\( l_p \)-Recovery of the Most Significant Subspace Among Multiple Subspaces with Outliers

Gilad Lerman · Teng Zhang

Abstract We assume data sampled from a mixture of \( d \)-dimensional linear subspaces with spherically symmetric distributions within each subspace and an additional outlier component with spherically symmetric distribution within the ambient space (for simplicity, we may assume that all distributions are uniform on their corresponding unit spheres). We also assume mixture weights for the different components. We say that one of the underlying subspaces of the model is most significant if its mixture weight is higher than the sum of the mixture weights of all other subspaces. We study the recovery of the most significant subspace by minimizing the \( l_p \)-averaged distances of data points from \( d \)-dimensional subspaces of \( \mathbb{R}^D \), where \( 0 < p \in \mathbb{R} \). Unlike other \( l_p \) minimization problems, this minimization is nonconvex for all \( p > 0 \) and thus requires different methods for its analysis. We show that if \( 0 < p \leq 1 \), then for any fraction of outliers, the most significant subspace can be recovered by \( l_p \) minimization with overwhelming probability (which depends on the generating distribution and its parameters). We show that when adding small noise around the underlying subspaces, the most significant subspace can be nearly recovered by \( l_p \) minimization for any \( 0 < p \leq 1 \) with an error proportional to the noise level. On the other hand, if \( p > 1 \) and there is more than one underlying subspace, then with overwhelming probability the most significant subspace cannot be recovered or nearly recovered. This last result does not require spherically symmetric outliers.
Keywords  Best approximating subspace · $l_p$ minimization · Robust statistics · Optimization on the Grassmannian · Geometric probability · Hybrid linear modeling

Mathematics Subject Classification  Primary 68Q32 · 62G35 · 60D05 · Secondary 62-07 · 68T10

1 Introduction

Principal Component Analysis (PCA) is arguably the most common tool in high dimensional data analysis. It approximates a given data set by a lower-dimensional subspace obtained from solving an $l_2$ optimization problem. While such an $l_2$ minimization can be easily implemented to run fast for moderate-size data, it is not robust to outliers. That is, the estimated subspace can significantly change when adding points sampled from a very different distribution. This obstacle motivated the developments of many algorithms for robust PCA, where some of them are based on $l_1$ minimization. Their robustness is often theoretically guaranteed when restricting both the distribution and the fraction of outliers.

Here, we study the robustness to outliers of a “geometric $l_1$ minimization” for subspace recovery. In fact, we discuss the robustness of the following geometric $l_p$ minimization for all $p > 0$: For a data set $X \subset \mathbb{R}^D$, it tries to minimize among all $d$-dimensional subspaces, $L \subseteq \mathbb{R}^D$, the quantity:

$$e_{l_p}(X, L) = \sum_{x \in X} \text{dist}(x, L)^p,$$

where $\text{dist}(x, L)$ denotes the Euclidean distance between a data point $x$ and the subspace $L$. In this paper, we restrict this minimization to $d$-dimensional linear subspaces (instead of affine), which we refer to as $d$-subspaces.

The geometric $l_1$ minimization is related to some of the recent attempts for robust PCA [23,27,43,44,46]. However, it is hard to implement it directly since it is not convex (the set of $d$-subspaces, over which the $l_1$ energy is minimized, is not convex). Nevertheless, the question of its robustness is fundamentally interesting. While the analysis in [22] implies such robustness when restricting the fraction of outliers, here we ask a more challenging question for the recovery of a single subspace: Can it be recovered by a sufficiently large sample when having no restriction on the fraction of outliers but on their distribution? One possible instance is when the outliers are spherically symmetric, i.e., invariant to rotations (or for simplicity uniformly distributed on the sphere). We make the problem more interesting by assuming points sampled from several multiple subspaces and outliers (where the distributions of all components are spherically symmetric), and we study the recovery of the most significant subspace by geometric $l_1$ (or $l_p$) minimization.
1.1 The Most Significant Subspace and Its Difference from the Global $l_0$ Subspace

Ideally, one may wish to recover the global $l_0$ subspace, that is, the subspace with the largest number of points, by geometric $l_1$ minimization (or $l_p$ geometric minimization with any $p \leq 1$). This is similar in spirit to basis pursuit, where $l_1$ minimization can be used to solve an $l_0$ minimization under some conditions [5,6,11,12].

However, in the setting of this paper it can be impossible to recover the global $l_0$ subspace by the global $l_p$ subspace for any $p > 0$. Indeed, we do not have any restriction on the relative orientation of the underlying subspaces of our model. Furthermore, unlike the $l_0$ energy (that is, number of points on the complement of a given subspace), the $l_p$ energy with $p > 0$ is a continuous function of the subspace $L$. Therefore, the global $l_p$ subspace is different than the global $l_0$ subspace in the case of two sufficiently close subspaces, whose union contain more than half of the points, and a global $l_0$ subspace, which is sufficiently far from these two subspaces. To clarify this point, let us assume for simplicity that $L_1^*$, $L_2^*$ and $L_3^*$ are $d$-subspaces in $\mathbb{R}^D$, where 40% of the points are on $L_1^*$, 30% on $L_2^*$ and 30% on $L_3^*$. Clearly $L_1^*$ is the global $l_0$ subspace. However, if $p > 0$ is fixed and $L_2^*$ and $L_3^*$ are sufficiently close to each other, their $l_p$ energies can be lower than that of $L_1^*$. Therefore, $L_1^*$ is not the global $l_p$ subspace.

As opposed to this example, we will not talk about the exact number of points on a subspace (or around it in a noisy setting), but assume an i.i.d. sample from a mixture measure of $K + 1$ components: $K$ of them along $d$-subspaces $\{L_i^*\}_{i=1}^K$ with weights $\{\alpha_i\}_{i=1}^K$ and another component of outliers with weight $\alpha_0$; more details of the distributions themselves are in Sect. 1.4. We say that $L_1^*$ is the most significant subspace if

$$\alpha_1 > \sum_{i=2}^K \alpha_i. \quad (2)$$

Unlike the condition of the global $l_0$ subspace, which translates here to $\alpha_1 > \max_{i=2}^K \alpha_i$, condition (2) is still valid for $l_p$ subspace recovery if $\{L_i^*\}_{i=2}^K$ are arbitrarily close to each other.

1.2 Background and Related Work

The $l_1$ norm has been widely used to form robust statistics [20,25,31]. The early principle of least absolute deviations for robust regression, minimizes the sum of absolute values of residuals. For example, in linear regression, it minimizes the sum of the absolute values of the deviations of the dependent variable observations from the fitted linear estimator based on the independent variable observations. It is a natural robust alternative for the least squares regression and actually emerged independently of least squares regression (see, e.g., historical review in [10,18,19]).

The sum of absolute values of residuals can also be used in total regression problems, where observational errors of both dependent and independent variables are taken into account. This is a robust alternative for the total least squares problem,
which can be described geometrically as minimizing \((1)\) with \(p = 2\). The robust version with sum of absolute values is equivalent to minimizing \((1)\) with \(p = 1\). Osborne and Watson [29], Späth and Watson [34], and Nyquist [28] suggested a procedure for solving the latter minimization problem over hyperplanes, that is, when the codimension of the subspaces is 1 (see also [3]). Watson [40,41] even suggested an orthogonal \(l_1\) procedure for fitting a surface to data. David and Semmes [7] proposed the minimization of \((1)\) for \(p \geq 1\) for a pure analytic setting, which is free of outliers. In the context of machine learning and data mining, Ding et al. [9] proposed the minimization of \((1)\) with \(p = 1\) as rotation-invariant robust PCA. They also proposed a numerical strategy for approximating a minimizer of \((1)\) when \(p = 1\), but without valid theoretical guarantees for such an approximation. Zhang et al. [47] have formulated an online procedure for this minimization, which can even approximate data by multiple subspaces. More recently, Lerman and Maunu [21] suggested another heuristic for \(0 < p < 2\). Both [47] and [21] lack substantial theoretical guarantees.

In [22], which followed this work, we analyzed the recovery of all underlying subspaces within outliers by minimizing a modified version of \((1)\) (adapted to multiple subspaces). In that work, the outlier distribution is rather general, but the fraction of outliers is restricted. Since we continued developing the current work, it includes improved estimates for some of the constants of [22].

Recently, several convex algorithms for robust PCA (with provable exact recovery) have been suggested [4,23,27,43,44,46]. In [23,27,43,44,46], the problem of fitting a subspace to data is translated into fitting a low-rank matrix to a given matrix, whose columns represent the data points, where outliers correspond to grossly corrupted columns. Both [46] and [23] propose a convex relaxation of the minimization in \((1)\). In the case of a pure inliers-outliers model (inliers lie exactly on a subspace and outliers in its complement) that satisfies certain combinatorial conditions, the subspace outputted by either [46] or [23] is the minimizer of \((1)\) when \(p = 1\). We also view the second (and last) term in the energy functions of Algorithm 1 of [43,44] and Algorithm 2 of [27] as analogs of the energy \((1)\). In the case of spherically symmetric outliers with no restriction of their fraction, it is currently unknown if exact recovery is guaranteed for any of the algorithms in [23,27,43,44,46], though we conjecture it is impossible. On the other hand, we show here that such guarantees exist for the geometric \(l_1\) minimization. To make the problem more challenging (so that the underlying subspace cannot be nearly recovered by PCA due to the spherically symmetric outliers), we find it interesting to ask about the geometric \(l_1\) recovery of the most significant subspace among multiple subspaces within spherically-symmetric outliers.

Hardt and Moitra [17] showed that it is small set expansion hard to exactly recover a \(d\)-subspace in \(\mathbb{R}^D\) with fraction of outliers larger than \((D - d)/D\) for all scenarios satisfying a rather general combinatorial condition. They also developed deterministic and random algorithms for achieving subspace recovery that can handle outliers with fraction at most \((D - d)/D\) for data satisfying their combinatorial condition. Our current work suggests a higher fraction of outliers (arbitrarily close to 100%), however, it does not contradict [17]. First of all, in the case of a single subspace \((K = 1)\), the recovery in our work only applies to spherically symmetric outliers and not to all scenarios satisfying the combinatorial condition of [17]. Second of
all, our work verifies exact recovery in probability, while the result of [17] requires deterministic satisfaction of all scenarios of their combinatorial condition. Third of all, the combinatorial condition of [17] may not be satisfied for our setting when \( K > 1 \), that is, when having multiple subspaces. At last, if the small set expansion problem has no efficient algorithm (which is unknown), then the result of Hardt and Moitra [17] implies the following fact: Any estimator that can exactly recover a subspace in all settings specified by their combinatorial condition with percentage of outliers larger than \((D-d)/D\) cannot be efficient. Since our minimizer involves a non-convex minimization, it is possible that there is no efficient algorithm for computing it.

There are several other nonconvex methods for subspace recovery that seem to work well with high percentages of outliers, in particular, higher than the ones guaranteed for convex methods like [23,27,43,44,46] or are highly common among practitioners (without theoretical guarantees), and we thus review some of them.

In the computer vision literature, a common procedure for subspace fitting uses the Random Sample Consensus (RANSAC) [15] heuristic. In theory, it may not exactly recover subspaces for any positive ratio of outliers. However, in practice it often nearly recovers subspaces when the ambient dimension is sufficiently small. It is possible that the random strategy of Hardt and Moitra [17] may serve as a good theoretically-guaranteed alternative for RANSAC. The RANSAC strategy repeatedly applies the following two steps: 1. randomly select a set of \( d \) independent vectors; 2. count the number of data points within a strip of width \( \epsilon \) around the \( d \)-subspace spanned by those \( d \) vectors (both \( \epsilon \) and the number of iterations of these two steps are parameters set by the user). The final output of this algorithm is the \( d \)-subspace maximizing the quantity computed in step 2.

Torr and Zisserman [37,38] suggested a RANSAC-type strategy, which selects a subspace (among the random set of candidates) by minimizing a supposedly robust variant of the \( l_2 \) distance from a subspace. This variant uses the square function until a fixed threshold and a constant function for larger values. However, by following the proof of Theorem 1.3 in this work (in particular, (103)), one can show that when \( K > 1 \), the subspace obtained by the minimizer of this variant is sufficiently far from the most significant subspace with probability 1.

There are nonconvex methods for removing outliers (or detecting the hidden low-dimensional structures) that can handle arbitrarily large fraction of outliers. For example, Arias-Castro et al. [1] 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 fraction of order \( 1 - O((\epsilon)^{m(D-d)/(d+m(D-d))}) \). Arias-Castro et al. [2] used higher-order spectral clustering affinities to remove outliers and thus detect differentiable surfaces (or certain unions of such surfaces) among uniform outliers, whose maximal fraction can be of a similar asymptotic order as that of the scan statistics.

The work of Soltanolkotabi and Candès [32], which appeared after the online release of the first version of this work, assumed a similar model to the one assumed here but without noise and including another assumption (when \( N \) approaches infinity this assumption becomes \( d < \frac{1}{\sqrt{5}}D \)); they established the exact recovery of all underlying subspaces (and not the most significant subspace) by the sparse subspace
clustering (SSC) algorithm [14]. They also proposed an additional step for removing outliers, which is not convex, and analyzed its performance when \( N \) lies in a certain interval. In fact, they analyzed the part of the SSC algorithm that forms an affinity matrix, where the affinities are obtained via convex optimization. The second part of SSC involves clustering the subspaces by this affinity matrix and is not convex. Soltanolkotabi et al. [33] also analyzed the stability to noise of the first and convex part of a modified version of the SSC algorithm without outliers and when \( d < c_0 D / \log(N) \).

Zhang et al. [48, 49] proposed a method for recovering multiple subspaces by globally incorporating information from several local best-fit subspaces. It can also be adapted for finding only the most significant subspace. Nevertheless, the full guarantees for recovering either the most significant subspace or all underlying subspaces have not been established yet.

1.3 Basic Conventions and Notation

We denote by \( G(D, d) \) the Grassmannian space, i.e., the set of all \( d \)-subspaces of \( \mathbb{R}^D \) with a manifold structure. The geodesic distance between \( F \) and \( G \) in \( G(D, d) \) is

\[
\text{dist}_G(F, G) = \sqrt{\sum_{i=1}^{d} \theta_i^2},
\]

where \( \{\theta_i\}_{i=1}^{d} \) are the principal angles between \( F \) and \( G \) (we review these angles and their relation to geodesics in Sect. 3.2.1). Following Sect. 3.9 of [26], we denote by \( \gamma_{D,d} \) the “uniform distribution on \( G(D, d) \)”. We designate a ball in \( G(D, d) \) by \( B_{G}(L, r) \) as opposed to a Euclidean ball in \( \mathbb{R}^D \), \( B(x, r) \). We refer to any of the global minimizers of (1) among \( L \in G(D, d) \) as a global \( l_p \) subspace. Similarly, local minimizers of (1) among \( L \in G(D, d) \) are local \( l_p \) subspaces.

We use “w.p.” as a shorthand for “with probability”. By saying “with overwhelming probability”, or in short “w.o.p.”, we mean that the underlying probability is at least \( 1 - Ce^{-N/C} \), where \( N \) is the size of the data set \( X \) and \( C \) is a constant independent of \( N \). When using this terminology, we will make sure to estimate the asymptotic dependence of \( C \) on \( D \) and \( d \) and use it to infer the asymptotic dependence of the minimal sample size \( N \) on \( D \) and \( d \); this way we make sure that the probabilistic estimate is not completely useless.

1.4 Setting of This Paper

We assume \( K \) distinct \( d \)-subspaces in \( \mathbb{R}^D \), which we denote by \( \{L_i^*\}_{i=1}^{K} \). Furthermore, we assume an i.i.d. data set \( X \subseteq \mathbb{R}^D \) of size \( N \) sampled from a mixture distribution \( \mu_{\epsilon} \) with components supported on each of the \( d \)-subspaces \( \{L_i^*\}_{i=1}^{K} \) as well as an outlier component and noise level \( \epsilon \geq 0 \). Our typical setting assumes spherically symmetric distributions within \( \{L_i^*\}_{i=1}^{K} \) and (for most of the discussion) spherically symmetric distribution of the outliers (within \( \mathbb{R}^D \)).
For simplicity of our presentation, we replace spherically symmetric distributions with uniform distributions on the sphere, though our analysis can be easily extended to the former distributions. Furthermore, one can always normalize the data to the sphere so that spherically symmetric distributions (or even more general distributions) are mapped to uniform distributions onto the sphere. Normalization of data to the unit sphere is a common practice for robust PCA algorithms [23,24] as well as algorithms for modeling data by multiple subspaces [45,47].

In the noiseless case ($\epsilon = 0$), we denote the $K + 1$ components of the mixture measure by $\{\mu_i\}_{i=0}^K$, where $\mu_0$ is the uniform distribution on $S^{D-1}$ (the $(D-1)$-dimensional unit sphere) that represents outliers and for $1 \leq i \leq K$, $\mu_i$ is the uniform distribution on $S^{D-1} \cap L_i^*$. For the noisy case, we assume that $\{\mu_i\}_{i=1}^K$ are contaminated by the noise distributions $\{\nu_i, \epsilon\}_{i=1}^K$ such that $\text{supp}(\mu_i + \nu_i, \epsilon) \subseteq S^{D-1}$ (that is, all points sampled from this noisy distribution also lie on the unit sphere), and for technical reasons, we assume that the $p$th moments of $\{\nu_i, \epsilon\}_{i=1}^K$ are smaller than $\epsilon^p$ for all $p \leq 1$ (when considering geometric $l_p$ minimization with $p \geq 1$ we only need this condition with $p = 1$ and when considering geometric $l_p$ minimization with $p < 1$ we only need this condition with the relevant value of $p$). If $\epsilon = 0$, then the latter model is consistent with the former one by letting $\{\nu_i, 0\}_{i=1}^K$ be the Dirac $\delta$ distributions at $0$.

For any noise level $\epsilon \geq 0$, the mixture distribution $\mu_\epsilon$ has the form

$$\mu_\epsilon = \alpha_0 \mu_0 + \sum_{i=1}^K \alpha_i (\mu_i + \nu_i, \epsilon), \quad (4)$$

where $\alpha_0 \geq 0$, $\alpha_i > 0 \ \forall \ 1 \leq i \leq K$ and $\sum_{i=0}^K \alpha_i = 1$. If $\epsilon = 0$, then for convenience we replace the notation $\mu_\epsilon$ by $\mu$, i.e.,

$$\mu = \sum_{i=0}^K \alpha_i \mu_i. \quad (5)$$

We refer to $\mu_\epsilon$ created according to this model as spherically uniform HLM (hybrid linear modeling) measure with noise level $\epsilon$ (sometimes we also add “w.r.t. $\{L_i^*\}_{i=1}^K$”). In part of our setting, the assumption on $\mu_0$ can be completely removed, while still assuming that $\{\mu_i\}_{i=1}^K$ are the same. In this case, we refer to $\mu_\epsilon$ as weakly spherically uniform HLM measure with noise level $\epsilon$.

1.5 Mathematical Problems of This Paper

We address here two mathematical problems. The simpler one is implicit in this introduction, though clear from the proofs. It asks whether the most significant subspace $L_1^*$ can be recovered when $\epsilon = 0$ by minimizing $E_\mu(\text{dist}^p(\mathbf{x}, L))$ over all $L \in G(D, d)$. The main problem can be formulated using the empirical distribution $\mu_N$ of i.i.d. sample of size $N$ from $\mu$. It asks whether $L_1^*$ can be recovered (w.o.p.) by minimizing
In the noiseless case and $0 < p \leq 1$, we can exactly recover the most significant subspace by $l_p$ minimization as follows.

**Theorem 1.1** If $\mu$ is a spherically uniform HLM measure on $\mathbb{R}^D$ with $K$ $d$-subspaces $\{L_i^*\}_{i=1}^K \subset G(D, d)$ and mixture coefficients $\{\alpha_i\}_{i=0}^K$ satisfying (2), $X$ is a data set of $N$ points identically and independently sampled from $\mu$ and $0 < p \leq 1$, then the probability that $L_1^*$ is a global $l_p$ subspace is at least

$$1 - C' \exp\left(-\frac{N}{C}\right),$$

where $C$ and $C'$ are constants depending on $D$, $d$, $K$, $p$, $\alpha_0$, $\alpha_1$, and $\min_{2 \leq i \leq K} (\text{dist}_G(L_1^*, L_i^*))$.

The asymptotic dependence of $C$ and $C'$ on $d$ and $D$ (when the rest of the parameters are fixed) can be expressed as follows:

$$C = O\left(d^{\max(13p, 2)} D^3 p\right) \quad \text{and} \quad C' = O\left(d^{d(d+1)/2} + d^{6.5d(D-d)} D^{1.5d(D-d)}\right).$$

The theorem guarantees exact recovery of $L_1^*$ w.o.p. for any percentage of outliers $\alpha_0 < 1$. However, the probability of this event depends (through the constants $C$ and $C'$) on the model parameters. Due to the nonconvexity of the underlying minimization, it is too complicated to estimate the parameters $C$ and $C'$, even for very special cases. However, the theorem states their asymptotic dependence on $d$ and $D$, which is later verified in Sect. 3.4.6. We also show in Sect. 3.4.6 that these estimates imply that $N = \Omega(d^{\max(13p, 2) + 1} D^3 p \max(D - d, d + 1) \log(D))^1$. This indicates some unnecessary oversampling for the single subspace recovery, but we believe that we may improve this estimate. Nevertheless, we currently view this estimate as “a sanity check” ensuring that the minimal $N$ has polynomial dependence on $D$ and $d$, where the polynomial in $D$ is of low order and the polynomial in $d$ is of moderate order at most.

