J. R. Kettenring. Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, 76(374):pp. 354–362, 1981. ISSN 01621459.

C. Ding, D. Zhou, X. He, and H. Zha. R1-PCA: rotational invariant L_1-norm principal component analysis for robust subspace factorization. In *ICML '06: Proceedings of the 23rd International Conference on Machine Learning*, pages 281–288, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2. doi: 10.1145/1143844.1143880.

M. Fornasier, H. Rauhut, and R. Ward. Low-rank matrix recovery via iteratively reweighted least squares minimization. *SIAM J. Optim.*, 21(4):1614–1640, 2011. ISSN 1052-6234. doi: 10.1137/100811404.

M. Hardt and A. Moitra. Algorithms and hardness for robust subspace recovery. In *COLT*, pages 354–375, 2013.

J. Ho, M. Yang, J. Lim, K. Lee, and D. Kriegman. Clustering appearances of objects under varying illumination conditions. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 11–18, 2003.

D. Hsu, S.M. Kakade, and Tong Zhang. Robust matrix decomposition with sparse corruptions. *Information Theory, IEEE Transactions on*, 57(11):7221 –7234, nov. 2011. ISSN 0018-9448. doi: 10.1109/TIT.2011.2158250.

P. J. Huber and E. M. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, 2nd edition, 2009. ISBN 978-0-470-12990-6. doi: 10.1002/9780470434697.

Q. Ke and T. Kanade. Robust subspace computation using L_1 norm. Technical report, Carnegie Mellon, 2003.

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. ISSN 00905364.

H. W. Kuhn. A note on Fermat's problem. *Mathematical Programming*, 4:98–107, 1973. ISSN 0025-5610. 10.1007/BF01584648.

N. Kwak. Principal component analysis based on L_1-norm maximization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(9):1672–1680, 2008. doi: 10.1109/TPAMI.2008.114.

C. L. Lawson. *Contributions to the Theory of Linear Least Maximum Approximation*. PhD thesis, University of California, Los Angeles, 1961.

K.-C. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27 (5):684–698, 2005. ISSN 0162-8828. doi: 10.1109/TPAMI.2005.92.

G. Lerman and T. Zhang. l_p-Recovery of the most significant subspace among multiple subspaces with outliers. *ArXiv e-prints*, December 2010. To Appear in *Constructive Approximation*.

G. Lerman and T. Zhang. Robust recovery of multiple subspaces by geometric l_p minimization. *Ann. Statist.*, 39(5):2686–2715, 2011. ISSN 0090-5364. doi: 10.1214/11-AOS914.

G. Lerman, M. McCoy, J. A. Tropp, and T. Zhang. Robust computation of linear models, or how to find a needle in a haystack. *ArXiv e-prints*, February 2012.

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. ISSN 01621459. doi: 10.2307/2288497.

L. Li, W. Huang, I. Gu, and Q. Tian. Statistical modeling of complex backgrounds for foreground object detection. *Image Processing, IEEE Transactions on*, 13(11):1459 – 1472, nov. 2004. ISSN 1057-7149. doi: 10.1109/TIP.2004.836169.

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. In *In Intl. Workshop on Comp. Adv. in Multi-Sensor Adapt. Processing, Aruba, Dutch Antilles*, 2009.

G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *ICML*, 2010.

G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):171 –184, 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2012.88.

H. P. Lopuhaä and P. J. Rousseeuw. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Ann. Statist.*, 19(1):229–248, 1991. ISSN 0090-5364.

R. A. Maronna. Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4(1):pp. 51–67, 1976. ISSN 00905364.

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. ISBN 978-0-470-01092-1; 0-470-01092-4.

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

S. Mendelson. A few notes on statistical learning theory. In *Lecture Notes in Computer Science*, volume 2600, pages 1–40. Springer-Verlag, 2003.

H. Nyquist. Least orthogonal absolute deviations. *Computational Statistics & Data Analysis*, 6(4):361 – 367, 1988. ISSN 0167-9473. doi: 10.1016/0167-9473(88)90076-X.

M. R. Osborne and G. A. Watson. An analysis of the total approximation problem in separable norms, and an algorithm for the total l_1 problem. *SIAM Journal on Scientific and Statistical Computing*, 6(2):410–424, 1985. doi: 10.1137/0906029.