Even though we cannot fully estimate the probability for a global minimum, we can still estimate the probability that $L_1^*$ is a local minimum when $K = 1$. For example, it follows from Theorem 2.2 (which appears later in Sect. 2.2) that if there are $qN$ i.i.d. samples from $\mu_1$ and $(1 - q)N$ i.i.d. samples from $\mu_0$, then $L_1^*$ is a local $l_1$ subspace with probability at least

$$1 - 2d^2 \exp\left(-\frac{q \cdot N}{8.01 \cdot d^4}\right) - 2d \exp\left(-\frac{q^2 \cdot N}{8 \cdot (1 - q) \cdot d^4 \cdot D}\right).$$

We further discuss this estimate in Sect. 2.2.

In the noisy case, exact asymptotic recovery is not possible in general (as we explain in Sect. 3.6.7), but we can extend the above formulation to near recovery.

**Theorem 1.2** If $\epsilon > 0$, $\mu_\epsilon$ is a spherically uniform HLM measure on $\mathbb{R}^D$ of noise level $\epsilon$ with $K$ $d$-subspaces $\{L_i^*\}_{i=1}^K \subset G(D, d)$ and mixture coefficients $\{\alpha_i\}_{i=0}^K$

---

1 We recall that $f = \Omega(g)$ if and only if $g = O(f)$.
satisfying (2), \( \mathcal{X} \) is a data set of \( N \) points sampled identically and independently from \( \mu_\epsilon \) and \( 0 < p \leq 1 \), then the global \( l_p \) subspace for \( \mu_\epsilon \) is in the ball \( B_G (L^*_1, f) \), where

\[
f = f(\epsilon, K, d, p, \alpha_0, \alpha_1, \mu_1) = \frac{\sqrt{d + p} \cdot \pi^{\frac{2p+1}{2p}} \cdot 4.55^{\frac{1}{p}} \cdot \epsilon}{(\alpha_0 + 2 \cdot \alpha_1 - 1) \cdot \pi^{\frac{1}{p}} \cdot 2^{\frac{3}{2}}}, \tag{6}
\]

w.p. at least

\[
1 - \exp(-N \epsilon^{2p} / 2)(C_2 \sqrt{d})^{d/(D-d)} / (2\epsilon^p)^{d/(D-d)}.
\tag{7}
\]

If \( K = 1 \), then the above statement extends to \( 1 < p < \infty \) with

\[
f = f(\epsilon, K, d, p, \alpha_1, \mu_1) = \begin{cases} \sqrt{d + p} \cdot \pi^{2p} \cdot 4.55^{\frac{1}{p}} \cdot \epsilon^{\frac{1}{p}} \cdot \pi^{\frac{1}{p}} & \text{if } 1 < p \leq 2, \\ (\alpha_0 + 2 \cdot \alpha_1 - 1)^{\frac{1}{p}} \cdot \pi^{\frac{1}{p}} \cdot 2^{\frac{3}{2}} & \text{if } p > 2, \end{cases}
\tag{8}
\]

and probability

\[
1 - \exp(-N \epsilon^2 \pi^2 / 2)(C_2 \sqrt{d})^{d/(D-d)} / (2\epsilon^p)^{d/(D-d)}.
\]

We note that if \( f \geq \pi \sqrt{d} \cdot \pi^{\frac{1}{p}} \cdot \pi^{\frac{1}{p}} \cdot 4.55^{\frac{1}{p}} \), then all principle angles are at most \( \pi/2 \), and thus \( B_G (L^*_1, f) = G(D, d) \). The theorem is thus only interesting when \( \epsilon \) is sufficiently small, in particular, when it satisfies the following bound, which ensures that \( f < \frac{\pi \sqrt{d}}{2} \):

\[
\epsilon < \begin{cases} \frac{\sqrt{2d} \cdot (\alpha_0 + 2 \cdot \alpha_1 - 1)^{\frac{1}{p}}}{\sqrt{d + p} \cdot \pi^{\frac{2p+1}{2p}} \cdot 4.55^{\frac{1}{p}}} & \text{if } p \leq 1, \\ \frac{(2d)^{\frac{1}{p}} \cdot (\alpha_0 + 2 \cdot \alpha_1 - 1)}{(d + p) \cdot \pi^{\frac{1}{p}} \cdot 4.55^{\frac{1}{p}}} & \text{if } 1 < p \leq 2 \text{ and } K = 1, \\ \frac{1}{4} & \text{if } p > 2 \text{ and } K = 1. \end{cases}
\tag{9}
\]

Finally, we formulate the impossibility of \( l_p \) recovery when \( p > 1 \) and \( K > 1 \) and thus demonstrate a phase transition at \( p = 1 \) when \( K > 1 \). This result does not require \( \mu_0 \) to be uniform on the sphere (or spherically symmetric).

**Theorem 1.3** Assume that \( \{L^*_i\}_{i=1}^K \) are \( K \) \( d \)-subspaces in \( \mathbb{R}^D \), which are identically and independently distributed according to \( \gamma_{D,d} \). For each \( \epsilon \geq 0 \) and a random sample of \( \{L^*_i\}_{i=1}^K \), let \( \mu_\epsilon \) be a weakly spherically uniform HLM measure on \( \mathbb{R}^D \) (w.r.t. \( \{L^*_i\}_{i=1}^K \)) of noise level \( \epsilon \) and let \( \mathcal{X} \) be a data set of \( N \) points sampled identically and independently from \( \mu_\epsilon \). If \( K > 1 \) and \( p > 1 \), then for almost every \( \{L^*_i\}_{i=1}^K \) (w.r.t. \( \gamma_{D,d}^K \)), there exist positive constants \( \delta_0 \) and \( \kappa_0 \), independent of \( N \), such that for any \( 0 \leq \epsilon < \delta_0 \), the global \( l_p \) subspace of \( \mathcal{X} \) is not in the ball \( B_G (L^*_1, \kappa_0) \) with overwhelming probability.
The overwhelming probability of Theorem 1.3 is not of practical interest, but for completeness we specify it later in (94). More importantly, in Sect. 3.6.6 we provide estimates for $\delta_0$ and $\kappa_0$, which are independent of $\epsilon$. They require some technical definitions, which we would rather avoid here. Instead, we exemplify them for the special case where $K = 2$, $d = 1$, $D = 2$, and $\mu_1$ and $\mu_2$ are uniform distributions on line segments centered on the origin and of length 2. Denoting by $\theta$ the angle between $L^*_{\mu_1}$ and $L^*_{\mu_2}$, the analysis in Sect. 3.6.6 implies the following lower bound for both $\kappa_0$ and $\delta_0$ in this special case:

$$\delta_0, \kappa_0 \geq \begin{cases} 
\frac{1}{8(p+1)^2} \cdot \alpha^2 \cdot \cos^2(\theta) \cdot \sin^2(p-1)(\theta) & \text{if } p \geq 2, \\
2 \cdot \frac{p-4}{p-1} \cdot \frac{1}{p-1} \cdot \alpha^2 \cdot \cos^p(\theta) \cdot \sin^p(\theta) & \text{if } 1 < p < 2.
\end{cases}$$

These lower bounds for $\delta_0$ and $\kappa_0$ approach zero when $\alpha_2$ approaches zero or when $\theta$ approaches 0 or $\pi/2$. We expect such a behavior since if $\alpha_2 = 0$, $\theta = 0$, or $\theta = \pi/2$, then for any $p > 1$, $L^*_{\mu_1}$ is the unique global $l_p$ minimizer w.o.p. We also comment that these bounds are not sharp (in particular, their discontinuity at $p = 2$ is artificial).

1.7 Relevance of Theory

As discussed in Sect. 1.2, the geometric $l_1$ minimization is a prototype for other robust and convex PCA algorithms [23, 27, 43, 44, 46]. Without any control on the fraction of outliers, no guarantees are known for the exact recovery of the other algorithms. We thus find it interesting to analyze the robustness of the geometric $l_1$ minimization to spherically uniform outliers (or spherically symmetric outliers) with no restriction of their fraction and with possibly other underlying subspaces. It is also interesting for us to quantify the phase transition of exact recovery at $p = 1$ (different phase transitions at $p = 1$ and $p = 0$ are discussed in Sects. 4.3 and 1.1, respectively). The analysis of the geometric $l_1$ minimization of this paper has inspired the different analysis of [23, 46] and is also directly used in [22]. Nevertheless, our setting is nonconvex, and we are not aware of theoretically guaranteed strategies to compute the global minimizer. It is possible that the ability to theoretically recover the global minimizer with an arbitrarily large fraction of outliers is closely related to the possible inefficiency of any algorithm that aims to compute this minimizer (see Sect. 4.4).

1.8 Additional Results and Structure of the Paper

Additional theory is reviewed in Sect. 2. In particular, Sect. 2.1 establishes some necessary and sufficient deterministic conditions for a $d$-subspace to be a local $l_p$ minimizer for a given data set; Sect. 2.2 uses these conditions to show that if one samples $N_0$ i.i.d. outliers from $\mu_0$ and $N_1$ i.i.d. inliers from $\mu_1$ and if $N_0 = o(N_1^2)$, then the global $l_0$ subspace (which is also the most significant subspace in this case) is a local $l_1$ subspace. On the other hand, it shows that in a general setting of a single underlying subspace with outliers, the global $l_0$ subspace is a local $l_p$ subspace w.p. 0 when $p > 1$.
and w.p. 1 when $0 < p < 1$; Sect. 2.3 demonstrates natural instances, distinct from the case of spherically uniform outliers (or spherically symmetric outliers), where the most significant subspace is neither a local $l_p$ subspace (even for $p = 1$) nor global one (even for $0 < p < 1$). We separately include all mathematical details verifying the theory of this paper in Sect. 3, while leaving some auxiliary verifications to the Appendix. Finally, Sect. 4 concludes this paper and discusses extensions of its results as well as open problems.

2 Additional Theory

2.1 Combinatorial Conditions for $l_0$ Subspaces Being Local $l_p$ Subspaces

2.1.1 Preliminary Notation

We denote the orthogonal group of $n \times n$ matrices by $O(n)$ and the semigroup of $n \times n$ nonnegative diagonal matrices by $D_+(n)$. We designate the projection from $\mathbb{R}^D$ onto the $d$-subspace $L$ by $P_L$ and the corresponding orthogonal projection by $P_L^\perp$. We represent them by $d \times D$ and $(D - d) \times D$ matrices, respectively. In only a few places in the text, we use $D \times D$ matrix representations instead and thus denote them by $\hat{P}_L$ and $\hat{P}_L^\perp$ instead (where $P_L^T P_L = P_L$ and $P_L^\perp P_L^\perp = \hat{P}_L^\perp$). The nuclear norm of $A$, which is denoted by $\|A\|_*$, is the sum of singular values of $A$. We define the scaled outlying “correlation” matrix $B_{L,\mathcal{X}}$ of a data set $\mathcal{X}$ and a $d$-subspace $L$ as follows:

$$B_{L,\mathcal{X}} = \sum_{x \in \mathcal{X} \setminus L} P_L(x) P_L^\perp(x)^T / \text{dist}(x, L). \quad (11)$$

That is, unlike the covariance matrix, which sums over all data points the rank one matrices $xx^T$, $B_{L,\mathcal{X}}$ sums over all outlying data points (i.e., $x \in \mathcal{X}$ not lying on $L$), the restriction of $xx^T$ to matrices with column space in $L$ and row space in the orthogonal complement of $L$, while scaling this product by the distance of $x$ to $L$, i.e., $\|P_L^\perp(x)\|_2$, where throughout the paper $\| \cdot \|$ denotes the Euclidean norm.

We exemplify $B_{L,\mathcal{X}}$ for a typical counterexample of robust recovery, which we discuss later in Sect. 2.3.

Example 1 Let $D = 2$, $d = 1$, $z = (t_0 \cos(\theta_0), t_0 \sin(\theta_0))^T$, where $t_0 > 0$ and $0 < \theta_0 \leq \frac{\pi}{2}$ and $\mathcal{X} = \{(a_1, 0)^T, (a_2, 0)^T, \ldots, (a_{N_1}, 0)^T, z\}$. That is, $\mathcal{X}$ is a set of $N_1 + 1$ points, where $N_1$ of them lie on the $x$-axis with magnitudes $|a_i|_{i=1}^{N_1}$ and one of them has an angle $\theta_0$ with the $x$-axis and magnitude $t_0$. We denote the $x$-axis by $L_x$ and the line passing through the origin and $z$ by $L_z$. We note that

$$B_{L_x,\mathcal{X}} = \sum_{x \in \mathcal{X} \setminus L_x} P_{L_x}(x) P_{L_x}^\perp(x)^T \text{dist}(x, L_x)^{-1} = P_{L_x}((t_0 \cos(\theta_0), t_0 \sin(\theta_0))^T) \frac{P_{L_x}^\perp((t_0 \cos(\theta_0), t_0 \sin(\theta_0))^T)^T}{\text{dist}((t_0 \cos(\theta_0), t_0 \sin(\theta_0))^T, L_x)} = t_0 \cos(\theta_0) t_0 \sin(\theta_0) / t_0 \sin(\theta_0) = t_0 \cos(\theta_0) \quad (12)$$
and
\[
\mathbf{B}_{L_z, X'} = \sum_{x \in X \setminus L_z} P_{L_x}(x) P_{L_x}^\perp(x)^T \text{dist}(x, L_z)^{-1} \\
= \sum_{i=1}^{N_1} P_{L_x}((a_i, 0)^T) P_{L_x}^\perp((a_i, 0)^T)^T / \text{dist}((a_i, 0)^T, L_z) \\
= \sum_{i=1}^{N_1} a_i \cos(\theta_0) a_i \sin(\theta_0) / |a_i \sin(\theta_0)| = \cos(\theta_0) \sum_{i=1}^{N_1} |a_i|. \tag{13}
\]

### 2.1.2 Conditions for a Local $l_p$ Minimizer

We formulate conditions for an arbitrary $d$-subspace $\hat{L}$ to be a local $l_p$ subspace, while distinguishing between three cases: $p = 1$, $0 < p < 1$, and $p > 1$.

**Theorem 2.1** If $\hat{L} \in G(D, d)$, $X_1 = \{x_i\}_{i=1}^{N_1} \subset \hat{L}$, $X_0 = \{y_i\}_{i=1}^{N_0} \subset \mathbb{R}^D \setminus \hat{L}$, and $X = X_0 \cup X_1$, then a sufficient condition for $L$ to be a local $l_1$ $d$-subspace is that for any $V \in O(d)$ and $C \in D_+(d)$,
\[
\sum_{i=1}^{N_1} \|CV P_{\hat{L}}(x_i)\| > \|CVB_{\hat{L}, X'}\|_*. \tag{14}
\]

Furthermore, a necessary condition is that for any $V \in O(d)$ and $C \in D_+(d)$,
\[
\sum_{i=1}^{N_1} \|CV P_{\hat{L}}(x_i)\| \geq \|CVB_{\hat{L}, X'}\|_*. \tag{15}
\]

**Proposition 2.1** If $\hat{L} \in G(D, d)$, $X_1 = \{x_i\}_{i=1}^{N_1} \subset \hat{L}$, $X_0 = \{y_i\}_{i=1}^{N_0} \subset \mathbb{R}^D \setminus \hat{L}$, $\text{Sp}([x_i]_{i=1}^{N_1}) = \hat{L}$, $X = X_0 \cup X_1$, and $p < 1$, then $\hat{L}$ is a local minimum of $e_{l_p}(X, L)$ among all $L \in G(D, d)$.

**Proposition 2.2** If $\hat{L} \in G(D, d)$, $X_1 = \{x_i\}_{i=1}^{N_1} \subset \hat{L}$, $X_0 = \{y_i\}_{i=1}^{N_0} \subset \mathbb{R}^D \setminus \hat{L}$, $X = X_0 \cup X_1$, and $p > 1$, then a necessary condition for $\hat{L}$ to be a local minimum of $e_{l_p}(X, L)$ among all $L \in G(D, d)$ is
\[
\sum_{i=1}^{N_0} P_{\hat{L}}(y_i) P_{\hat{L}}^\perp(y_i)^T \text{dist}(y_i, \hat{L})^{p-2} = 0. \tag{16}
\]

This statement is also true when $X_1 = \emptyset$ and $0 < p \leq 1$.

The above conditions follow from differentiating the corresponding energy function (along geodesics) and using the resulting derivative to form necessary and sufficient
conditions for local minimum (see their proof in Sect. 3.2). However, intuitively it is hard to explain their expressions without going through all calculations. Instead, we exemplify them as follows.

**Example 2** We simplify the conditions of Theorem 2.1 and Propositions 2.1 and 2.2 for the special case of Example 1.

**The Case** $p = 1$:

Let us first simplify (14) (or equivalently (15)) in this example. If $\dot{L} = L_x$, then the set of inliers and outliers are $\mathcal{X}_1 = \{(a_i, 0)^T\}_{i=1}^{N_1}$ and $\mathcal{X}_0 = \{z\}$, respectively. Since $d = 1$, then $V \in O(d)$ is either 1 or $-1$ and $C$ is a positive constant $c$. The LHS of (14) thus has the form

$$\sum_{x \in \mathcal{X}_1} \|CVP_L(x)\| = c \sum_{i=1}^{N_1} |a_i|$$

and computing $B_{L, \mathcal{X}}$ as in (12), the RHS has the form

$$\|CVB_{L, \mathcal{X}}\|_* = c t_0 \cos(\theta_0).$$

Therefore, a sufficient condition for $L_x$ to be a local $l_1$ line is

$$\sum_{i=1}^{N_1} |a_i| > t_0 \cos(\theta_0).$$

If $\dot{L} = L_z$, then $\mathcal{X}_1 = \{z\}$ and $\mathcal{X}_0 = \{(a_i, 0)^T\}_{i=1}^{N_1}$. Applying (13) and following similar calculations as above, we have that a sufficient condition for $L_z$ to be a local $l_1$ line is

$$\cos(\theta_0) \sum_{i=1}^{N_1} |a_i| < t_0.$$

If, on the other hand, $\dot{L}$ does not pass through any point in $\mathcal{X}$, then $\mathcal{X}_1 = \emptyset$ and $\mathcal{X}_0 = \mathcal{X}$. Therefore, the LHS of (14) is 0, and thus (14) never holds.

All the above conditions are also necessary when their inequalities are not strict (see (15)).

We thus note that if $\theta_0 = \pi/2$, then both $L_x$ and $L_z$ are the only two local $l_1$ lines (assuming the obvious conditions: $t_0 > 0$ and $\sum_{i=1}^{N_1} |a_i| > 0$). If, on the other hand, $0 < \theta_0 < \pi/2$, then $L_x$ is a local $l_1$ line if $\sum_{i=1}^{N_1} |a_i|/t_0 > \cos(\theta_0)$, and $L_z$ is a local $l_1$ line if $\sum_{i=1}^{N_1} |a_i|/t_0 < 1/\cos(\theta_0)$ (we also recall that for necessary conditions, we relax the strict inequalities). Therefore, for fixed $0 < \theta_0 < \pi/2$, at least one of $L_x$ or $L_z$ is a local $l_1$ line, and there are no other local minimizers. If $t_0$ is sufficiently large, then $L_z$ is the global $l_1$ line, and if $t_0$ is sufficiently small, then $L_x$ is the global $l_1$ line.
The Case $0 < p < 1$:

We note that Proposition 2.1 implies that both $L_x$ and $L_z$ are local $l_p$ lines (as long as $N_1 \neq 0$ and one of the $a_i$’s is not zero).

The Case $p > 1$:

We express the necessary condition of Proposition 2.2 in our setting. If $\hat{L} = L_x$, then the LHS of (16) is $t_0^p \cos(\theta_0) (\sin(\theta_0))^{p-1}$. Therefore, (16) holds in this case only when $\theta_0 = \pi/2$ (recall that $0 < \theta_0 \leq \pi/2$). Similarly, if $\hat{L} = L_z$, then the LHS of (16) is $\sum |a_i|^p \cos(\theta_0) \sin(\theta_0)^{p-1}$, and thus also in this case (16) holds only when $\theta_0 = \pi/2$.

At last, if $\hat{L}$ has an angle $\theta$ with the $x$-axis, where $-\pi/2 < \theta \neq 0, \theta_0 < \pi/2$, that is, $\hat{L}$ is any line but not $L_x$ or $L_z$, then (16) holds only when

$$\cos(\theta) \sin(\theta) |\sin(\theta)|^{p-2} \sum_{i=1}^{N_1} |a_i| + t_0 \cos(\theta - \theta_0) \sin(\theta - \theta_0) |\sin(\theta - \theta_0)|^{p-2} = 0.$$ (17)

We first note that if $\theta_0 = \pi/2$, then the LHS of (17) is either positive or negative, and thus $\hat{L}$ is not a local minimum. If, on the other hand, $\theta_0 \neq \pi/2$, then since both $L_z$ and $L_x$ are not local minimizers (see above), there then exists $\theta$ such that $\hat{L} \equiv \hat{L}(\theta)$ is a local minimizer (a continuous function over the Grassmannian has at least one local minimizer). If $0 < \theta < \pi/2$ or $-\pi/2 < \theta < 0$, then the LHS of (17) is either positive or negative. It is thus necessary that $0 < \theta < \theta_0$. That is, a local minimizer $\hat{L}$ must lie between $L_x$ and $L_z$. Furthermore, $\hat{L} \in G(D, d)$ is a local $l_p$ minimum w.p. 0 (w.r.t. $\gamma_{D, d}$), since $0 < \theta < \theta_0$ satisfies (17) w.p. 0.

We emphasize that for $p > 1$, we only specified a necessary condition. In particular, when $\theta_0 = \pi/2$, we suspect that almost always only one of the subspaces $L_z$ and $L_x$ can be a local subspace. Indeed, when $p = 2$ (and $\theta_0 = \pi/2$), it follows from basic eigenvalue analysis of the covariance matrix that the following holds: If $t_0$ is sufficiently small, then $L_x$ is the only global (or local) $l_2$ subspace; if $t_0$ is sufficiently large, then $L_z$ is the only global (or local) $l_2$ subspace; and for a unique choice of $t_0$ (given the other parameters), both $L_x$ and $L_z$ are the global minimizers.

2.2 Local $l_p$ Subspaces for Probabilistic Settings with a Single Subspace

We exemplify how to use the conditions of Sect. 2.1.2 in a probabilistic setting of i.i.d. samples from a uniform HLM measure with a single underlying subspace (i.e., $K = 1$). More precisely, we assume that $\mu_0$ and $\mu_1$ are uniform on $S^{D-1}$ and $S^{D-1} \cap L_1^*$, respectively, where $L_1^* \in G(D, d)$ is fixed, and sample i.i.d. inliers from $\mu_1$ and i.i.d. outliers from $\mu_0$ (instead of using mixture weights).

Since $K = 1$, $L_1^*$ is both the most significant subspace and the global $l_0$ subspace w.o.p. For any $p > 0$, we determine whether $L_1^*$ is also a local $l_p$ subspace w.o.p. Our proofs appear in Sect. 3.3.

We first claim that for $p = 1$, the global $l_0$ subspace is a local $l_p$ subspace w.o.p. as long as the fraction of inliers is larger than 0 (assuming that $N$ is sufficiently large).
Theorem 2.2 If \( L_1^* \in G(D, d) \), \( \mathcal{X} \) is a data set in \( \mathbb{R}^D \) of \( N_0 + N_1 \) points, where \( N_0 \) of them are uniformly sampled from \( S^{D-1} \) and \( N_1 \) of them are uniformly sampled from \( S^{D-1} \cap L_1^* \), then \( L_1^* \) is a local \( l_1 \) subspace of \( \mathcal{X} \) w.p. at least

\[
1 - 2d^2 \exp \left( -\frac{N_1}{8.01 \cdot d^4} \right) - 2d D \exp \left( -\frac{N_1^2}{8 \cdot d^4 \cdot D \cdot N_0} \right).
\]

(18)

We note that if \( N_0 = o(N_1^2) \), then \( L_1^* \) is a local \( l_1 \) subspace of \( \mathcal{X} \) w.o.p. However, when \( N_1 \ll \sqrt{N_0} \), then the lower bound for the probability in (18) is actually negative and thus meaningless.

We observe that the asymptotic requirement \( N_0 = o(N_1^2) \) allows any fraction of outliers lower than 1 when \( N \to \infty \). Indeed, if \( 0 \leq \alpha < 1 \) is fixed, \( N_0 = \alpha N \), and \( N_1 = (1 - \alpha) N \), then \( N_0 = o(N_1^2) \) is equivalent to \( \alpha = o(N \cdot (1 - \alpha)^2) \), which is satisfied when \( N \to \infty \).

We emphasize, however, that this recovery of local minima with arbitrarily high percentage of outliers requires a significantly large number of inliers. Indeed, the first exponent in (18) implies that \( N_1 = \Omega(d^4) \). Moreover, the second exponent in (18) implies that \( N_1 = \Omega(d^2 \sqrt{D} \sqrt{N_0}) \).

For comparison, the S-REAPER algorithm [23] can recover the global \( l_1 \) minimizer when \( d < (D - 1)/2 \) with \( N_1 = \Omega(d) \) and \( N_0 = \Omega(D) \), which are significantly smaller (see [23, Theorem 1.1]). However, in this case the asymptotic fraction of outliers (when \( N \) approaches infinity) is restricted as follows: \( N_0 / N < D / (D + 30d) \) (it is possible that 30 can be reduced to a number closer to 1). We remark that in this case with no noise, the minimizer of S-REAPER is an orthogonal projector and not a relaxation of it, and thus it reveals the global \( l_1 \) minimizer. Furthermore, while [23, Theorem 1.1] assumes normal distributions for the inliers and outliers, the S-REAPER normalizes the data points to the sphere, and thus it also applies to our case of spherically uniform distributions.

Next we discuss the case where \( p \neq 1 \). If \( p > 1 \), then Proposition 2.2 implies that under a rather general setting, the global \( l_0 \) subspace is not a local \( l_p \) subspace w.p. 1. Indeed, it is rather unlikely to satisfy (16). We clarify this idea by showing in Sect. A.5 that if \( p > 1 \), the inliers are sampled from the single subspace \( L_1^* \), the outlier distribution does not concentrate on any subspace, and \( D > d - 1 \), then w.p. 1 \( L_1^* \) is not a local \( l_p \) minimizer.

If, on the other hand, \( 0 < p < 1 \), then Proposition 2.1 implies that w.o.p. \( L_1^* \) is a local \( l_p \) subspace. In fact, this proposition suggests the weakest condition one would expect for a subspace to be a local minimizer, that is, being spanned by the points it contains.

The phase transition phenomenon demonstrated above at \( p = 1 \) for the global \( l_0 \) subspace (or most significant subspace) to be a local \( l_p \) is rather artificial in the current setting with \( K = 1 \). Indeed, when \( p > 1 \), the distance between the global \( l_0 \) subspace and the global \( l_p \) subspace (which is also a local \( l_p \) subspace) approaches 0 as \( N \) approaches infinity. Moreover, Theorem 1.2 shows that this formal phase transition also breaks down with noise. Nevertheless, Theorems 1.1 and 1.3 indicate that there is a clear phase transition for a spherically uniform HLM model with \( K > 1 \).
2.3 Counterexamples for Robustness of Best $l_p$ Subspaces

We discuss here basic situations, where global $l_p$ $d$-subspaces are not robust to outliers for all $0 < p < \infty$. More precisely, we show how a single outlier can completely change the underlying subspace. These cases differ from our underlying model of spherically uniform outliers (or spherically symmetric outliers). In all examples below, we assume a single underlying subspace and thus discuss the global $l_0$ subspace instead of the most significant subspace. While we describe a probabilistic setting to sample the data, we only care about a single counterexample sampled this way. We thus do not bother about statements in high probability (even though they are correct), but a positive statement for at least one of the sampled data sets.

A typical example includes $N^1$ points sampled identically and independently from a uniform distribution on $B(0, \epsilon) \cap L^* \subseteq \mathbb{R}^D$, where $L^*$ is a $d$-subspace of $\mathbb{R}^D$, and an additional outlier located on a unit vector orthogonal to $L^*$. By choosing $\epsilon$ sufficiently small, e.g., $\epsilon \leq N_1^{-1/p}$, the global $l_p$ subspace passes through the single outlier and is orthogonal to the initial $d$-subspace for all $p > 0$, which is the global $l_0$ $d$-subspace.

If $p = 1$, then the global $l_0$ $d$-subspace in this example is still a local $l_1$ subspace (as explained in Example 2 for the special cases $d = 1$ and $D = 2$). Nevertheless, if the outlier is located instead on a unit vector having an elevation angle with the original $d$-subspace less than $\pi/2$, then $\epsilon$ can be chosen so that the global $l_0$ subspace is not even a local $l_1$ subspace (see again Example 2). However, if $0 < p < 1$, then Proposition 2.1 implies that the global $l_0$ subspace is still a local $l_p$ subspace in both examples.

Similarly, it is not hard to produce examples of data points on the unit sphere of $\mathbb{R}^D$ where the global $l_0$ subspace is still not a global $l_p$ subspace for all $p > 0$. It is important for us to point this out, since for simplicity we formulated the theory for data lying on the unit sphere, and by normalizing the data sets in the examples above to the unit sphere, they may not form counterexamples anymore. For simplicity, we give a counterexample when $D = 3$ and $d = 2$. We uniformly sample $N^1$ inliers ($N_1^1 > 2$) from an arc on the great circle in the $xy$-plane with the following parametrization: $(\cos \theta, \sin \theta, 0)$, where $\theta \in [-\epsilon, \epsilon]$. We also fix an outlier $(x_0, y_0, z_0) \in S^2$ such that $z_0 \neq 0$. For any fixed $p > 0$ and $\epsilon$ sufficiently small, the 2-subspace spanned by $(x_0, y_0, z_0)$ and $(1, 0, 0)$ (which is the center of the arc) results in a smaller $l_p$ energy than that of the global $l_0$ subspace (i.e., the $xy$-plane). That is, the global $l_0$ subspace is not the global $l_p$ subspace.

3 Verification of Theory

We describe here the proofs of the theorems and propositions of this paper according to the following order of sections: Sects. 2.1, 2.2 and 1.

3.1 Preliminaries

3.1.1 Basic Notation and Conventions

We denote the Frobenius dot product and norm by $\langle A, B \rangle_F$ and $\|A\|_F$, that is, $\langle A, B \rangle_F = \text{tr}(A^T B)$ and $\|A\|_F = \sqrt{\langle A, A \rangle_F}$. The $n \times n$ identity matrix is writ-
ten as $I_n$. We denote the subset of $D_+(n)$ with Frobenius norm 1 by $ND_+(n)$. If $m > n$, we let $O(m, n) = \{X \in \mathbb{R}^{m \times n} : X^TX = I_n\}$, whereas if $n > m$, $O(m, n) = \{X \in \mathbb{R}^{m \times n} : XX^T = I_m\}$.

We sometimes apply the energy (1) to a single point $x$, while using the notation: $e_{l_p}(x, L) \equiv e_{l_p}(\{x\}, L)$.

### 3.1.2 Auxiliary Lemmata

We formulate several technical lemmata, which will be proved in Appendices A.3–A.6.

**Lemma 3.1** If $L_1, \hat{L}_1 \in G(D, d)$, $p > 0$, and $\mu_1$ is a uniform measure on $L_1 \cap S^{D-1}$, then

$$E_{\mu_1}(e_{l_p}(x, \hat{L}_1)) > \begin{cases} \pi^{-p/2} \cdot 2^p \cdot d^{-p} \cdot \text{dist}_G(L_1, \hat{L}_1)^p & \text{if } p \geq 2, \\ 0.88 \cdot 2^{3p/2} \cdot \pi^{-(2p+1)/2} \cdot (d + p)^{-p/2} \cdot \text{dist}_G(L_1, \hat{L}_1)^p & \text{if } p < 2. \end{cases}$$

**Lemma 3.2** For any $x \in \mathbb{R}^D$ and $L_1, L_2 \in G(D, d)$,

$$|\text{dist}(x, L_1) - \text{dist}(x, L_2)| \leq \|x\| \text{dist}_G(L_1, L_2).$$

**Lemma 3.3** If $L_1, L_2 \in G(D, d)$, $\mu_1$ and $\mu_2$ are uniform measures on $L_1 \cap S^{D-1}$ and $L_2 \cap S^{D-1}$, respectively, and $p \leq 1$, then for any $\hat{L} \in G(D, d)$,

$$\mathbb{E}_{\mu_1}(\text{dist}(x_1, \hat{L})^p) + \mathbb{E}_{\mu_2}(\text{dist}(x_2, \hat{L})^p) \geq \mathbb{E}_{\mu_1}(\text{dist}(x_1, L_1)^p) + \mathbb{E}_{\mu_2}(\text{dist}(x_2, L_1)^p) \quad \text{for } i = 1, 2. \quad (19)$$

### 3.2 Proofs for the Theory of Sect. 2.1: Combinatorial Conditions via Calculus on the Grassmannian

#### 3.2.1 Preliminaries: Principal Angles, Principal Vectors, Representation of the Grassmannian, and Geodesics on the Grassmannian

We frequently use here principal angles and for completeness we present one of their equivalent definitions (Sect. 12.4.3 of [16] provides additional background on principal angles). For two $d$-subspaces $F$ and $G$ with corresponding orthonormal bases stored as columns of the matrices $Q_F, Q_G \in \mathbb{R}^{D \times d}$, respectively, the principal angles $\pi/2 \geq \theta_1 \geq \theta_2 \geq \cdots \geq \theta_d \geq 0$ are obtained by

$$\theta_i = \arccos(\sigma_{d-i}(Q_G^T Q_F)), \quad i = 1, \ldots, d, \quad (20)$$

where $\sigma_{d-i}(Q_G^T Q_F)$ is the $(d - i)$th singular value of the matrix $Q_G^T Q_F$. We remark that we order the principal angles decreasingly, unlike the common agreement [16] (Sect. 12.4.3), where $\sigma_{d-i}$ in (20) is replaced by $\sigma_i$.

We denote by $k = k(F, G)$ the largest number such that $\theta_k \neq 0$, so that $\theta_1 \geq \cdots \geq \theta_k > \theta_{k+1} = \cdots = \theta_d = 0$. We refer to this number as the interaction dimension.
and reserve the index $k$ for denoting it (the subspaces $F$ and $G$ will be clear from
the context). We recall that the principal vectors $\{v_i\}_{i=1}^d$ and $\{v'_i\}_{i=1}^d$ of $F$ and $G$, respectively, are two orthogonal bases for $F$ and $G$ satisfying
\[
\langle v_i, v'_i \rangle = \cos(\theta_i) \quad \text{for } i = 1, \ldots, d,
\]
and
\[
v_i \perp v'_j \quad \text{for all } 1 \leq i \neq j \leq k.
\]
We define the complementary orthogonal system $\{u_i\}_{i=1}^d$ for $G$ with respect to $F$ by the formula:
\[
\begin{cases}
v'_i = \cos(\theta_i)v_i + \sin(\theta_i)u_i, & i = 1, 2, \ldots, k, \\
u_i = v_i, & i = k + 1, \ldots, d.
\end{cases}
\]
Clearly,
\[
u_i \perp v_j \quad \text{for all } 1 \leq i, j \leq k.
\]
We note that $F + G$ can be decomposed using these principal vectors as follows:
\[
F + G = \text{Sp}(v_1, u_1) \bigoplus \text{Sp}(v_2, u_2) \bigoplus \cdots \bigoplus \text{Sp}(v_k, u_k) \bigoplus (F \cap G),
\]
where $\bigoplus$ denotes an orthogonal sum (i.e., any two subspaces of the sum are orthogonal). Therefore, the interaction between $F$ and $G$ can be described only within these 2-dimensional subspaces $\text{Sp}(v_i, u_i)$ (equivalently, $\text{Sp}(v_i, v'_i)$) via the principal angles. This idea is also motivated by purely geometric intuition in Sect. 2 of [42].

It follows from [42, Theorem 9] that if the largest principal angle between $F$ and $G$ is less than $\pi/2$, then there is a unique geodesic line between them. Following [13, Theorem 2.3], we can parametrize this line from $F$ to $G$ by the following function $L: [0,1] \rightarrow G(D, d)$, which is expressed in terms of the principal angles $\{\theta_i\}_{i=1}^d$ of $F$ and $G$, the principal vectors $\{v_i\}_{i=1}^d$ of $F$, and the complementary orthogonal system $\{u_i\}_{i=1}^d$ of $G$ with respect to $F$:
\[
L(t) = \text{Sp}((\cos(t\theta_i)v_i + \sin(t\theta_i)u_i)_{i=1}^d).
\]
The length of this geodesic line is clearly expressed by the distance $\text{dist}G$ of (3). We remark that (22) only holds when equipping the Grassmannian with this distance.

3.2.2 Proof of Theorem 2.1

In order to establish quantitative conditions for $\dot{L}$ to be a local minimizer of $e_{l_1}(\mathcal{X}, L)$ among all $d$-subspaces in $G(D, d)$, we arbitrarily fix a $d$-subspace $\hat{L} \in B_G(\hat{L}, 1)$ and

\textcopyright Springer
check the sign of the derivative of the $l_1$ energy when restricted to the geodesic line from $\hat{L}$ to $\check{L}$. If this derivative is positive, then $\check{L}$ is a local $l_1$ subspace. Similarly, if $\hat{L}$ is a local $l_1$ subspace, then this derivative is nonnegative.

The restriction of $\hat{L}$ to $B_G(\hat{L}, 1)$ implies that $\theta_1 \leq 1$, and thus by [42, Theorem 9], this geodesic line (connecting $\hat{L}$ and $\check{L}$) is unique. We parametrize it by the function $L: [0,1] \to G(D,d)$ of (22), where here $\{\theta_i\}_{i=1}^d$ are the principal angles between $\hat{L}$ and $\check{L}$, $\{v_j\}_{j=1}^d$ are the principal vectors of $\hat{L}$, and $\{u_j\}_{j=1}^d$ are the complementary orthogonal system for $\check{L}$ with respect to $\hat{L}$. The necessary and sufficient conditions for $\hat{L}$ to be a local $l_1$ subspace will be formulated in terms of the sign of the derivative of $e_{l_1}(X,L(t))$: $[0,1] \to \mathbb{R}$ at $t = 0$. Clearly, this derivative only exists from the right, however, our notation throughout the paper does not emphasize it (also when $l_1$ is replaced with $l_p$).

We follow by simplifying the expression for the function $e_{l_1}(X,L(t))$ and its derivative according to $t$. We denote the projection from $\mathbb{R}^D$ onto $\text{Sp}(v_j, u_j)$, where $1 \leq j \leq d$, by $P_j$ and the projection from $\mathbb{R}^D$ onto ($\hat{L} + \check{L}$) by $P^\perp$ and use this notation to express the following components of the function $e_{l_1}(X_0, L(t))$ for $i = 1, \ldots, N_0$ (we later express the components of $e_{l_1}(X_1, L(t))$):

$$
\text{dist}(y_i, L(t)) = \sqrt{\sum_{j=1}^d \text{dist}^2(P_j(y_i), L(t)) + \text{dist}^2(P^\perp(y_i), L(t))}
$$

$$
= \sqrt{\sum_{j=1}^d ((-\sin(t\theta_j)v_j + \cos(t\theta_j)u_j) \cdot y_i)^2 + \text{dist}^2(P^\perp(y_i), L(t))}. \quad (23)
$$

We differentiate\textsuperscript{2} the expression for $\text{dist}(y_i, L(t))$ in (23) for all $1 \leq i \leq N_0$ as follows (note that we use the fact that $\text{dist}^2(P^\perp(y_i), L(t))$ is independent of $t$):

$$
\frac{d}{dt} (\text{dist}(y_i, L(t))) = -\sum_{j=1}^d \theta_j \left( (\cos(t\theta_j)v_j + \sin(t\theta_j)u_j) \cdot y_i \right) \left( (-\sin(t\theta_j)v_j + \cos(t\theta_j)u_j) \cdot y_i \right) / \text{dist}(y_i, L(t)). \quad (24)
$$

At $t = 0$, it becomes

$$
\left. \frac{d}{dt} (\text{dist}(y_i, L(t))) \right|_{t=0} = -\sum_{j=1}^d \theta_j (v_j \cdot y_i)(u_j \cdot y_i) / \text{dist}(y_i, L(0)). \quad (25)
$$

We form the following matrices: $C = \text{diag}(\theta_1, \theta_2, \ldots, \theta_d)$, $\check{V} \in O(d, D)$ with $j$th row $v_j^T$, and $\check{U} \in O(d, D)$ with $j$th row $u_j^T$. We then reformulate (25) using these

\textsuperscript{2} Since $0 \leq t \leq 1$, the derivative at $t = 0$ is from the right (and the derivative at $t = 1$ is from the left).
matrices as follows:

\[
\frac{d}{dt} \left( \text{dist}(y_i, L(t)) \right) \bigg|_{t=0} = -\frac{\text{tr}(C\tilde{V}y_i y_i^T \tilde{U}^T)}{\text{dist}(y_i, L)}.
\] (26)

Similarly, we express the components of \(e_{l_1}(\mathcal{X}_i, L(t))\) for all \(x_i \in \hat{L}\), where \(i = 1, 2, \ldots, N_1\), by

\[
\text{dist}(x_i, L(t)) = \left( \sum_{j=1}^{d} |(v_j \cdot x_i)|^2 \sin^2(t\theta_j) \right)^{1/2}
\]

and differentiate these expressions as follows:

\[
\frac{d}{dt} \left( \text{dist}(x_i, L(t)) \right) = \sum_{j=1}^{d} \theta_j |v_j \cdot x_i|^2 \sin(t\theta_j) \cos(t\theta_j) \frac{1}{\text{dist}(x_i, L(t))}.
\] (27)

At \(t = 0\), these derivatives become

\[
\frac{d}{dt} \left( \text{dist}(x_i, L(t)) \right) \bigg|_{t=0} = \sum_{j=1}^{d} |(v_j \cdot x_i)|^2 \theta_j^2 = \|C\tilde{V}x_i\|.
\] (28)

Combining (26) and (28) and using

\[
A := \sum_{i=1}^{N_0} y_i y_i^T / \text{dist}(y_i, \hat{L}),
\]

we obtain the following expression for the derivative of the \(l_1\) energy of (1):

\[
\frac{d}{dt} \left( e_{l_1}(\mathcal{X}, L(t)) \right) \bigg|_{t=0} = \sum_{i=1}^{N_1} \|C\tilde{V}x_i\| - \text{tr}(C\tilde{V}A\tilde{U}^T).
\] (29)

Replacing \(\tilde{V}\) with \(V \in O(d)\), whose \(j\)th row is \(P_L(v_j)^T\), and \(\tilde{U}\) with \(U \in \mathbb{R}^{d \times (D-d)}\), where \(U^T = [U_1, U_2]\), \(U_1 \in O(D - d, k)\), whose \(j\)th row is \(P_L^T(u_j)^T\), and \(U_2 = 0_{(D-d) \times (d-k)}\), we may rewrite this expression as follows:

\[
\frac{d}{dt} \left( e_{l_1}(\mathcal{X}, L(t)) \right) \bigg|_{t=0} = \sum_{i=1}^{N_1} \|CV P_L x_i\| - \text{tr}(CVB_{L,\mathcal{X}} U^T).
\] (30)

We note that

\[
\max_{U^T}(\text{tr}(CVB_{L,\mathcal{X}} U^T)) = \|CVB_{L,\mathcal{X}}\|_*.
\] (31)
Indeed, denoting the thin SVD decomposition of $\text{CVB}_{\dot{L},X}$ by $U_0 \Sigma_0 V_0^T$, we have that
\[
\text{tr}(\text{CVB}_{\dot{L},X} U^T) = \text{tr}(U_0 \Sigma_0 V_0^T U^T) = \text{tr}(\Sigma_0 V_0^T U^T U_0) \leq \text{tr}(\Sigma_0) = \|\text{CVB}_{\dot{L},X}\|_*.
\] (32)
and equality is achieved in (32) when $U^T = V_0 U_0^T$. The theorem is thus concluded by combining (30) and (31).