A. M. Ostrowski. *Solution of Equations and Systems of Equations*. Academic Press, Second edition, September 1966. ISBN 0471889873.

M. Soltanolkotabi and E. J. Candès. A geometric analysis of subspace clustering with outliers. *Ann. Stat.*, 40(4):2195–2238, 2012. doi: 10.1214/12-AOS1034.

H. Späth and G. A. Watson. On orthogonal linear approximation. *Numer. Math.*, 51: 531–543, October 1987. ISSN 0029-599X. doi: 10.1007/BF01400354.

C. V. Stewart. Robust parameter estimation in computer vision. *SIAM Reviews*, 41:513–537, 1999.

T. Tao. *Topics in Random Matrix Theory*, volume 132 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012. ISBN 978-0-8218-7430-1.

M. Tipping and C. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2):443–482, 1999.

F. De La Torre and M. J. Black. Robust principal component analysis for computer vision. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 362 –369 vol.1, 2001. doi: 10.1109/ICCV.2001.937541.

F. De La Torre and M. J. Black. A framework for robust subspace learning. *International Journal of Computer Vision*, 54:117–142, 2003. ISSN 0920-5691. doi: 10.1023/A:1023709501986.

P. Tseng. Nearest q-flat to m points. *Journal of Optimization Theory and Applications*, 105:249–252, 2000. ISSN 0022-3239. 10.1023/A:1004678431677.

D. E. Tyler. A distribution-free M-estimator of multivariate scatter. *Ann. Statist.*, 15(1): 234–251, 1987. ISSN 0090-5364. doi: 10.1214/aos/1176350263.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012.

H. Voss and U. Eckhardt. Linear convergence of generalized weiszfeld's method. *Computing*, 25:243–251, 1980. ISSN 0010-485X. doi: 10.1007/BF02242002.

L. Wang and A. Singer. Exact and stable recovery of rotations for robust synchronization. *Information and Inference*, 2013. doi: 10.1093/imaiai/iat005.

G. A. Watson. *Some Problems in Orthogonal Distance and Non-Orthogonal Distance Regression*. Defense Technical Information Center, 2001. URL `http://books.google.com/books?id=WKKWGwAACAAJ`.

G. A. Watson. On the gauss-newton method for l_1 orthogonal distance regression. *IMA Journal of Numerical Analysis*, 22(3):345–357, 2002. doi: 10.1093/imanum/22.3.345.

E. Weiszfeld. Sur le point pour lequel la somme des distances de n points donne's est minimum. *Tohoku Math. J.*, 43:35–386, 1937.

H. Xu, C. Caramanis, and S. Mannor. Principal component analysis with contaminated data: The high dimensional case. In *COLT*, pages 490–502, 2010a.

H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via outlier pursuit. In *NIPS*, pages 2496–2504, 2010b.

H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via outlier pursuit. *Information Theory, IEEE Transactions on*, PP(99):1, 2012. ISSN 0018-9448. doi: 10.1109/TIT.2011.2173156.

L. Xu and A.L. Yuille. Robust principal component analysis by self-organizing rules based on statistical physics approach. *Neural Networks, IEEE Transactions on*, 6(1):131–143, 1995. ISSN 1045-9227. doi: 10.1109/72.363442.

T. Zhang. Robust subspace recovery by geodesically convex optimization. *ArXiv e-prints*, 2012.

T. Zhang, A. Szlam, and G. Lerman. Median K-flats for hybrid linear modeling with many outliers. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on Computer Vision*, pages 234–241, Kyoto, Japan, 2009. doi: 10.1109/ICCVW.2009.5457695.

T. Zhang, A. Szlam, Y. Wang, and G. Lerman. Randomized hybrid linear modeling by local best-fit flats. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1927 –1934, jun. 2010. doi: 10.1109/CVPR.2010.5539866.

T. Zhang, A. Szlam, Y. Wang, and G. Lerman. Hybrid linear modeling via local best-fit flats. *International Journal of Computer Vision*, 100:217–240, 2012. ISSN 0920-5691. doi: 10.1007/s11263-012-0535-6.