The theorem is now easily concluded. Indeed, if (14) is satisfied, then it follows from (30) and (32) that the derivative of $e_{l_1}(X, L(t))$ at $t = 0$ is positive, and thus $\dot{L}$ is a local $l_1$ subspace. If, on the other hand, $\dot{L}$ is a local $l_1$ subspace, then the derivative of $e_{l_1}(X, L(t))$ at $t = 0$ is nonnegative for any geodesic line. It thus follows from (30) and (31) that (15) is satisfied.

3.2.3 Simultaneous Proof for Both Propositions 2.1 and 2.2

For the $d$-subspace $\dot{L}$ and an arbitrary $d$-subspace $\hat{L} \in \mathbb{B}_G(\dot{L}, 1)$, we form the geodesic line parametrization $L(t)$ and the corresponding matrices $C, \tilde{V}, \tilde{U}, V, U$ as in the proof of Theorem 2.1.

We assume first that $p > 1$ (and thus start with proving the main part of Proposition 2.2). We note for $z \in \mathbb{R}^D$,
\[
\frac{d}{dt} \text{dist}(z, L(t))^p = p \text{dist}(z, L(t))^{p-1} \frac{d}{dt} \text{dist}(z, L(t)),
\] (33)
where if $z = x_i$, $i = 1, 2, \ldots, N_1$, or $z = y_i$, $i = 1, 2, \ldots, N_0$, then the derivative in the RHS of (33) can be formulated using (24) or (27), respectively. Applying (25), (28), (33) and the fact that $\text{dist}(x_i, \dot{L}) = 0$, for $i = 1, 2, \ldots, N_1$, we obtain that
\[
\frac{d}{dt} \left( e_{l_p}(X, L(t)) \right) \bigg|_{t=0} = -p \sum_{i=1}^{N_0} \text{dist}(y_i, \dot{L})^{p-2} \text{tr}(C\tilde{V}y_i y_i^T \tilde{U}^T)
= -p \sum_{i=1}^{N_0} \text{dist}(y_i, \dot{L})^{p-2} \text{tr}(C V P_{\dot{L}}(y_i) P_{\perp L}(y_i)^T U^T).\] (34)

If $\dot{L}$ is a local minimum of $e_{l_p}(X, L)$, then the LHS of (34) is nonnegative. Fixing $C = V = I_d$ in the RHS of (34) and using its nonnegativity and then applying (31), we conclude that
\[
0 \geq \max_U p \sum_{i=1}^{N_0} \text{dist}(y_i, \dot{L})^{p-2} \text{tr}(P_{\dot{L}}(y_i) P_{\perp L}(y_i)^T U^T) = p \left\| \sum_{i=1}^{N_0} \text{dist}(y_i, \dot{L})^{p-2} P_{\dot{L}}(y_i) P_{\perp L}(y_i)^T \right\|_*
\] (35)
and consequently that (16) holds. That is, Proposition 2.2 is proved when \( p > 1 \).

Proposition 2.2 can be similarly proved when \( \chi_1 = \emptyset \) and \( 0 < p \leq 1 \). Indeed, (34) still holds in this case (\( \chi = \chi_0 \)).

Next, assume that \( p < 1 \). We note that the derivative of \( e_{l_p}(\chi, L(t)) \) at \( t = 0 \) is only defined when \( p \geq 1 \) (indeed, in view of (28) the limit of the derivative in (33) when \( t \to 0 \) and \( z = x_i, i = 1, 2, \ldots, N_1 \), is infinite). To overcome this, we use the following derivative according to the variable \( t^p \):

\[
\frac{d}{dt^p} \left( \text{dist}(z, L(t)^p) \right) \bigg|_{t=0} = \lim_{t \to 0} \left( \frac{t^{1-p}}{p} \cdot \frac{d}{dt} \left( \text{dist}(z, L(t))^p \right) \right).
\]

(37)

It follows from (33), (37), and (25) that

\[
\frac{d}{dt^p} \left( \text{dist}(y_i, L(t))^p \right) \bigg|_{t=0} = \lim_{t \to 0} \left( t^{1-p} \cdot p \cdot \text{dist}(y_i, L(t))^{p-1} \right) \cdot \frac{d}{dt} \text{dist}(y_i, L(t)) \bigg|_{t=0} = 0.
\]

(38)

Furthermore, it follows from (33), (37), and (28) (and also its derivation from (27)) that

\[
\frac{d}{dt^p} \left( \text{dist}(x_i, L(t))^p \right) \bigg|_{t=0} = \left( \lim_{t \to 0} \text{dist}(x_i, L(t))/t \right)^{p-1} \cdot \frac{d}{dt} \text{dist}(x_i, L(t)) \bigg|_{t=0} = \|CV_P L(x_i)\|^p.
\]

(39)

Combining (38) and (39), we obtain that

\[
\frac{d}{dt^p} \left( e_{l_p}(\chi, L(t)) \right) \bigg|_{t=0} = \sum_{i=1}^{N_1} \|CV_P L(x_i)\|^p.
\]

(40)

Now, if \( \text{Sp}([x_i]_{i=1}^{N_1}) = \hat{L} \), then there exists \( 1 \leq j \leq N_1 \) such that \( v_j^T x_j \neq 0 \), and thus \( \|CV_P L(x_i)\| = \|CV_Y x_i\| \geq \theta_1 \|v_j^T x_i\| > 0 \). Combining this observation with (40), we conclude that \( \hat{L} \) is a local minimum of \( e_{l_p}(\chi, L(t)) \) and thus prove Proposition 2.1.

3.3 Proof of Theorem 2.2: Combination of Combinatorial Estimates (Sect. 3.2) with Probabilistic Estimates

To find the probability that \( L_1^* \) is a local \( l_1 \) subspace, we will estimate the probabilities of large LHS and small RHS of (14) for arbitrary \( \hat{L} \in B_G(L_1^*, 1) \). We denote the \( N_1 \)
inliers and $N_0$ outliers by $\{x_i\}_{i=1}^{N_1}$ and $\{y_i\}_{i=1}^{N_0}$, respectively. Due to the homogeneity of (14) in $C$, we will assume without loss of generality that $\|C\|_2 = 1$, i.e., $\theta_1 = 1$.

We start with estimating the probability that the RHS of (14) is small. Applying the above assumption that $\|C\|_2 = 1$, we have that

$$\|CVB_{L_1^*, X}\|_F \leq \|VB_{L_1^*, X}\|_F = \|B_{L_1^*, X}\|_F$$

and consequently

$$\Pr \left( \frac{\|CVB_{L_1^*, X}\|_*}{N_0} < \epsilon \right) \geq \Pr \left( \frac{\|CVB_{L_1^*, X}\|_F}{N_0} < \frac{\epsilon}{\sqrt{d}} \right) \geq \Pr \left( \frac{\|B_{L_1^*, X}\|_F}{N_0} < \frac{\epsilon}{\sqrt{d}} \right) \geq \Pr \left( \frac{\max_{1 \leq p,l \leq d} |(B_{L_1^*, X})_{p,l}|}{N_0} < \frac{\epsilon}{d \sqrt{D}} \right).$$

We further estimate this probability by Hoeffding’s inequality as follows: we view the matrix $B_{L_1^*, X}$ as the sum of random variables $PL_1^*(y_i)P_{L_1^*}^T(y_i)/\|P_{L_1^*}^T(y_i)\|$, $i = 1, \ldots, N_0$. Since the distribution of outliers is uniform on the unit sphere, the coordinates of both $PL_1^*(y_i)$ and $P_{L_1^*}^T(y_i)/\|P_{L_1^*}^T(y_i)\|$ have expectations 0 and take values in $[-1,1]$. We can thus apply Hoeffding’s inequality to the sum defining $B_{L_1^*, X}$ and consequently obtain that

$$\Pr \left( \frac{\max_{1 \leq p,l \leq d} |(B_{L_1^*, X})_{p,l}|}{N_0} < \frac{\epsilon}{d \sqrt{D}} \right) \geq 1 - 2d D \exp \left( - \frac{N_0 \epsilon^2}{2d^2 D} \right). \quad (41)$$

Next, we estimate the probability that the LHS of (14) is sufficiently large. We first note that

$$\sum_{i=1}^{N_1} \|CP_{L_1^*}(x_i)\| \geq \sum_{i=1}^{N_1} |\theta_1 v_1^T P_{L_1^*}(x_i) |$$

$$= \sum_{i=1}^{N_1} |v_1^T P_{L_1^*}(x_i) |$$

$$\geq \sqrt{ \sum_{i=1}^{N_1} |v_1^T P_{L_1^*}(x_i) |^2} \geq \min_t \left( \sum_{i=1}^{N_1} P_{L_1^*}(x_i) P_{L_1^*}(x_i)^T \right). \quad (42)$$

Second of all, since $\mu_1$ is uniform on $L_1^* \cap S^{D-1}$,

$$E_{\mu_1} (P_{L_1^*}(x) P_{L_1^*}(x)^T) = \delta_* I_d, \text{ where } \delta_* = \frac{1}{d}. \quad (43)$$
We will prove in Sect. A.7 the following statement:

if \[ \max_{1 \leq j \leq d} \sigma_j \left( \sum_{i=1}^{N_1} P_{L_i^*}(x_i) P_{L_i^*}(x_i)^T - \delta^*_d I_d \right) < \eta, \]

then \[ \min_{1 \leq j \leq d} \sigma_j \left( \sum_{i=1}^{N_1} P_{L_i^*}(x_i) P_{L_i^*}(x_i)^T \right) > \delta^*_d - \eta. \] (44)

We combine (42)–(44) and Hoeffding’s inequality to obtain the following probabilistic estimate for the LHS of (14):

\[ \Pr \left( \sum_{i=1}^{N_1} \| CV P_{L_i^*}(x_i) \|_{N_1} > \delta^*_d - \eta \right) \]
\[ \geq \Pr \left( \min_{1 \leq j \leq d} \sigma_j \left( \frac{\sum_{i=1}^{N_1} P_{L_i^*}(x_i) P_{L_i^*}(x_i)^T}{N_1} \right) > \delta^*_d - \eta \right) \]
\[ \geq \Pr \left( \max_{1 \leq j \leq d} \sigma_j \left( \frac{\sum_{i=1}^{N_1} P_{L_i^*}(x_i) P_{L_i^*}(x_i)^T}{N_1} - \delta^*_d I_d \right) < \eta \right) \]
\[ \geq \Pr \left( \left\| \sum_{i=1}^{N_1} P_{L_i^*}(x_i) P_{L_i^*}(x_i)^T \right\|_{N_1} - \delta^*_d I_d \right) < \eta \right) \]
\[ \geq \Pr \left( \max_{1 \leq p,l \leq d} \left( \frac{\sum_{i=1}^{N_1} P_{L_i^*}(x_i) P_{L_i^*}(x_i)^T}{N_1} - \delta^*_d I_d \right)_{1 \leq p,l \leq d} \right) < \eta \right) \]
\[ \geq 1 - 2d^2 \exp \left( \frac{-N_1 \eta^2}{2d^2} \right). \] (45)

From (41) and (45), (14) is valid with probability at least

\[ 1 - 2d^2 \exp \left( \frac{-N_1 \eta^2}{2d^2} \right) - 2dD \exp \left( \frac{-N_0 \epsilon^2}{2d^2 D} \right) \quad \forall \epsilon, \eta \text{ s.t. } \eta + \frac{N_0 \epsilon}{N_1} < \frac{1}{d}. \] (46)

We can choose \( \epsilon = N_1/2dN_0, \eta = 1/d\sqrt{4.005} \) and obtain that if \( N_0 = o(N_1^2) \), then (14) is valid with the probability specified in (18).

3.4 Proof of Theorem 1.1: From Local Probabilistic Estimates to Global Ones

3.4.1 Outline of the Proof

The proof verifies three different propositions and then combines them to conclude Theorem 1.1. We use the following notation: For any subspace \( \hat{L} \in G(D, d) \) such that \( \text{dist}_G(\hat{L}, L^*_1) = 1 \), we let \( L(t) : [0, 1] \rightarrow G(D, d) \) denote the parametrization of the
geodesic line from \( L_1^* \) to \( \hat{L} \) such that \( L(0) = L_1^* \) and \( L(1) = \hat{L} \). Using this notation and the setting of Theorem 1.1, the propositions are formulated as follows:

**Proposition 3.1** For any fixed \( 0 < p \leq 1 \), there exists \( \gamma_1 = \gamma_1(p, D, d, \alpha_1, \alpha_0) \) such that w.o.p. for any \( \hat{L} \in G(D, d) \) satisfying \( \text{dist}_G(\hat{L}, L_1^*) = 1 \) with the corresponding geodesic parametrization \( L(t) \) from \( L_1^* \) to \( \hat{L} \):

\[
\left. \frac{d}{dt} \left( \frac{\sum_{x \in \mathcal{X}} \text{dist}(x, L(t))^p}{N} \right) \right|_{t=0} > \gamma_1.
\]

**Proposition 3.2** For any fixed \( 0 < p \leq 1 \), there exists \( 0 < \gamma_2 = \gamma_2(p, D, d, \alpha_1, \alpha_0) < 1 \) such that w.o.p. for any \( t_0 \in [0, \gamma_2] \) and any \( \hat{L} \in G(D, d) \) satisfying \( \text{dist}_G(\hat{L}, L_1^*) = 1 \) with the corresponding geodesic parametrization \( L(t) \) from \( L_1^* \) to \( \hat{L} \):

\[
\left. \frac{d}{dt} \left( \frac{\sum_{x \in \mathcal{X}} \text{dist}(x, L(t))^p}{N} \right) \right|_{t=0} - \left. \frac{d}{dt} \left( \frac{\sum_{x \in \mathcal{X}} \text{dist}(x, L(t))^p}{N} \right) \right|_{t=t_0} < \frac{\gamma_1}{2},
\]

where \( \gamma_1 \) is the constant guaranteed by Proposition 3.1 for this value of \( p \).

**Proposition 3.3** For any fixed \( 0 < p \leq 1 \) and \( \gamma_2 \), the constant guaranteed by Proposition 3.2:

\( L_1^* \) is a global \( l_p \) subspace w.o.p. in \( G(D, d) \setminus B_G(L_1^*, \gamma_2) \).

Equation (50) implies that w.o.p. \( L_1^* \) is the global \( l_p \) subspace in \( B_G(L_1^*, \gamma_2) \). Combining it with (49), we conclude Theorem 1.1.

We prove Proposition 3.1 when \( p = 1 \) in Sect. 3.4.2 and when \( 0 < p < 1 \) in Sect. 3.4.3; Proposition 3.2 in Sect. 3.4.4; and Proposition 3.3 in Sect. 3.4.5. Finally, Sect. 3.4.6 estimates the asymptotic dependence of the overwhelming probability in Theorem 1.1 and the minimal size \( N \) on \( d \) and \( D \).

### 3.4.2 Proof of Proposition 3.1 for \( p = 1 \)

We decompose the sampled data set as follows: \( \mathcal{X} = \bigcup_{i=0}^{K} \mathcal{X}_i \), where \( \mathcal{X}_i \) is the set of points sampled from \( \mu_i \) for all \( 0 \leq i \leq K \). It follows from (14) that the event in (47)
is the same as the event
\[
\sum_{x \in \mathcal{X}} \|CVP_{L_1^*}(x)\|_N - \|CVB_{L_1^*, \mathcal{X} \setminus \mathcal{X}_1}\|_* > \gamma_1 \\
\forall C \in \text{ND}_+(d) \text{ and } V \in \text{O}(d).
\] (51)

We will prove (51) in two steps. In the first step, we will fix matrices \(C_0 \in \text{ND}_+(d)\) and \(V_0 \in \text{O}(d)\) and show that
\[
\sum_{x \in \mathcal{X}_1} \|C_0V_0P_{L_1^*}(x)\| - \|C_0V_0B_{L_1^*, \mathcal{X} \setminus \mathcal{X}_1}\|_* > 2\gamma_1, \\
\text{w.p.} \geq 1 - (2D^2 + 1) \exp(-2N\gamma_1^2),
\] (52)
where
\[
\gamma_1 : = \beta_0 \min_{C \in \text{ND}_+(d), V \in \text{O}(d)} E_{\mu_1} \|CVP_{L_1^*}(x)\|/6 \quad \text{and} \quad \beta_0 = \alpha_1 - \sum_{j=2}^K \alpha_j.
\]

In the second step, we will combine a covering argument and (52) to prove (51).

\textbf{Step 1: Proof of (52)}

We will first verify the following two probabilistic inequalities:
\[
\|C_0V_0B_{L_1^*, \mathcal{X}_0}\|_* < 2\gamma_1 \quad \text{w.p.} \ 1 - 2D^2 \exp(2\gamma_1^2 N)
\] (53)
and
\[
\sum_{x \in \mathcal{X}_1} \|C_0V_0P_{L_1^*}(x)\| - \sum_{x \in \mathcal{X} \setminus (\mathcal{X}_1 \cup \mathcal{X}_0)} \|C_0V_0P_{L_1^*}(x)\|_N > 4\gamma_1, \\
\text{w.p.} \geq 1 - \exp(-2N\gamma_1^2).
\] (54)

\textbf{Part I of Step 1: Proof of (53)}

We define \(J_0(x) = I(x \in \mathcal{X}_0)P_{L_1^*}(x)P_{L_1^*}^T(x)/\text{dist}(x, L_1^*)\). We note that its elements lie in \([-1, 1]\) and \(E_{\mu_0}(J_0(x)) = 0\). Indeed, setting \(R_{L_1^*}(x) = \hat{P}_{L_1^*}(x) - \hat{P}_{L_1^*}(x)\) (i.e., \(R_{L_1^*}(x)\) is the reflection of \(x\) w.r.t. the \(d\)-subspace \(L_1^*\)), we obtain that
\[
2E_{\mu_0}(J_0(x)) = E_{\mu_0}(J_0(x)) - E_{\mu_0}(J_0(R_{L_1^*}(x))) = 0,
\]
where the first equality is clear since \(P_{L_1^*}(x)P_{L_1^*}^T(x) - P_{L_1^*}(R_{L_1^*}(x))P_{L_1^*}^T(R_{L_1^*}(x))^T\) and the second one follows from the symmetry of \(\mu_0\). Therefore, combining the fact
that
\[ D_{i,j} = e_i^T D e_j \leq \max_{u,v \in \mathbb{R}^D} u^T D v / \| u \| \| v \| = \| D \|_*, \]
for any \( D \in \mathbb{R}^{D \times D} \) and 1 \( \leq i, j \leq N \), and Hoeffding’s inequality for the random variable \( J_0(x) \), we establish the following inequality, which clearly implies (53):
\[
\Pr \left( \| \sum_{x \in \mathcal{X}_0} P_{L_1^*}(x) P_{L_1^*}^{-1}(x)^T / \text{dist}(x, L_1^*) \|_* / N < 2\gamma_1 \right) \\
\geq \Pr \left( \| \sum_{x \in \mathcal{X}_0} P_{L_1^*}(x) P_{L_1^*}^{-1}(x)^T / \text{dist}(x, L_1^*) \|_\infty / N < 2\gamma_1 \right) \\
\geq 1 - 2D^2 \exp(2\gamma_1^2 N). \tag{55}
\]

**Part II of Step 1: Proof of (54)**

We define the random variable \( J_1(x) = (I(x \in \mathcal{X}_1) - I(x \in \mathcal{X} \setminus (\mathcal{X}_1 \cup \mathcal{X}_0))) \| C_0 V_0 P_{L_1^*}(x) \| \), and using the spherical symmetry of \( \{\mu_i\}_{i=1}^K \), we have
\[
E_{\mu}(J_1(x)) = E_{\mu_N}\left( \sum_{x \in \mathcal{X}_1} \| C_0 V_0 P_{L_1^*}(x) \| - \sum_{x \in \mathcal{X} \setminus (\mathcal{X}_1 \cup \mathcal{X}_0)} \| C_0 V_0 P_{L_1^*}(x) \| \right) \\
= \alpha_1 E_{\mu_1} \| C_0 V_0 P_{L_1^*}(x) \| - \sum_{j=2}^K \alpha_j E_{\mu_j} \| C_0 V_0 P_{L_1^*}(x) \| \\
\geq \alpha_1 E_{\mu_1} \| C_0 V_0 P_{L_1^*}(x) \| - \sum_{j=2}^K \alpha_j E_{\mu_1} \| C_0 V_0 P_{L_1^*}(x) \| \\
= \beta_0 E_{\mu_1} \| C_0 V_0 P_{L_1^*}(x) \| \\
\geq 6\gamma_1. \tag{56}
\]

We conclude (54) by applying Hoeffding’s inequality to the random variable \( J_1(x) \), while using the facts that its expectation is larger than \( 6\gamma_1 \) and its values are in \([-1, 1]\).

**Part III of Step 1: Conclusion of (52) via (53) and (54)**

We first observe that
\[
\| C_0 V_0 B_{L_1^*, \mathcal{X}\setminus \mathcal{X}_1} \|_* \leq \| C_0 V_0 B_{L_1^*, \mathcal{X} \setminus (\mathcal{X}_1 \cup \mathcal{X}_0)} \|_* + \| C_0 V_0 B_{L_1^*, \mathcal{X}_0} \|_* \tag{57}
\]
Applying (57) and (58), we bound the LHS of (52) by the difference between the LHS of (53) and the LHS of (54) as follows:

\[
\sum_{x \in X} \left| \langle C_0 V_0 P_{L_1^*}(x) - C_0 V_0 B_{L_1^*}(x), x \rangle \right| = \sum_{x \in X} \left| \langle C_0 V_0 P_{L_1^*}(x) P_{L_1^*}(x)^T / \| P_{L_1^*}(x) \|_* \rangle \right| \\
\leq \sum_{x \in X} \left| \langle C_0 V_0 B_{L_1^*}(x), x \rangle \right|.
\]

Equation (52) is thus an immediate consequence of (53), (54), and (59).

**Step 2: Conclusion of (51) via (52) and a covering argument**

We recall that (51) needs to be verified for all matrices \( C \in \text{ND}_+(d) \) and \( V \in \text{O}(d) \). We define

\[
\text{dist}_{\text{ND}_+(d) \times \text{O}(d)}((C_1, V_1), (C_2, V_2)) := \max(\|C_1 - C_2\|_2, \|V_1 - V_2\|_2)
\]

and note that whenever \( \text{dist}_{\text{ND}_+(d) \times \text{O}(d)}((C_1, V_1), (C_2, V_2)) < \gamma_1 / 2 \) and \( x \in B(0, 1) \), we have that

\[
\|C_1 V_1 P_{L_1^*}(x) - C_2 V_2 P_{L_1^*}(x)\| \\
= (\|C_1 V_1 P_{L_1^*}(x)\| - \|C_2 V_1 P_{L_1^*}(x)\|) + (\|C_2 V_1 P_{L_1^*}(x)\| - \|C_2 V_2 P_{L_1^*}(x)\|) \\
\leq \|C_1 - C_2\|_2 + \|C_2\|_2 \|V_1 - V_2\|_2 \\
\leq \gamma_1.
\]
Combining (52) and (61), we obtain that for \((C, V)\) in a ball in \(\text{ND}_+(d) \times O(d)\) of radius \(\gamma_1/2\) and center \((C_0, V_0)\),

\[
\sum_{x \in X_1} \| CV P_{L_1^+}(x) \| - \sum_{x \in X_1 \setminus X_1} \| CV P_{L_1^+}(x) \| > \gamma_1 \text{ w.p. } 1 - \exp(-2N\gamma_1^2).
\]

(62)

We easily extend (62) for all pairs of matrices \((C, V)\) in the compact space \(\text{ND}_+(d) \times O(d)\) (with the distance specified in (60)). Indeed, it follows from [36, Theorem 7] that \(O(d)\) can be covered by \(C_1^{d(d-1)/2}/(\gamma_1/2)^{d(d-1)/2}\) balls of radius \(\gamma_1/2\) for some \(C_1' > 0\) (note that the dimension of \(O(d)\) is \(d(d-1)/2\)). Since \(\text{ND}_+(d)\) is isomorphic to \(\mathbb{S}^{d-1}\), it follows from [39, Lemma 5.2] that it can be covered by \(3^d/(\gamma_1/2)^d\) balls of radius \(\gamma_1/2\). Therefore, the product space \(\text{ND}_+(d) \times O(d)\) with norm defined in (60) can be covered by \(C_1^{d(d+1)/2}/(\gamma_1/2)^{d(d+1)/2}\) balls of radius \(\gamma_1/2\), where \(C_1 := \max(C_1', 3)\), and consequently

\[
(51) \text{ is valid for any } C \in \text{ND}_+(d) \text{ and } V \in O(d)
\]

w.p. \(1 - C_1^{d(d+1)/2} \exp(-2N\gamma_1^2)/(\gamma_1/2)^{d(d+1)/2}\),

(63)

which means that (47) with \(p = 1\) holds with the probability specified in (63).

3.4.3 Proof of Proposition 3.1 for \(0 < p < 1\)

When \(0 < p < 1\), it follows from (40) and Hoeffding’s inequality that (47) holds for any \(C \in \text{ND}_+(d)\) and \(V \in O(d)\) w.p. \(1 - \exp(-2N\gamma_1^2)\), where

\[
\gamma_1 := \alpha_1 \cdot \min_{C \in \text{ND}_+(d), V \in O(d)} E_{\mu_0} (\| CV P_{L_1^+}(x) \|^p)/2.
\]

Following the same covering argument as in the proof of (63), we conclude that (47) holds with the same probability specified in (63) (though \(\gamma_1\) is defined differently for \(p = 1\) and \(0 < p < 1\)).

3.4.4 Proof of Proposition 3.2

We verify (48) by separating \(X\) into three parts: \(X_1 = X \cap L_1^+, \hat{X} := \{ x \in X \setminus X_1 : \text{dist}(x, L_1^+) \leq 2\gamma_3\} (\gamma_3 \text{ will be clarified later})\), and \(X' \setminus (X_1 \cup \hat{X})\). Specifically, we will prove that there exists \(0 < \gamma_2 < 1\) such that for any \(t_0 \in [0, \gamma_2]\),

\[
\frac{1}{N} \sum_{x \in X_1} \left( \frac{d}{dt} \text{dist}(x, L(t))^p \right|_{t=0} - \frac{d}{dt} \text{dist}(x, L(t))^p \right|_{t=t_0} \right) < \frac{\gamma_1}{6},
\]

(64)

\[
\frac{1}{N} \sum_{x \in X'} \left( \frac{d}{dt} \text{dist}(x, L(t))^p \right|_{t=0} - \frac{d}{dt} \text{dist}(x, L(t))^p \right|_{t=t_0} \right) < \frac{\gamma_1}{6}.
\]

(65)
and
\[
\frac{1}{N} \sum_{x \in \mathcal{X} \setminus (\mathcal{X}_1 \cup \mathcal{X}_2)} \left( \frac{d}{dt^p} \text{dist}(x, L(t))^p \right)_{t=0} - \frac{d}{dt^p} \text{dist}(x, L(t))^p \bigg|_{t=0} < \frac{\gamma_1}{6}. \quad (66)
\]

We prove (64) and (66) deterministically and (65) w.o.p. Then (48) follows from the summation of (64), (65) and (66).

In order to prove (64), we uniformly bound from above the terms of the sum in (64) by a term of order \(O(t_0^2)\). For simplicity, let us assume that \(p = 1\). It follows from (27) and the fact that the sinc function is decreasing that for any \(x \in \mathcal{X}_1\) and any \(0 \leq t_0 \leq 1\), the derivative \(\frac{d}{dt} \text{dist}(x, L(t))\) at \(t = t_0\) is bounded below by
\[
\frac{\sin t_0}{t_0} \sum_{j=1}^{d} \theta_j |v_j \cdot x|^2 t_0 \theta_j \cos t_0 \leq \frac{\sin t_0 \cos t_0}{t_0} \sum_{j=1}^{d} (v_j \cdot x)^2 \theta_j^2. \quad (67)
\]

We note that \(\sqrt{\sum_{j=1}^{d} (v_j \cdot x)^2} \theta_j^2 \leq 1\) (indeed, the assumption \(\text{dist}_G(\hat{L}, L^*_1) \leq 1\) implies that \(\sum_{i=1}^{d} \theta_i^2 = 1\)). Combining this observation with (28) and (67) we derive the following bound on the terms in the sum of (64) when \(p = 1\):
\[
\frac{d}{dt^p} \text{dist}(x, L(t)) \bigg|_{t=0} - \frac{d}{dt^p} \text{dist}(x, L(t)) \bigg|_{t=0} \leq \left( 1 - \frac{\sin t_0 \cos t_0}{t_0} \right) = O(t_0^2). \quad (68)
\]

Similarly, one can also uniformly bound these terms by an \(O(t_0^2)\) term when \(p < 1\). Therefore, one can choose a sufficiently small \(\gamma_2\) such that (64) holds.

Next, we verify (65). Here we can bound the terms of the sum in (65) by 2 (using an additional assumption; see below). However, we cannot bound them by a term that approaches zero when \(t_0\) approaches zero. We thus control w.o.p. the fraction of the cardinality of \(\hat{X}\) over \(N\) by a sufficiently small constant. We fix \(\gamma_3 \equiv \gamma_3(D, d, \gamma_1) \equiv \gamma_3(D, d, \alpha_0, \alpha_1, p)\) a sufficiently small constant such that
\[
\mu(x \in S^{D-1} : 0 < \text{dist}(x, L^*_1) \leq 2 \gamma_3) \leq \gamma_1/24. \quad (69)
\]

By applying Hoeffding’s inequality to the indicator function of \(\hat{X}\), \(I_{\hat{X}}(x)\), while using the facts that \(E(I_{\hat{X}}(x)) = \mu(x : x \in \hat{X}) \leq \gamma_1/24\) and \(I_{\hat{X}}(x)\) takes values in \([0, 1]\), we bound the size of \(\hat{X}\) as follows:
\[
\frac{\#(\hat{X})}{N} = \frac{\#(\hat{X})}{\#(\hat{X})} \leq \frac{\gamma_1/12}{\#(\hat{X})} \leq \frac{\gamma_1/24}{N} \leq \frac{\gamma_1/12}{\#(\hat{X})} = \gamma_1/12. \quad (70)
\]

Now for \(x \in \hat{X}\), we claim that the derivative according to \(t^p\) of \(\text{dist}(x, L(t))^p\) takes values in \([-1, 1]\) (this requires an additional assumption when \(p < 1\). When
3.4.5 Proof of Proposition 3.3

Applying Lemma 3.3, we obtain that for all $2 \leq i \leq K$,

$$E_{\mu_1} (\text{dist}(x, L)^p - \text{dist}(x, L_1^*)^p) + E_{\mu_i} (\text{dist}(x, L)^p - \text{dist}(x, L_1^*)^p) \geq 0. \quad (72)$$

Further application of Lemma 3.1 with $L \in G(D, d) \setminus B_G(L_1^*, \gamma_2)$ results in the inequality

$$E_{\mu_1} (\text{dist}(x, L)) > 0.88 \cdot 2^{3p} \cdot \pi^{-(2p+1)/2} \cdot (d + p)^{-p/2} \cdot \gamma_2^p. \quad (73)$$

Now, combining (72) and (73), we have that

$$E_{\mu} (\text{dist}(x, L)^p - \text{dist}(x, L_1^*)^p)$$

$$= \sum_{i=2}^{K} \alpha_i ((E_{\mu_1} (\text{dist}(x, L)^p - \text{dist}(x, L_1^*)^p) + E_{\mu_i} (\text{dist}(x, L)^p - \text{dist}(x, L_1^*)^p))$$

$$+ \beta \gamma E_{\mu_1} (\text{dist}(x, L)^p - \text{dist}(x, L_1^*)^p) \geq \beta_0 \cdot 0.88 \cdot 2^{3p} \cdot \pi^{-(2p+1)/2} \cdot (d + p)^{-p/2} \cdot \gamma_2^p. \quad (74)$$
We define
\[ \gamma_4 = 0.88 \cdot 2^{3p} \cdot \pi^{-(2p+1)/2} \cdot (d + p)^{-p/2} \cdot \gamma_2^p \] (75)
and note that it depends on \( d, D, K, \alpha_0, \alpha_1, \) and \( \min_{2 \leq i \leq K} (\text{dist}_G(L_i^+, L_i^+)) \). Applying Hoeffding’s inequality to \( \text{dist}(x, L) - \text{dist}(x, L_i^+) \), whose absolute values are uniformly bounded by 1 and its expectation is at least \( \gamma_4 \) (which follows from (74) and (75)), we obtain that for any \( L \in G(D, d) \setminus B_G(L_1^+, \gamma_2) \),
\[ e_{lp}(\mathcal{X}, L) - e_{lp}(\mathcal{X}, L_1^+) > \gamma_4 N/2 \text{ w.p. } \geq 1 - \exp(-N \gamma_4^2/8). \] (76)

By Lemma 3.2, we have that for any \( L' \in G(D, d) \) satisfying \( \text{dist}_G(L, L') < (\gamma_4/4)^{1/p} \) and any \( x \in B(0, 1) \),
\[ |\text{dist}(x, L')^p - \text{dist}(x, L)^p| < \gamma_4/4. \]

Consequently, for any \( L \in G(D, d) \setminus B_G(L_1^+, \gamma_2) \) and all \( L' \in B_G(L_1^+, \gamma_4/4)^{1/p} \),
\[ e_{lp}(\mathcal{X}, L') - e_{lp}(\mathcal{X}, L_1^+) > 0 \text{ w.p. } \geq 1 - \exp(-N \gamma_4^2/8). \] (77)

We can cover \( G(D, d) \setminus B_G(L_1^+, \gamma_2) \) by \( (C_2 \sqrt{d})^{d(D-d)}/\gamma_4^{d(D-d)/p} \) balls of radius \( (\gamma_4/4)^{1/p} \) (this follows from Remark 8.4 of [35]). Now, for each such ball, we have that (76) is valid for its center w.p. \( 1 - \exp(-N \gamma_4^2/8) \) and consequently (77) is valid for subspaces in that ball with the same probability. We thus conclude that
\[ (49) \text{ holds w.p. } 1 - \exp(-N \gamma_4^2/8)(C_2 \sqrt{d})^{d(D-d)/p} / \gamma_4^{d(D-d)}. \] (78)

### 3.4.6 Dependence of the Probability and \( N \) on \( d \) and \( D \)

Applying the union bound for the events specified in (47), (48), and (49), whose probabilities are specified in (63), (70), and (78), respectively, we conclude that \( L_1^+ \) is a global \( l_1 \) subspace w.p. at least
\[ 1 - C_1^{d(d+1)/2} \exp(-2N \gamma_1^2/(\gamma_1/2)^{d(d+1)/2}) - \exp(-2N \gamma_1/24)^2 - \exp(-N \gamma_4^2/8)(C_2 \sqrt{d})^{d(D-d)/p} / \gamma_4^{d(D-d)}. \] (79)

We bound (79) from below by \( 1 - C' \exp(-N/C) \), where \( C = 1/ \min(2(\gamma_1/24)^2, \gamma_4^2/8) \) and \( C' = C_1^{d(d+1)/2}/(\gamma_1/2)^{d(d+1)/2} + 1 + (C_2 \sqrt{d})^{d(D-d)/p} / \gamma_4^{d(D-d)}. \)

We cannot formulate nice expressions for \( \gamma_1 \) and \( \gamma_4 \); however, we can express their dependence on \( D \) and \( d \) as follows (assuming the rest of the parameters are fixed). The definition of \( \gamma_1 \) in (56) implies that it is in the order of \( 1/d \). In order to estimate \( \gamma_4 \), we first need to estimate \( \gamma_3 \) and \( \gamma_2 \). The defining equation of \( \gamma_3 \), i.e., (69), implies that \( \gamma_3 \) is in the order of \( d^{-1} D^{-1/2} \) (a rigorous argument appears in Sect. A.2). We claim that \( \gamma_2 \) is in the order of \( d^{-6} D^{-1.5} \). Indeed, when proving (64), we
required that $\gamma_2 C(d)/\gamma_2^3 < \gamma_1/6$ and $C(d) = O(d^2)$. Finally, applying (75) and the estimate above for $\gamma_2$, we conclude that $\gamma_4$ is in the order of $d^{-6.5} p D_{-1.5p}$. Therefore, $C = O(d^\max(13p,2) D_{-3p})$ and $C' = O(d^{d(d+1)/2} + d^{6.5d(D-d)} D_{1.5d(D-d)})$.

We can use these estimates for $C$ and $C'$, and thus for the probability $1 - C' \exp(-N/C)$, to obtain an estimate of the dependence of the minimal size $N$ on $D$ and $d$ in the asymptotic case. Assume that $N(D \to \infty$ and $N/(d^\max(13p,2)+1) D_{3p}$ max$(D-d, d+1) \log(D)) \to \infty$, then the probability $1 - C' \exp(-N/C)$ approaches 0. That is, asymptotically $N = \Omega(d^\max(13p,2)+1) D_{3p}$ max$(D-d, d+1) \log(D))$. This estimate, which indicates significant oversampling for the single subspace recovery, is not tight, and tighter estimates are left for future work.

3.5 Proof of Theorem 1.2: Stability Analysis

3.5.1 Reduction of Theorem 1.2

We first explain how to reduce the proof of Theorem 1.2 when $0 < p \leq 1$ to the verification of a simpler statement. We then adapt this idea for proving the same theorem when both $p > 1$ and $K = 1$.

In order to prove Theorem 1.2 when $0 < p \leq 1$, i.e., prove that the global minimum of $e_{L_p}(X, L)$ is in $B_G(L^*_1, f)$ w.o.p., we only need to show that there exists a constant $\rho_1 > 0$ such that for any $L \notin B_G(L^*_1, f)$,

$$E_{\mu_\epsilon}(e_{L_p}(x, L)) > E_{\mu_\epsilon}(e_{L_p}(x, L^*_1)) + \rho_1. \quad (80)$$

Indeed, we cover the compact space $G(D, d) \setminus B_G(L^*_1, f)$ with small balls of radius $\rho_1/2$. Then by using (80) and Hoeffding’s inequality, we obtain that $e_{L_p}(X, L) > e_{L_p}(X, L^*_1)$ for any $L$ in each such ball w.o.p. Therefore, $e_{L_p}(X, L) > e_{L_p}(X, L^*_1)$ for $L \in G(D, d) \setminus B_G(L^*_1, f)$ w.o.p. Equivalently, $G(D, d) \setminus B_G(L^*_1, f)$ does not contain the global minimum of $e_{L_p}(X, L)$ w.o.p. By a similar argument as in Sect. 3.4.5, the probability is at least $1 - \exp(-N\rho_1^2/8)(C_2 \sqrt{d}) d^{d(D-d)}/p / \rho_1^d(D-d)$.

We will prove (80) with

$$\rho_1 = 2\epsilon^p \quad (81)$$

and thus obtain the probability specified in (7).

We further reduce (80) by using the measure $\mu$ instead of $\mu_\epsilon$ (see Sect. 1.4). Combining the triangle inequality and the concavity of $x^p$, we obtain that

$$|E_{\mu_\epsilon + \nu_\epsilon}(e_{L_p}(x, L)) - E_{\mu_\epsilon}(e_{L_p}(x, L))| = |E_{\mu_\epsilon + \nu_\epsilon}(\|P_{L^*_1}(x)\|^p - \|P_{L^*_1}^+(\hat{P}_{L^*_1}(x))\|^p)|$$

$$\leq E_{\mu_\epsilon + \nu_\epsilon} \|P_{L^*_1}^+(\hat{P}_{L^*_1}(x))\|^p$$

$$\leq E_{\mu_\epsilon + \nu_\epsilon} \|P_{L^*_1}(x)\|^p$$

$$= E_{\nu_\epsilon} \|x\|^p \leq \epsilon^p. \quad (82)$$
Summing (82) over all $1 \leq i \leq K$, we have
\[ |E_{\mu_i}(e_{l_p}(x, L)) - E_{\mu}(e_{l_p}(x, L))| \leq \epsilon^p. \tag{83} \]

Hence, in order to prove (80) and thus Theorem 1.2 for $p \leq 1$, the following equation is sufficient:
\[ E_{\mu}(e_{l_p}(x, L)) > E_{\mu}(e_{l_p}(x, L^*_1)) + \rho_1 + 2\epsilon^p, \quad \text{for any } L \in G(D, d) \setminus B_G(L^*_1, f). \tag{84} \]

We can similarly reduce Theorem 1.2 when $K = 1$ and $p > 1$. However, (82) needs to be modified since $x^p$ is not concave when $p > 1$. For this purpose, we note that for any $x_1, x_2 \in B(0, 1)$,
\[ \text{dist}(x_1, L^*_1)^p - \text{dist}(x_2, L^*_1)^p < 1 - (1 - \text{dist}(x_1, x_2))^p < p \cdot \text{dist}(x_1, x_2). \tag{85} \]

Indeed, when $p = 1$, (85) is immediate (it is equivalent to $\|P_{L^*_1}(x_2 - x_1)\| \leq \|x_2 - x_1\|$), and it extends to $p > 1$ by the following proposition: if $0 \leq y_1, y_2 \leq 1$, $y_1 - y_2 < \eta$ and $p > 1$, then $y_1^p - y_2^p < 1 - (1 - \eta)^p$. Combining (85) with the derivation of (82), we conclude the following analog of (82) in the current case:
\[ |E_{\mu_i}(e_{l_p}(x, L)) - E_{\mu}(e_{l_p}(x, L))| \leq p \cdot \epsilon. \tag{86} \]

Consequently, we reduce (80) and (81) (and thus Theorem 1.2) when $K = 1$ and $p > 1$ to the following equations:
\[ E_{\mu}(e_{l_p}(x, L)) > E_{\mu}(e_{l_p}(x, L^*_1)) + \rho_1 + 2 p \epsilon, \quad \text{for any } L \in G(D, d) \setminus B_G(L^*_1, f) \tag{87} \]

and
\[ \rho_1 = 2 \cdot p \cdot \epsilon. \tag{88} \]

### 3.5.2 Proof of (84) and (87) and Conclusion of Theorem 1.2

We arbitrarily fix $L \in G(D, d) \setminus B_G(L^*_1, f)$. We assume first that $0 < p \leq 1$ and apply Lemma 3.3 to obtain that
\[
\begin{align*}
E_{\mu - (\alpha_1 - \sum_{i=2}^K \alpha_i)\mu_1} e_{l_p}(x, L) &- E_{\mu - (\alpha_1 - \sum_{i=2}^K \alpha_i)\mu_1} e_{l_p}(x, L^*_1) \\
&= \sum_{i=2}^K \alpha_i \left( E_{\mu_1 + \mu_i} e_{l_p}(x, L) - E_{\mu_1 + \mu_i} e_{l_p}(x, L^*_1) \right) \\
&\geq 0.
\end{align*}
\]

\( \square \) Springer
Consequently, we prove (84) with $\rho_1 = 2\epsilon_p$ by Lemma 3.1 as follows:

\[
E_\mu(e_{l_p}(x, L)) - E_\mu(e_{l_p}(x, L_1^*)) \\
\geq \left( \alpha_1 - \sum_{i=2}^{K} \alpha_i \right) E_{\mu_1}(e_{l_p}(x, L)) \\
\geq \left( \alpha_1 - \sum_{i=2}^{K} \alpha_i \right) \cdot f_p \cdot 2^3 p/2 \cdot 0.88 \\
= 4\epsilon_p,
\]

where the second inequality applies Lemma 3.1 and the last equality uses the fact that the term $\alpha_0 + 2 \cdot \alpha_1 - 1$ in the definition of $f$ equals $(\alpha_1 - \sum_{i=2}^{K} \alpha_i)$. Equation (6) is obtained by solving for $f$ in the last equality of (89).

Equation (87) (with $p > 1$) follows from the same argument of (89), where $\epsilon_p$ is now replaced by $p\epsilon$. Equation (8) is deduced in a similar way to (6), while using (88) instead of (81).

3.6 Proof of Theorem 1.3: Symmetry Arguments

3.6.1 Structure of the Proof

We proceed with several reductions of the statement of the theorem. The first reduction (see Sect. 3.6.2) practically states that it is enough to prove w.p. 1 (under the measure $\gamma_{K}^{D,d}$) that $L_1^*$ is not a global $l_p$ subspace in expectation, or equivalently, $L_1^* \neq \arg \min_{L \in G(D,d)} E_\mu(e_{l_p}(x, L))$. In order to be able to prove this, we condition the probability measure on other events and thus “reduce randomness”. In the second reduction (see Sect. 3.6.3), we condition on $L_1^*, L_3^*, L_4^*, \ldots, L_K^*$, and in the third reduction, (see Sect. 3.6.4) we condition on the principal vectors and principal angles of $L_2^*$. We then prove the final reduced statement in Sect. 3.6.5. Finally, Sect. 3.6.6 estimates the sizes of $\delta_0$ and $\kappa_0$, and Sect. 3.6.7 uses the results of this section to show that exact asymptotic recovery is impossible in our setting with $K > 1$ and any noise level $\epsilon > 0$.

3.6.2 First Reduction of Theorem 1.3

Theorem 1.3 states that the global $l_p$ subspace is not in $B_G(L_1^*, \kappa_0)$ w.o.p. for almost every $\{L_1^*\}_{i=1}^{K} \in G(D, d)^K$. We claim that it reduces to (or equivalently, implied by) the following statement:

\[
\gamma_{K}^{D,d}\left( \{L_1^*\}_{i=1}^{K} \subset G(D, d) : L_1^* = \arg \min_{L \in G(D,d)} E_\mu(e_{l_p}(x, L)) \right) = 0. \tag{90}
\]

Indeed, if (90) is satisfied, then for $L_0 = \arg \min_{L \in G(D,d)} E_\mu(e_{l_p}(x, L))$ and any $K$ $d$-subspaces $\{L_1^*\}_{i=1}^{K}$ in a subset of $G(D, d)^K$ with nonzero $\gamma_{K}^{D,d}$ measure, the constant
\[ \zeta_1 := E_{\mu}(e_{l_p}(x, L_1^*)) - E_{\mu}(e_{l_p}(x, L_0)) \]

is positive.

For any \( L^* \in B_G(L_1^*, \kappa_0) \) and \( x \in \text{supp}(\mu) \subseteq B(0, 1) \),

\[
\text{dist}(x, L^*)^p - \text{dist}(x, L_1^*)^p \leq 1^p - (1 - \text{dist}_G(L^*, L_1^*))^p \leq p \cdot \text{dist}_G(L^*, L_1^*),
\]

and therefore

\[
E_{\mu}(e_{l_p}(x, L^*)) > E_{\mu}(e_{l_p}(x, L_1^*)) - \kappa_0 \cdot p.
\] (91)

Letting \( \delta_0 = \kappa_0 = \frac{\zeta_1}{4}p \epsilon \), we obtain from (86) (using the fact that \( \epsilon < \delta_0 \)) and (91) that

\[
E_{\mu}(e_{l_p}(x, L^*)) - E_{\mu}(e_{l_p}(x, L_1^*)) > E_{\mu}(e_{l_p}(x, L_0)) - E_{\mu}(e_{l_p}(x, L_0)) - 2\delta_0 p > E_{\mu}(e_{l_p}(x, L_1^*)) - E_{\mu}(e_{l_p}(x, L_0)) - 2\delta_0 p - \kappa_0 p = \frac{\zeta_1}{4}.
\]

Therefore, by Hoeffding’s inequality:

\[
e_{l_p}(X, L^*) - e_{l_p}(X, L_0) > \frac{\zeta_1 N}{8} \text{ w.p. } 1 - \exp(-\frac{N\zeta_1^2}{32}).
\] (92)

At last, we prove w.o.p. that

\[
e_{l_p}(X, L^*) - e_{l_p}(X, L_0) > 0 \text{ for all } L^* \in B_G(L_1^*, \kappa_0).
\] (93)

To do this, we cover \( B_G(L_1^*, \kappa_0) \) with small balls of radius \( \zeta_1/16 \) so that \( e_{l_p}(X, L) > e_{l_p}(X, L_0) \) for all \( L \) in each such ball w.o.p. Therefore, \( e_{l_p}(X, L) > e_{l_p}(X, L_0) \) for all \( L \in B_G(L_1^*, \kappa_0) \) w.o.p. Equivalently, \( B_G(L_1^*, \kappa_0) \) will not contain the global minimum of \( e_{l_p}(X, L) \) w.o.p. This implies Theorem 1.3.

We note that the number of covering balls can be the \((\zeta_1/16)\)-covering number of \( G(D, d) \), which is \((C_2\sqrt{d})^{D(D-d)} / (\zeta_1/16)^{D(D-d)} \) (see Section 3.4.5). The combination of this observation with the probabilistic estimate in (92) implies the following expression for the probability of (93) (which is the unspecified failure probability of (1.3)):

\[
1 - \frac{(C_2\sqrt{d})^{D(D-d)}}{(\zeta_1/16)^{D(D-d)}} \exp\left(-\frac{N\zeta_1^2}{32}\right),
\] (94)

where \( \zeta_1 \) is later estimated in (105).

3.6.3 Second Reduction of Theorem 1.3

We define the operator

\[
D_{L,x,p} = P_L(x)P_L^T \text{dist}(x, L)^{(p-2)}
\] (95)
and the function

\[ h(L_i^*, L_i^*) = E_{\alpha_i}(D_{L_i^*}, x, p), \quad 0 \leq i \neq 1 \leq K. \]

In view of Proposition 2.2, (90) follows from the condition

\[ \gamma_{D,d}^K \left( \{L_i^*\}_{i=1}^K \subset G(D, d) : E_{\mu} \left( D_{L_i^*}, x, p \right) = 0 \right) = 0, \quad (96) \]

which we rewrite as follows:

\[
\gamma_{D,d}^K \left( \{L_i^*\}_{i=1}^K \subset G(D, d) : E_{\mu} \left( D_{L_i^*}, x, p \right) = 0 \right) \\
= \gamma_{D,d}^K \left( \{L_i^*\}_{i=1}^K \subset G(D, d) : \sum_{i=0}^{K} \alpha_i \mu_i \left( D_{L_i^*}, x, p \right) = 0 \right) \\
= \gamma_{D,d}^K \left( \{L_i^*\}_{i=1}^K \subset G(D, d) : \sum_{i=0}^{K} \alpha_i h(L_i^*, L_i^*) = 0 \right) = 0. \quad (97)
\]

Since \( \{L_i^*\}_{i=1}^K \) are identically and independently distributed according to \( \gamma_{D,d} \), Fubini’s Theorem implies that (97) follows from the equation

\[ \gamma_{D,d} \left( L_2^* \in G(D, d) : h(L_1^*, L_2^*) = H(L_1^*, L_3^*, \ldots, L_K^*) \right) = 0, \quad (98) \]

where

\[ H(L_1^*, L_3^*, \ldots, L_K^*) = -\sum_{i=0}^{K} \frac{\alpha_i h(L_1^*, L_i^*)}{\alpha_2}. \quad (99) \]

3.6.4 Third Reduction of Theorem 1.3

We denote the principal angles between \( L_2^* \) and \( L_1^* \) by \( \{\theta_j\}_{j=1}^d \), the principal vectors of \( L_2^* \) and \( L_1^* \) by \( \{\hat{v}_j\}_{j=1}^d \) and \( \{v_j\}_{j=1}^d \), respectively, and the complementary orthogonal system for \( L_2^* \) w.r.t. \( L_1^* \) by \( \{u_j\}_{j=1}^d \). Note that \( h(L_1^*, L_2^*) \), as a function of \( x \), maps \( \text{Sp}(\{u_j\}_{i=1}^d) \) to \( \text{Sp}(\{v_j\}_{i=1}^d) \). Now, transforming \( x \in L_2^* \cap B(0, 1) \) to \( \{a_i\}_{i=1}^d \) in a \( d \)-dimensional unit ball by \( x = \sum_{i=1}^d a_i \hat{v}_i \), we have that for any \( 1 \leq i_1, i_2 \leq d \),

\[
\mathbf{v}_{i_1}^T h(L_1^*, L_2^*) \mathbf{u}_{i_2} = E_{\mu_2} \left( \mathbf{v}_{i_1}^T \hat{P}_{L_1^*}(x) \hat{P}_{L_2^*}^\perp(x) \mathbf{u}_{i_2} \text{dist}(x, L_1^*)^{p-2} \right) \\
= \int_{\sum_{i=1}^d a_i^2 \leq 1} \cos(\theta_{i_1}) a_{i_1} \sin(\theta_{i_2}) a_{i_2} \left( \sum_{i=1}^d a_i^2 \sin^2 \theta_i \right)^{p-2} \text{d}\mu_2.
\]
When $i_1 \neq i_2$, the function

$$\cos(\theta_{i_1})a_{i_1} \sin(\theta_{i_2})a_{i_2} \left( \sum_{i=1}^{d} a_i^2 \sin^2 \theta_i \right)^{\frac{p-2}{4}}$$

is odd w.r.t. $a_{i_1}$, and consequently

$$v_{i_1}^T h(L_1^*, L_2^*) u_{i_2} = \int_{\sum_{i=1}^{d} a_i^2 \leq 1} \cos(\theta_{i_1})a_{i_1} \sin(\theta_{i_2})a_{i_2} \left( \sum_{i=1}^{d} a_i^2 \sin^2 \theta_i \right)^{\frac{p-2}{4}} d\mu_2 = 0.$$  

Therefore, when we form $V$ and $U$ as in (26), the $d \times d$ matrix $h(L_1^*, L_2^*)^T U$ is diagonal with the elements

$$\int_{\sum_{i=1}^{d} a_i^2 \leq 1} \cos(\theta_j) \sin(\theta_j) a_j^2 \left( \sum_{i=1}^{d} a_i^2 \sin^2 \theta_i \right)^{\frac{p-2}{4}} d\mu_2, \quad j = 1, \ldots, d.$$  

Notice that $V h(L_1^*, L_2^*) = h(L_1^*, L_2^*) = h(L_1^*, L_2^*)^T U$ and that $h(L_1^*, L_2^*)$ has the following singular values, where $j = 1, \ldots, d$:

$$\lambda_j(h(L_1^*, L_2^*)) = \int_{\sum_{i=1}^{d} a_i^2 \leq 1} \cos(\theta_j) \sin(\theta_j) a_j^2 \left( \sum_{i=1}^{d} a_i^2 \sin 2\theta_i \right)^{\frac{p-2}{4}} d\mu_2.$$  

We arbitrarily fix $L_1^*, L_2^*, L_4^*, \ldots, L_K^*$ and denote the singular values of $H$ (which is defined in (99)) by $\{\sigma_i\}_{i=1}^{D}$ and observe that (98) is implied by the following equation:

$$\gamma_{D, d} \left( L_2^* \in G(D, d) : \lambda_1(h(L_1^*, L_2^*)) \in \{\sigma_i\}_{i=1}^{D} \right) = 0,$$

which we express as:

$$\gamma_{D, d} \left( \int_{\sum_{i=1}^{d} a_i^2 \leq 1} \cos(\theta_1) \sin(\theta_1) a_1^2 \left( \sum_{i=1}^{d} a_i^2 \sin^2 \theta_i \right)^{\frac{p-2}{4}} d\mu_2 \in \{\sigma_i\}_{i=1}^{D} \right) = 0.$$  

(101)
3.6.5 Proof of (101) and Conclusion of Theorem 1.3

We first conclude (101) when \( p = 2 \). In this case,

\[
\lambda_1(h(L_1^*, L_2^*)) \equiv \int \cos(\theta_1) \sin(\theta_1) a_1^2 \left( \sum_{i=1}^{d} a_i^2 \sin^2 \theta_i \right)^{\frac{p-2}{2}} d\mu_2
\]

is a monotone function of \( \theta_1 \) on \([0, \pi/4]\) as well as \([\pi/4, \pi/2]\). That is, the requirement that \( \lambda_1(h(L_1^*, L_2^*)) \in \{\sigma_i\}_{i=1}^{D} \) can occur only at discrete values of \( \theta_1 \) (at most \( 2D \)) and consequently has \( \gamma_{D,d} \) measure 0, that is, (101) (and consequently (90)) is verified in this case.

If \( p \neq 2 \) and \( \{\theta_i\}_{i=1}^{d-1} \) are fixed, then

\[
\int \cos(\theta_1) \sin(\theta_1) a_1^2 \left( \sum_{i=1}^{d} a_i^2 \sin^2 \theta_i \right)^{\frac{p-2}{2}} d\mu_2
\]

is a monotone function of \( \theta_d \). Following a similar argument, we obtain that

\[
\gamma_{D,d} \left( \lambda_1(h(L_1^*, L_2^*)) \in \{\sigma_i\}_{i=1}^{D} | \{\theta_i\}_{i=1}^{d-1} \right) = 0.
\]

Combining (104) with Fubini’s Theorem, we conclude (101).

3.6.6 Remark on the Size of \( \delta_0 \) and \( \kappa_0 \)

The above constants \( \delta_0 \) and \( \kappa_0 \) depend on other parameters of the underlying spherically uniform HLM model, in particular the underlying subspaces \( \{L_i^*\}_{i=1}^{K} \). We recall that \( \kappa_0 = \delta_0 = \frac{\zeta_1}{4p} \), where \( \zeta_1 = E_{\mu}(e_{\ell_p}(x, L_1^*)) - \min_{L \in G(D,d)} E_{\mu}(e_{\ell_p}(x, L)) \). Therefore, in order to bound \( \kappa_0 \) and \( \delta_0 \) from below, we bound \( \zeta_1 \) from below as follows:

\[
\zeta_1 \geq \begin{cases} 
\frac{p}{2} \| E_{\mu}(D_{L_1^*, x, p}) \|_{F}^2 & \text{if } p \geq 2, \\
(p - 1) \frac{1}{p - 2} \frac{1}{2^{\frac{p-2}{p-1}}} \| E_{\mu}(D_{L_1^*, x, p}) \|_{F}^{\frac{p}{p-1}} & \text{if } 1 < p < 2.
\end{cases}
\]

We include the proof of (105) in Sect. A.8.

We derive (10) from (105) as follows. We recall that (10) applies to the case where \( K = 2, \alpha_0 = 0, \dim(L_1^*) = \dim(L_2^*) = 1, D = 2, \) and where \( \mu_1 \) and \( \mu_2 \) are uniform
distributions on line segments centered on the origin and of length 2 within $L_1^*$ and $L_2^*$. If $\theta$ is the angle between $L_1^*$ and $L_2^*$, then

$$\|E_\mu(D_{L_1^*,x,p})\| = \alpha_2 \cos(\theta) \sin(\theta)^{p-1}/(p + 1).$$  \hspace{1cm} (106)

The lower bound for both $\kappa_0$ and $\delta_0$ in (10) thus follows from (105), (106), and the fact that $\kappa_0 = \delta_0 = \zeta_1/4p$.

3.6.7 Implication of Proof: A Counterexample for Exact Asymptotic Recovery

Theorem 1.2 established near recovery of $L_1^*$ for a spherically uniform HLM measure $\mu_\epsilon$ when $\epsilon > 0$ and $0 < p \leq 1$. It is sometimes more desirable to have exact asymptotic $l_p$ recovery of $L_1^*$. It means that if $X = \{x_1, x_2, \ldots, x_N\}$ is an i.i.d. sample from $\mu_\epsilon$ and $L_1^*(N)$ is the minimizer of $e_1^p(X, L_1^*)$, then $L_1^*(N)$ converges to $L_1^*$ w.p. 1 as $N$ approaches infinity. However, this is generally not true for any $p > 0$ when $K > 1$ and $\epsilon > 0$. Indeed, we provide here a simple counterexample, whose verification follows the proof of Theorem 1.3.

We assume a measure $	ilde{\mu} = \alpha_1 \tilde{\mu}_1 + \sum_{i=2}^{K} \alpha_i \mu_i$, where $\{\mu_i\}_{i=2}^{K}$ are the uniform measures on $S^{D-1} \cap L_i^*$ and $\tilde{\mu}_1$ is the uniform measure on the strip $\{x \in S^{D-1} : \text{dist}(x, L_1^*) \leq \epsilon\}$. The symmetry of this strip w.r.t. $L_1^*$ implies that

$$\int P_{L_1^*}(x) P_{L_1^*}(x)^T \text{dist}(x, L_1^*)^{p-2} d\tilde{\mu}_1(x) = 0.$$  \hspace{1cm} (107)

Besides, it follows from Proposition 2.2 that a necessary condition for $L_1^*$ to be a local $l_p$ subspace in expectation is

$$\int P_{L_1^*}(x) P_{L_1^*}(x)^T \text{dist}(x, L_1^*)^{p-2} d\mu(x) = 0.$$  \hspace{1cm} (108)

Combining (107) and (108), we conclude that

$$\sum_{i=0,i=2}^{K} \int P_{L_1^*}(x) P_{L_1^*}(x)^T \text{dist}(x, L_1^*)^{p-2} d\mu_i(x) = 0.$$  \hspace{1cm} (109)

However, the proof of (97) implies that the measure $\gamma_{D,d}^K$ of (109) w.r.t. $\{L_i^*\}_{i=1}^{K}$ is zero. That is, a.e. $L_1^*$ (w.r.t. $\gamma_{D,d}^K = 1$) is not the global $l_p$ subspace in expectation. Consequently, a.e. $L_1^*$ is not the asymptotic global $l_p$ subspace (since exact asymptotic recovery is stronger than recovery in expectation).

4 Discussion

We studied the effectiveness of $l_p$ minimization for recovering and nearly recovering the most significant subspace within outliers w.o.p. Our setting assumed i.i.d. sampling...
from a spherically uniform HLM measure (and sometimes weakly spherically uniform HLM measure) with noise level $\epsilon \geq 0$. A restricted setting is necessary, and indeed we described some typical cases where the global $l_p$ subspace is different than the most significant subspace for all $0 < p < \infty$. In our particular study of significantly large fraction of outliers, we need the rather strong restriction of spherically symmetric outliers, which is not necessary when limiting this fraction (see, e.g., [22]).

Our analysis provided some guarantees for the robustness to spherically uniform outliers (or spherically symmetric outliers) of the single subspace recovery advocated in [9]. The recovery established here is for the theoretical minimizer of the energy and not for any algorithmic output. Both [46] and [23] followed some basic ideas of this paper in their analysis of a convex relaxation of (1) when $p = 1$, while incorporating many other ideas. However, the theoretical guarantees of the latter works require a bound on the fraction of outliers, and it is unclear if their algorithms can always recover the most significant subspace in our setting when $K > 1$.

We further discuss possible and impossible extensions of this theory, some other implications, and open problems.

4.1 Beyond Spherically Uniform Distributions

We can easily replace spherically uniform distributions with sub-Gaussian spherically symmetric distributions. For this purpose, we may apply the Hoeffding-type inequality for sub-Gaussian measures of Proposition 5.10 in [39]. Alternatively, if the data is projected onto the unit sphere, then spherically symmetric distributions (and even some more general distributions) are mapped into spherically uniform distributions.

We may even relax the spherical symmetry of $\{\mu_i\}_{i=1}^K$ within $\{L_i^*\}_{i=1}^K$ and require instead approximate spherical symmetry within $\{L_i^*\}_{i=1}^K$. That is, we require for $i = 2, \ldots, K$ that there exist $\{\tilde{\mu}_i\}_{i=1}^K$ spherically symmetric distributions within $L_i^*$ such that the derivatives $f_i := d\mu_i / d\tilde{\mu}_i$, $i = 1, \ldots, K$ are bounded away from 0 and $\infty$. In this case, (2) is replaced with

$$\alpha_1 > \sum_{i=2}^K \frac{\sup(f_i)}{\inf(f_1)} \alpha_i.$$  \hfill (110)

On the other hand, a symmetry-type property of $\mu_0$ is crucial for the proof of Theorems 1.1 and 1.2, unless one may tolerate a restricted fraction of outliers [22].

In Theorem 2.2, it is enough to assume that $\mu_0$ is symmetric with respect to $L_1^*$. It is even possible to assume a slightly weaker assumption: $E_{\mu_0}(D_{L_1^*x,p}) = 0$, where $D_{L_1^*x,p}$ is defined in (95).

4.2 Affine Subspaces

We restrict the theory of this paper to linear subspaces, since affine subspaces do not fit within the framework of spherically uniform (or spherically symmetric) measures. The common strategy of using homogenous coordinates, which transform $d$-dimensional
affine subspaces in $\mathbb{R}^D$ to $(d + 1)$-dimensional linear subspaces in $\mathbb{R}^{D+1}$, is not useful to us since it distorts the structure of both noise and outliers. On the other hand, the theory of [22] can be generalized to affine subspaces (see Sect. 5.6 of [22]).

4.3 $p = 1$ Versus $0 < p < 1$

Our main theorems do not distinguish between $p = 1$ and $0 < p < 1$. However, Proposition 2.1 shows that many subspaces can be local $l_p$ subspaces when $p < 1$ (in particular, $d$-subspaces spanned by subsets of outliers). Such wealth of local minima clearly does not occur when $p = 1$. An open problem is to estimate the number and depth of local minima when $p = 1$ for spherically uniform HLM measures.

4.4 The Nonconvexity of Our Strategy

Our setting is nonconvex, and we are not aware of theoretically guaranteed strategies to compute the global minimizer. Ding et al. [9], Zhang et al. [47] and Lerman and Maunu [21] suggested heuristic methods without theoretical guarantees to approximate a minimizer to our problem. It will be interesting to develop even partial theoretical guarantees for any of these strategies or a newer one. It will also be interesting to understand whether it is advantageous to try to compute the minimizer of (1) when $p = 1$ instead of computing the minimizer of a convex relaxation of this energy.

In Sect. 1.2, we discussed the result of Hardt and Moitra [17], which implies that if the small set expansion problem has no efficient algorithm (which is unknown), then under some circumstances (different than the ones here) subspace recovery with sufficiently high percentage of outliers cannot be done by an efficient algorithm. It is interesting to know if it is true that any procedure that can exactly recover the underlying subspace in our setting with arbitrarily large percentage of outliers must be inefficient. If it is true, we are then curious about the upper bound on the fraction of outliers in our setting. After all, [46] and [23] indicated higher fraction of outliers than Hardt and Moitra [17] for our setting with $K = 1$.

Acknowledgments  This work was supported by NSF grants DMS-09-15064 and DMS-09-56072. It is inspired by our collaboration with Arthur Szlam on efficient and fast algorithms for hybrid linear modeling, which apply geometric $l_1$ minimization. We thank the anonymous reviewer for many insightful comments and suggestions that significantly improved the presentation of this work, John Wright for referring us to [37,38] as well as for relevant questions which we address in Sect. 4 and Vic Reiner, Stanislaw Szarek, and J. Tyler Whitehouse for commenting on an earlier version of this manuscript. Thanks to the Institute for Mathematics and its Applications (IMA) for holding a workshop on multi-manifold modeling that GL co-organized and TZ participated in.

Appendix A: Supplementary Details

A.1. The auxiliary function $\psi$

We define the function $\psi$ and bound it from above. We later use this function and its upper bound to estimate $\gamma_3$ (in Sect. A.2).
We assume that $L$ is a $d$-subspace of $\mathbb{R}^D$, where $0 \leq d \leq D$, and that $v$ is the uniform measure on $L \cap S^{D-1}$. We define

$$
\psi_v(t) = v(x \in L : |x^Tv| < t),
$$

(111)

where $v$ is an arbitrarily fixed vector in $L \cap S^{D-1}$ (since $v$ is uniform on $S^{D-1} \cap L$, $\psi_v$ is independent of $v$).

We establish the following upper bound on $\psi_v$:

$$
\psi_v(t) < \sqrt{\frac{\pi d^2}{2}} t.
$$

(112)

Let us denote the surface area measure on $S^{d-1}$ by $\text{Area}_{d-1}$. Using this notation, we conclude (112) as follows:

$$
\psi_{\mu_1}(t) = \text{Area}_{d-1} \left\{ x \in S^{d-1} : |x_1| < t \right\} / \text{Area}_{d-1} \left\{ S^{d-1} \right\} = \frac{\int_{\frac{\pi}{2}}^{\frac{\pi}{2} \cdot \cos^{-1}(t)} \sin^{d-2}(\theta) d\theta}{\int_{0}^{\frac{\pi}{2}} \sin^{d-2}(\theta) d\theta} \cdot \frac{\int_{0}^{\frac{\pi}{2}} \sin^{d-2}(\theta) d\theta}{\int_{\frac{\pi}{2}}^{\frac{\pi}{2} \cdot \cos^{-1}(t)} \sin^{d-2}(\theta) d\theta} = \frac{\sin^{-1}(t)}{\sqrt{\pi d}} \leq \frac{\pi d}{\sqrt{2}}.
$$

We remark that the second equality follows from the well-known formula for the surface area of the spherical cap of “half-angle” $\beta$, $\text{Cap}(\beta) \subset S^{D-1}$:

$$
\text{Area}_{d-1}(\text{Cap}(\beta)) = C(d) \cdot \int_{0}^{\frac{\beta}{2}} \sin^{d-2}(\theta) d\theta,
$$

where $d$ does not use the value of $C(d)$ since it cancels in both the numerator and denominator; the fourth (and last) equality follows from a basic trigonometric identity (for the numerator) and the following well-known integration formula (for the denominator): $\int_{0}^{\frac{\pi}{2}} \sin^{d-2}(\theta) d\theta = B((d-1)/2, 1/2)/2 = \sqrt{\pi} \Gamma((d-1)/2)/(2 \Gamma(d/2))$; and the last inequality is obtained by applying the inequality $\sin^{-1}(t) \leq \pi t/2$ for $0 \leq t \leq \pi/2$ (for the numerator) and the following immediate consequence of Gautschi–Kershaw’s inequality [30] for the gamma function:

$$
\Gamma\left(\frac{d}{2}\right) / \Gamma\left(\frac{d-1}{2}\right) \leq \sqrt{d/2}
$$

(for the denominator), which is obtained by substituting $s = 0.5$, $x = d/2 - 1$ in (1) of [30] and using a looser upper bound.

A.2 Asymptotic Dependence of $\gamma_3$ on $D$ and $d$

We upper bound the constant $\gamma_3$, which is determined by (69), by applying the function $\psi$ and its upper bound in (112). We note that for any $v \in S^{D-1}$ orthogonal to $L_1^*$,

$$
\{ x \in S^{D-1} : 0 < \text{dist}(x, L_1^*) < 2\gamma_3 \} \subset \{ x \in S^{D-1} : 0 < |x^Tv| < 2\gamma_3 \}.
$$

(113)

Therefore, we arbitrarily fix $v \in S^{D-1} \cap L_1^\perp$ (we will adapt this choice throughout the construction) and estimate a constant $\gamma_3$ that will satisfy the equation

$$
\psi_v(t) < \sqrt{\frac{\pi d^2}{2}} t.
$$
Indeed, it follows from (113) and the fact that dist(x, L_1) = 0 if and only if x ∈ supp(μ_1) that γ_3 satisfies (114), then it also satisfies (69).

Since α_0 + ∑_{i=2}^{K} α_i < 1, we only need to find γ_3 such that

\[
\max_{i=0,2,3,...,K} \mu_i(\{x \in \mathbb{S}^{D-1} : |x^T v| < 2\gamma_3\}) \leq \gamma_1/24. \tag{115}
\]

Applying (112) (with L = \mathbb{R}^D, where L is the subspace defining v), we obtain that

\[
\mu_0(\{x \in \mathbb{S}^{D-1} : |x^T v| < 2\gamma_3\}) = \psi_{\mu_0}(2\gamma_3) < \gamma_3 \sqrt{2\pi D}. \tag{116}
\]

Fixing 2 ≤ i ≤ K and applying again (112) (with L = L_i^*), we obtain that

\[
\mu_i(\{x \in \mathbb{S}^{D-1} : |x^T y| = 2\gamma_3\}) \leq \mu_i(\{x \in L_i^* : |x^T (P_{L_i^*} y)| < 2\gamma_3\})
\]

\[= \mu_i(\{x \in L_i^* : |x^T (P_{L_i^*} y)/\|P_{L_i^*} y\| < 2\gamma_3/\|P_{L_i^*} y\|\})
\]

\[= \psi_{\mu_i}(2\gamma_3/\|P_{L_i^*} y\|) < \gamma_3 \sqrt{2\pi d}/\|P_{L_i^*} y\|. \tag{117}
\]

Since the subspaces \{L_i^*\}_{i=1}^{K} are distinct, we may adapt v such that \|P_{L_i^*} v\| ≠ 0 for all 2 ≤ i ≤ K (we discuss the optimal choice of v below). Combining (116) and (117) and using the fact that (115) implies (114) and thus (69), we conclude that γ_3 = γ_1 \min_{i=2}^{K} \|P_{L_i^*} v\|/(24\sqrt{2\pi D}) will satisfy (115).

We can choose v to maximize the \min_{i=2}^{K} \|P_{L_i^*} v\|, and therefore

\[
\gamma_3 = \gamma_1 \max_{v \in L_1^*, \|v\|=1} \min_{i=2,...,K} \|P_{L_i^*} v\|/(24\sqrt{2\pi D}).
\]

We remark that \max_{v \in L_1^*, \|v\|=1} \min_{i=2}^{K} \|P_{L_i^*} v\| is similar to \min_{i=2}^{K} \text{dist}_G(L_1^*, L_i^*), since both of them measure the difference between L_1^* and \{L_i^*\}_{i=2}^{K}, and, in particular, their value is 0 only when L_1^* = L_i^* for some i ≥ 2.

A.3 Proof of Lemma 3.1

We assume first that p = 2. We denote the principal angles between L_1 and \tilde{L}_1 by \{θ_i\}_{i=1}^{d} and the principle vectors of L_1 and \tilde{L}_1 by \{v_i\}_{i=1}^{d} and \{\tilde{v}_i\}_{i=1}^{d}, respectively. We express every point x in L_1 by x = (x_1, x_2, ..., x_d) = (v_1^T x, v_2^T x, ..., v_d^T x). We note that

\[
\text{dist}(x, \tilde{L}_1)^2 = \sum_{i=1}^{d} x_i^2 \sin^2 \theta_i \geq \frac{4}{\pi^2} \sum_{i=1}^{d} x_i^2 \theta_i^2. \tag{118}
\]
Combining (118) with the observation that $E_{\mu_1}(x_i^2) = 1/d$ for all $1 \leq i \leq d$, we conclude Lemma 3.1 in this case as follows:

$$E_{\mu_1}(e_{l_2}(x, \hat{L}_1)) = E_{\mu_1}(\text{dist}(x, \hat{L}_1))^2 \geq E_{\mu_1}(\frac{4}{\pi^2} \sum_{i=1}^{d} x_i^2 \theta_i^2)$$

$$= \frac{4}{\pi^2} d \sum_{i=1}^{d} \theta_i^2 = \frac{4}{\pi^2} \cdot \text{dist}_G(L_1, \hat{L}_1)^2. \quad (119)$$

Next, we assume that $p > 2$. Applying (119) and Jensen's inequality with the convex function $\phi(x) = x^{p/2}$, we conclude Lemma 3.1 in this case as follows:

$$E_{\mu_1}(e_{l_p}(x, \hat{L}_1)) \geq \left(E_{\mu_1}(e_{l_2}(x, \hat{L}_1))\right)^{\frac{p}{2}} \geq \pi^{-p} \cdot 2^p \cdot d^{-\frac{p}{2}} \cdot \text{dist}_G(L_1, \hat{L}_1)^p.$$

Finally, we assume that $0 < p < 2$. Using the above parametrization $x = (x_1, x_2, \ldots, x_d)$ for points in $L_1^* \cap S^{D-1}$, we view the restriction of $\mu_1$ onto $L_1^*$ (expressed in these coordinates) as the uniform measure onto $S^{d-1}$. It follows from (118) that

$$E_{\mu_1}(e_{l_p}(x, \hat{L}_1)) \geq E_{\mu_1}\left(\frac{4}{\pi^2} \sum_{i=1}^{d} x_i^2 \theta_i^2\right)^{p/2}. \quad (120)$$

The main argument of the proof, which we delay to Sect. A.3.1, is to verify (via Karamata’s inequality) that when $\text{dist}_G(L_1, \hat{L}_1)$ (equivalently, $\sum_{i=1}^{d} \theta_i^2$) is fixed, then

$$E_{\mu_1}\left(\sum_{i=1}^{d} x_i^2 \theta_i^2\right)^{p/2} \geq E_{\mu_1}(x_1^p) \cdot \left(\sum_{i=1}^{d} \theta_i^2\right)^{p/2} = E_{\mu_1}(x_1^p) \cdot \text{dist}_G(L_1, \hat{L}_1)^p. \quad (121)$$

We estimate $E_{\mu_1}(x_1^p)$ as follows:

$$E_{\mu_1}(x_1^p) = \frac{\int (\sin \theta)^{d-2} (\cos \theta)^p \, d\theta}{\int (\sin \theta)^{d-2} \, d\theta} = \frac{B\left(\frac{d-1}{2}, \frac{p+1}{2}\right)}{B\left(\frac{d-1}{2}, \frac{1}{2}\right)} = \frac{\Gamma\left(\frac{p+1}{2}\right) \Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d-p+1}{2}\right) \Gamma\left(\frac{d}{2}\right)} > 0.88 \cdot \left(\frac{d + p - 0.6}{2}\right)^{-p/2}. \quad (122)$$

The last inequality uses the following equalities and inequalities: $\Gamma(1/2) = \sqrt{\pi}$; $\Gamma\left(\frac{p+1}{2}\right) \geq 0.88$, which follows from the well-known estimate $\min_{x \geq 0} \Gamma(x) \approx 0.885603$ (see, e.g., [8]); and the inequality $\Gamma\left(\frac{d+p}{2}\right) / \Gamma\left(\frac{d}{2}\right) < \left(\frac{d+p}{2}\right)^{p/2}$, which follows from Gautschi–Kershaw’s inequality [30] (indeed, apply (1) of [30] with $x = (d + p - 2)/2$ and $s = 1 - p/2$, while using a looser upper bound, and then invert the inequality while taking the power of -1 of both its LHS and RHS).
Therefore, the case $0 < p < 2$ is concluded by combining (120), (121) (which is proved in the following subsection), and (122).

A.3.1. Proof of (121)

We will prove a more general statement, which requires the following notation: For $1 \leq j \leq d$ and $1 \leq i \leq d$,

$$\theta_{i, j} = \begin{cases} \sqrt{\sum_{i=1}^{j} \theta_{i, j}^2} & \text{if } i = 1, \\ 0 & \text{if } 2 \leq i \leq j, \\ \theta_i & \text{if } j + 1 \leq i \leq d. \end{cases}$$

The more general statement is

$$E_{\mu_1} \left( \sum_{i=1}^{d} x_i^2 \theta_{i, j}^2 \right)^{p/2} \geq E_{\mu_1} \left( \sum_{i=1}^{d} x_i^2 \theta_{i, j+1}^2 \right)^{p/2} \quad \text{for } 1 \leq j \leq d - 1. \quad (123)$$

Clearly, successive application of (123) implies (121).

In order to prove (123), we introduce additional notation, formulate two sequences with the majorization property, and then apply Karamata’s inequality. For $1 \leq j \leq d - 1$, let

$$x_{i, j} = \begin{cases} x_i & \text{if } i \neq 1, j + 1, \\ x_{j+1} & \text{if } i = 1, \\ x_1 & \text{if } i = j + 1. \end{cases}$$

It follows from elementary algebraic manipulations that for any $1 \leq j \leq d$,

$$\sum_{i=1}^{d} x_i^2 \theta_{i, j+1}^2 + \sum_{i=1}^{d} x_i^2 \theta_{i, j+1}^2 = \sum_{i=1}^{d} x_i^2 \theta_{i, j+1}^2 + \sum_{i=1}^{d} x_i^2 \theta_{i, j+1}^2. \quad (124)$$

One can also verify that

$$\max \left( \sum_{i=1}^{d} x_i^2 \theta_{i, j+1}^2, \sum_{i=1}^{d} x_i^2 \theta_{i, j+1}^2 \right) \geq \max \left( \sum_{i=1}^{d} x_i^2 \theta_{i, j}^2, \sum_{i=1}^{d} x_i^2 \theta_{i, j}^2 \right). \quad (125)$$

This is done by showing (again by simple algebra) that each one of the terms in the argument of the maximum function in the LHS of (125) is controlled by one of the terms in the RHS of (125). Equations (124) and (125) imply that $(\sum_{i=1}^{d} x_i^2 \theta_{i, j+1}^2, \sum_{i=1}^{d} x_i^2 \theta_{i, j+1}^2)$ majorizes $(\sum_{i=1}^{d} x_i^2 \theta_{i, j}^2, \sum_{i=1}^{d} x_i^2 \theta_{i, j}^2)$. Combining this observation, the concavity of $f(x) = x^{p/2}$, and Karamata’s inequality, we conclude that
\[
\left( \sum_{i=1}^{d} x_{i,j}^2 \theta_{i,j}^2 \right)^{p/2} + \left( \sum_{i=1}^{d} x_{i,j}^2 \theta_{i,j}^2 \right)^{p/2} \geq \left( \sum_{i=1}^{d} x_{i,j}^2 \theta_{i,j}^2 \right)^{p/2} + \left( \sum_{i=1}^{d} x_{i,j}^2 \theta_{i,j}^2 \right)^{p/2}.
\]

(126)

Integrating (126) over \( \mu_1 \) and using the invariance of \( \mu_1 \) to permutations of \( x \) (in particular, invariance to replacing \( x_i \) with \( x_{i,j} \) for all \( 1 \leq i \leq d \)), we obtain (123) and consequently (121).

A.4. Proof of Lemma 3.2

We denote the principal angles between the \( d \)-subspaces \( L_1 \) and \( L_2 \) by \( \theta_1 \geq \theta_2 \geq \theta_3 \geq \cdots \geq \theta_d \). Arbitrarily choosing \( Q_1, Q_2 \in O(D, d) \), representing \( L_1, L_2 \), respectively, we note that

\[
|\text{dist}(x, L_1) - \text{dist}(x, L_2)| = |\|x - xQ_1Q_1^T\| - \|x - xQ_2Q_2^T\|| \\
\leq \|x - xQ_1Q_1^T - x + xQ_2Q_2^T\| \\
\leq \|x\| \left\| Q_1Q_1^T - Q_2Q_2^T \right\|_F \\
= \|x\| \sqrt{\sum_{i=1}^{d} \sin(\theta_i)^2} \leq \|x\| \sum_{i=1}^{d} \theta_i^2 = \|x\| \text{dist}_S(L_1, L_2).
\]

A.5. Local \( l_p \) subspace for \( p > 1 \) and \( K = 1 \)

**Proposition A.1** Assume that \( D > d + 1 \), \( L_1^* \in G(D, d) \), \( \mu_0 \) is a distribution on \( \mathbb{R}^D \) such that \( \mu_0((L_1^*)) \neq 0 \) for any affine subspace \( L \), where \( L \subset \mathbb{R}^D \), \( \mu_1 \) a distribution on \( L_1^* \) and \( \mu = \alpha_0 \mu_0 + \alpha_1 \mu_1 \), where \( \alpha_0, \alpha_1 \) are nonnegative numbers summing to 1. If \( X \) is a data set sampled identically and independently from \( \mu \) and \( p > 1 \), then the probability that \( L_1^* \) is a local \( l_p \) subspace of \( X \) is 0.

Let \( \{y_i\}_{i=1}^{N_0} \) denote the i.i.d. outliers sampled from \( \mu_0 \). We will prove that for any \( V \in \mathbb{R}^{d \times D-d} \),

\[
\mu_0 \left( y_1 \in \mathbb{R}^D : P_{L_1}(y_1) P_{L_1}^\perp(y_1)^T \operatorname{dist}(y_1, L_1^*)^{-p/2} = V \right) = 0. \quad (127)
\]

Proposition A.1 follows by substituting

\[
V = -\sum_{i=2}^{N_0} P_{L_1}(y_i) P_{L_1}^\perp(y_i)^T \operatorname{dist}(y_i, L_1^*)^{-p/2}
\]

in (127) and applying Proposition 2.2.

We may assume that \( y_1 \notin L_1^* \cup L_1^\perp \) since \( \mu_0((L_1^*)) = \mu_0((L_1^\perp)) = 0 \). We note that for any \( y_1 \notin L_1^* \cup L_1^\perp \), the rank of \( P_{L_1}(y_1) P_{L_1}^\perp(y_1)^T \) is 1. Therefore, (127) is obvious if \( \operatorname{rank}(V) \neq 1 \). Furthermore, if \( \ker(V) \ni L_1^* \), then (127) is also obvious since the kernel of \( P_{L_1}(y_1) P_{L_1}^\perp(y_1)^T \) contains \( L_1^* \).

Finally, we assume that \( \operatorname{rank}(V) = 1 \) and \( \ker(V) \ni L_1^* \) and denote \( v = \ker(V) \perp \). Applying the assumption that proper affine subspaces of \( \mathbb{R}^D \) have measure \( \mu_0 \) zero and
the assumption $D > d + 1$, we obtain that $\mu_0(\text{Sp}(L_1^*, v)) = 0$. We thus conclude (127) (and consequently Proposition A.1) as follows:

$$\mu_0 \left( y_1 \in \mathbb{R}^D : P_{L_1^*}(y_1) P_{L_1^*}^T(y_1) \text{dist}(y_1, L_1^*)^{p-2} = v \right)$$

$$\leq \mu_0 \left( y_1 \in \mathbb{R}^D : P_{L_1^*}(y_1) = cv \text{ for some } c \in \mathbb{R} \right)$$

$$= \mu_0 \left( y_1 \in \mathbb{R}^D : y_1 \in \text{Sp}(L_1^*, v) \right) = 0.$$

A.6. Proof of Lemma 3.3

We assume without loss of generality that $i = 1$ in (19). We thus need to prove that for all $\hat{L} \in G(D, d)$,

$$\mathbb{E}_{\mu_1}(\text{dist}(x_1, \hat{L})^p) + \mathbb{E}_{\mu_2}(\text{dist}(x_2, \hat{L})^p) \geq \mathbb{E}_{\mu_1}(\text{dist}(x_1, L_1)^p) + \mathbb{E}_{\mu_2}(\text{dist}(x_2, L_1)^p).$$

We denote the principal angles between $L_1$ and $L_2$ by $\{\theta_i\}_{i=1}^d$, the principle vectors of $L_1$ and $L_2$ by $\{v_i\}_{i=1}^d$ and $\{\hat{v}_i\}_{i=1}^d$, and the complementary orthogonal system for $L_2$ w.r.t. $L_1$ by $\{u_i\}_{i=1}^d$.

We notice that we can restrict the set of subspaces $\hat{L}$ satisfying (128). First of all, we only need to consider subspaces $\hat{L} \subseteq L_1 + L_2$.

Indeed, the LHS of (128) is the same if we replace $\hat{L}$ by $\hat{L} \cap (L_1 + L_2)$.

Second of all, we claim that it is sufficient to assume that

$$\text{Sp}(\hat{v}_i, v_i) \not\subseteq \hat{L} \text{ for all } 1 \leq i \leq k.$$

(130)

We first show this for $i = 1$. We suppose on the contrary to (130) that $\hat{v}_1, v_1 \in \hat{L}$. Since $\hat{L}$ is $d$-dimensional, there exists $2 \leq j \leq d$ (assume without loss of generality $j = 2$) such that it does not contain both $\hat{v}_j$ and $v_j$. For any pair of points $x = \sum_{i=3}^d a_i v_i \in L_1$ and $\hat{x} = \sum_{i=1}^d a_i \hat{v}_i \in L_2$,

$$\text{dist}(x, \hat{L}) = \sqrt{\sin(\theta_2)^2 a_2^2 + \tau_1^2} \quad \text{and} \quad \text{dist}(\hat{x}, \hat{L}) = \sqrt{\sin(\theta_1)^2 a_1^2 + \tau_2^2},$$

where

$$\tau_1 = \text{dist} \left( \sum_{i=3}^d a_i v_i, \hat{L} \right) \quad \text{and} \quad \tau_2 = \text{dist} \left( \sum_{i=3}^d a_i \hat{v}_i, \hat{L} \right).$$
Now, for \( \tilde{L} = \text{Sp}(\hat{L} \setminus \{v_1, \hat{v}_1\}, v_1, v_2) \), we obtain that

\[
\text{dist}(\hat{x}, \tilde{L}) = \sqrt{\sin(\theta_1)^2 a_1^2 + \sin(\theta_2)^2 a_2^2 + \tau^2} \quad \text{and} \quad \text{dist}(x, \tilde{L}) = \tau_1.
\]

Therefore,

\[
\text{dist}(x, \tilde{L})^p + \text{dist}(\hat{x}, \tilde{L})^p \leq \text{dist}(x, \hat{L})^p + \text{dist}(\hat{x}, \hat{L})^p,
\]

and by direct integration, we have that

\[
\mathbb{E}_{\mu_1}(\text{dist}(x_1, \tilde{L})^p) + \mathbb{E}_{\mu_2}(\text{dist}(x_2, \tilde{L})^p)
\]

\[\leq \mathbb{E}_{\mu_1}(\text{dist}(x_1, \hat{L})^p) + \mathbb{E}_{\mu_2}(\text{dist}(x_2, \hat{L})^p). \quad (131)\]

Since \( \tilde{L} \) satisfies (130) for \( i = 1 \) and satisfies (131), we conclude that proving (128) only for \( \hat{L} \) satisfying (130) with \( i = 1 \) implies it for all \( \hat{L} \in G(D, d) \). Similarly, we can assume that \( \hat{L} \) satisfies (130) for all \( 1 \leq i \leq k \), by verifying (131) for \( \tilde{L} = \text{Sp}(\hat{L} \setminus \{v_i, \hat{v}_i\}, v_i, v_j) \) for some \( 1 \leq j \neq i \leq k \) such that \( \text{Sp}(\hat{v}_j, v_j) \not\subseteq \hat{L} \).

It follows from (129) and (130) that \( \hat{L} \) can be represented as follows:

\[
\hat{L} = \text{Sp}(v_1^*, v_2^*, \ldots, v_d^*),
\]

where

\[
v_i^* = \cos \theta_i^* v_i + \sin \theta_i^* u_i.
\]

Thus, for any pair of points \( x = \sum_{i=1}^d a_i v_i \in L_1 \) and \( \hat{x} = \sum_{i=1}^d a_i \hat{v}_i \in L_2 \),

\[
\text{dist}(x, \hat{L}) = \sqrt{\sum_{i=1}^d \sin^2 \theta_i^* a_i^2},
\]

\[
\text{dist}(\hat{x}, \hat{L}) = \sqrt{\sum_{i=1}^d \sin^2 (\theta_i - \theta_i^*) a_i^2},
\]

\[
\text{dist}(x, L_1) = 0, \quad \text{and} \quad \text{dist}(\hat{x}, L_1) = \sqrt{\sum_{i=1}^d \sin^2 \theta_i a_i^2}. \quad (133)
\]

Applying (132), (133), the triangle inequality (for “sine vectors” in \( \mathbb{R}^d \)) and then the subadditivity of the sine function, we obtain that

\[
\text{dist}(x, \hat{L}) + \text{dist}(\hat{x}, \hat{L}) \geq \sqrt{\sum_{i=1}^d (\sin \theta_i^* + \sin (\theta_i - \theta_i^*))^2 a_i^2}
\]

\[
\geq \sqrt{\sum_{i=1}^d \sin^2 \theta_i a_i^2} = \text{dist}(\hat{x}, L_1) + \text{dist}(x, L_1).
\]
Since $p \leq 1$, this inequality clearly implies that
\[
\text{dist}(x, \hat{L})^p + \text{dist}(\hat{x}, \hat{L})^p \geq \text{dist}(\hat{x}, L_1)^p = \text{dist}(\hat{x}, L_1)^p + \text{dist}(x, L_1)^p.
\]  
(134)

We conclude (128) by appropriately integrating (134) and consequently prove the lemma.

A.7. Proof of (44)

We set $B = \sum_{i=1}^{N_1} P_{L_1^*}(x_i) P_{L_1^*}(x_i)^T$, and note that if $\max_{1 \leq j \leq d} \sigma_j (B - \delta_* I_d) < \eta$, then
\[
\frac{\|Bv - \delta_* v\|}{\|v\|} < \eta \text{ for all } v \in \mathbb{R}^d \setminus \{0\},
\]
and consequently
\[
\delta_* - \eta < \frac{\|Bv\|}{\|v\|} \text{ for all } v \in \mathbb{R}^d \setminus \{0\},
\]
that is, $\min_{1 \leq j \leq d} \sigma_j (B) > \delta_* - \eta$.

A.8. Proof of (105)

We first prove the following two lemmata.

**Lemma A.1** For $p > 1$ and any $x, y \in \mathbb{B}(0, 1)$,
\[
\|\|x\|^p - x\| - \|y\|^p - y\| \leq \begin{cases} 2^{3-p} \|x - y\|^{p-1} & \text{if } 1 < p \leq 2, \\ (p-1)\|x - y\| & \text{if } p > 2. \end{cases}
\]

**Proof** First we consider the case where either $\|x\| = 1$ or $\|y\| = 1$. Without loss of generality we assume that $\|x\| = 1$. When $p > 2$,
\[
\|\|x\|^p - x\| - \|y\|^p - y\| = \|x - y\|^{p-2} \frac{\|x - y\| + \|y - \|y\|^p - y\|}{\|x - y\|} \\
\leq \|x - y\| \frac{1 - \|y\|^{p-1}}{1 - \|y\|} \leq (p-1)\|x - y\|,
\]
where the second inequality follows from the identity $1 - \|y\| + \|y - \|y\|^p - y\| = 1 - \|y\|^{p-1}$, the inequality $\|x - y\| \geq 1 - \|y\|$, and the fact that the function $f(t) = (t+c)/t$ is nonincreasing for $c \geq 0$. 

Springer
On the other hand, when $1 < p \leq 2$,

$$
\|\|x\|^{p-2}x - \|y\|^{p-2}y\| = \|x - y\|^{p-2}y
\leq |x - y| + \|y - \|y\|^{p-2}y\| \leq 2\|x - y\|,
$$

(135)

where the last inequality of (135) follows from the inequality

$$
\|y - \|y\|^{p-2}y\| \leq \frac{\|y\|}{\|y\|} - \|y\| \leq \|x - y\|,
$$

(136)

which we explain as follows. Since $y$, $\|y\|^{p-2}y$, and $y/\|y\|$ lie on the same line through the origin, and since $\|y\| \leq \|\|y\|^{p-2}y\| \leq 1$, $\|y\|^{p-2}y$ is located between $y$ and $y/\|y\|$, and this clarifies the first inequality in (136). The second inequality in (136) follows from the following observation: $\|y/\|y\| - y\| = 1 - \|y\| = \|x + \|y\\| \leq \|x - y\|$.

The main idea of the proof for the general case is to arbitrarily fix $\|x - y\|$ and maximize $\|\|x\|^{p-2}x - \|y\|^{p-2}y\|$. We transform the problem into maximization over the two variables $r = \log(\|x\|/\|y\|)$ and $t = 2x^Ty/(\|x\|\|y\|)$ of the function

$$
h(r, t) := \frac{e^{(p-1)r} + e^{-(p-1)r} + t}{e^r + e^{-r} + t} = \left(\frac{\|\|x\|^{p-2}x - \|y\|^{p-2}y\|}{\|x - y\|^{p-1}}\right)^2,
$$

when $\|x - y\| > 0$ is fixed (if $\|x - y\| = 0$ then Lemma A.1 is trivial).

We first find the boundary of the domain of this function when $c_0 := \|x - y\|$ is fixed. We then maximize the function on the boundary and later find a local maximizer within the interior of this domain. The variable $t$ obtains values in $[-2, 2]$. For any fixed $t$, we find the values that $r$ may obtain. We note that $\|x\|^2 + \|y\|^2 - t\|x\|\|y\| = c_0^2$ and $e^{2r} + 1 - te^r = c_0^2/\|y\|^2$. Since $\|y\| \leq 1$, if $t$ is fixed and $r \leq 0$, then $r$ is in the domain $e^{2r} + 1 - te^r \geq c_0^2$, whose boundary is $e^{2r} + 1 - te^r = c_0^2$. That is, when $r \leq 0$, (i.e., $\|x\| \leq \|y\|$), then $\|y\| = 1$. Similarly, when $r \geq 0$, the boundary of the domain of $h(r, t)$ corresponds to the case $\|x\| = 1$.

Next, we verify (105) for points on the boundary of the domain of $h(r, t)$ (it is sufficient to verify it for maximizers on this boundary). For fixed $-2 < t < 2$, points on the boundary correspond to $\|x\| = 1$ or $\|y\| = 1$, and we have already verified (105) in this case. We also need to consider the boundary points $t = -2$ or $t = 2$, equivalently, $x/\|x\| = -y/\|y\|$ or $x/\|x\| = y/\|y\|$. We thus find the maximal values of $h(r, 2)$ and $h(r, -2)$ (when its denominator is fixed). The function $\sqrt{h(r, -2)}$ (i.e., with $x$ and $y$ satisfying $x/\|x\| = -y/\|y\|$) is equivalent to

$$
\frac{a^{p-1} + b^{p-1}}{(a + b)^{p-1}}, \text{ where } a = \|x\| \text{ and } b = \|y\|.
$$

Its maximum is obtained when $a = b$ if $1 < p \leq 2$ and when $a = 0$ or $b = 0$ if $p > 2$. The function $\sqrt{h(r, 2)}$ (i.e., with $x$ and $y$ satisfying $x/\|x\| = y/\|y\|$) is equivalent to

$$
\frac{a^{p-1} - b^{p-1}}{(a - b)^{p-1}}.
$$
Using the convexity/concavity of the power function $x^{p-1}$ for different values of $p$, we note that if $p \geq 2$, then its maximum is obtained when $a = 1$ or $b = 1$, and if $1 < p \leq 2$, then its maximum is obtained when $b = 0$. It is immediate to note that (105) is satisfied when $a = 0$ (i.e., $x = 0$) or $b = 0$ (i.e., $y = 0$). We have also verified above that it is satisfied when $a = b$ and $1 < p \leq 2$. Indeed, $\|x - y\| \leq \|x\| + \|y\| = 2\|x\|$, and thus $\|x - y\|^{p-2} \geq (2\|x\|)^{p-2}$, which implies that

\[
2^{3-p}\|x - y\|^{p-1} = 2^{3-p}\|x - y\|^{p-2}\|x - y\| \geq 2^{3-p}(2\|x\|)^{p-2}\|x - y\|
= 2\|x\|^{p-2}\|x - y\|
= 2\||x|^p - 2 - \|y\|^{p-2}\|y\| \geq \|x|^p - 2 - \|y\|^{p-2}\|y\|.
\]

We therefore verified (105) for points corresponding to the boundary of $h$.

Finally, we consider the interior of the domain of $h$. If $(r_0, t_0)$ is a local maximizer of $h(r, t)$, then

\[
0 = \frac{d}{dr} h(r, t) \bigg|_{(r, t)=(r_0, t_0)} = \frac{(e^{r_0} + e^{-r_0} + t_0) - (p - 1)(e^{(p-1)r_0} + e^{-(p-1)r_0} + t_0)}{(e^{r_0} + e^{-r_0} + t_0)^p}
\]

and $(e^{r_0} + e^{-r_0} + t_0) = (p - 1)(e^{(p-1)r_0} + e^{-(p-1)r_0} + t_0)$. Therefore,

\[
h(r_0, t_0) = \frac{p - 1}{(e^{r_0} + e^{-r_0} + t_0)^{p-2}}.
\]

Furthermore, its maximal value (when $t_0$ is fixed) is obtained when $r_0 = 0$ or $r_0 = \infty$ or $r_0 = -\infty$. Equivalently, it is obtained when $a = b$ or $a = 0$ or $b = 0$. To conclude the proof, we only need to verify that (105) is satisfied when $a = b$ and any $1 < p \leq 2$ (all the other cases were discussed above). In this case, we use the fact that $a \leq 1$ and $ap^{-2} \leq 1$ and consequently note that

\[
\||x|^p - 2 - \|y\|^{p-2}\|y\| = ap^{-2}\|x - y\| \leq \|x - y\| \leq (p - 1)\|x - y\|.
\]

\[
\Box
\]

**Lemma A.2** If $f, g : \mathbb{R} \to \mathbb{R}, \ g(0) = 0$, $g$ is increasing, and

\[
|f'(x_1) - f'(x_2)| \leq g(|x_1 - x_2|) \text{ for any } x_1, x_2 \in \mathbb{R}, \quad (137)
\]

then the following inequality is satisfied for all $x_0 \in \mathbb{R}$ and for $\hat{x} := \min_{x \in \mathbb{R}} f(x)$:

\[
f(x_0) - f(\hat{x}) \geq |f'(x_0)|g^{-1}(|f'(x_0)|) - \int_0^{\hat{x}} g(x) \, dx.
\]

\[\square\ Springer\]
Proof Without loss of generality we assume that \( f'(x_0) \geq 0 \). Applying this assumption, (137) and the definition of \( \hat{x} \), we conclude the lemma as follows:

\[
\begin{align*}
    f(x_0) - f(\hat{x}) & \geq f(x_0) - f(x_0 - g^{-1}(f'(x_0))) \\
    &= \int_{x_0-g^{-1}(f'(x_0))}^{x_0} f'(x) \, dx \\
    &\geq \int_{x_0-g^{-1}(f'(x_0))}^{x_0} (f'(x_0) - g(x_0 - x)) \, dx \\
    &= f'(x_0)g^{-1}(f'(x_0)) - \int_0^{g^{-1}(f'(x_0))} g(x) \, dx.
\end{align*}
\]

\[
\square
\]

To prove (105), we restrict \( E_\mu(e_{t_p}(x, L)) \) to a geodesic line \( L : [0, \infty) \to G(D, d) \) with \( L(0) = L^*_1 \). Then we use the following inequality to find the lower bound of \( \zeta_1 \):

\[
\zeta_1 = E_\mu(e_{t_p}(x, L^*_1)) - \min_{L \in G(D, d)} E_\mu(e_{t_p}(x, L)) \\
\geq E_\mu(e_{t_p}(x, L(0))) - \min_{t \geq 0} E_\mu(e_{t_p}(x, L(t))). \tag{138}
\]

The lower bound of the RHS of (138) will be obtained by applying Lemma A.2 to \( f(t) = E_\mu(e_{t_p}(x, L(t))) \) with a specific \( L(t) \).

We choose this \( L(t) \) such that \( \text{dist}_G(L(0), L(1)) = 1 \) and

\[
\left. \frac{d}{dt} E_\mu(e_{t_p}(x, L(t))) \right|_{t=0} = -p \| E_\mu(D_{L^*_1, x, p}) \|_F. \tag{139}
\]

To show that this is possible, we recall (see (34)) that

\[
\left. \frac{d}{dt} E_\mu(e_{t_p}(x, L(t))) \right|_{t=0} = -p \text{tr}(CE_\mu(D_{L^*_1, x, p})U^T), \tag{140}
\]

where \( \|C\|_F = 1 \) (since we use the distance defined in (3)). Let us denote the thin SVD of \( E_\mu(D_{L^*_1, x, p}) \) by \( V_0 \Sigma_0 U_0^T \). We choose the matrices \( V, U, \) and \( C \), which determine \( L(t) \), as follows: \( V = V_0^T \), \( U = U_0^T \), and \( C = \Sigma_0/\| \Sigma_0 \|_F \). This choice indeed implies (139) as a consequence of (140) and the following observation:

\[
p \text{tr}(CE_\mu(D_{L^*_1, x, p})U^T) = p \text{tr}(\Sigma_0^2)/\| \Sigma_0 \|_F = p \| \Sigma_0 \|_F = p \| E_\mu(D_{L^*_1, x, p}) \|_F.
\]

We proceed by finding \( g \) for \( f(t) = E_\mu(e_{t_p}(x, L(t))) \) so that (137) is satisfied. It follows from (140) that for \( t_2 > t_1 \geq 0 \),
\[
\left| f'(t_2) - f'(t_1) \right| \leq p E_{\mu} \left( C, V(D_{L(t_1)}, x, p - D_{L(t_2)}, x, p) U^T \right)_{F} \\
\leq p E_{\mu} \| D_{L(t_1)}, x, p - D_{L(t_2)}, x, p \|_F.
\] (141)

Combining the following observations:

\[
\| P_{L(t_1)}(x) - P_{L(t_2)}(x) \| \leq \| P_{L(t_1)} - P_{L(t_2)} \| \leq \text{dist}_{G}(L(t_1), L(t_2)) = t_2 - t_1, \\
\| P_{L(t_1)}(x) \| \text{dist}(x, L(t_1))^{(p-2)} \| \leq 1 \text{ and } \| P_{L(t_2)}(x) \| \leq 1,
\]

with the following consequence of Lemma A.1:

\[
\| P_{L(t_1)}(x) \| \text{dist}(x, L(t_1))^{(p-2)} - P_{L(t_2)}(x) \| \text{dist}(x, L(t_2))^{(p-2)} \| \\
\leq \begin{cases} 
2^{3-p} \| P_{L(t_1)}(x) - P_{L(t_2)}(x) \|^{p-1} & \text{if } 1 < p \leq 2, \\
(p - 1) \| P_{L(t_1)}(x) - P_{L(t_2)}(x) \| & \text{if } p \geq 2,
\end{cases}
\]

we obtain that

\[
\| D_{L(t_1)}, x, p - D_{L(t_2)}, x, p \|_F = \| P_{L(t_1)}(x) P_{L(t_1)}(x)^T \| \text{dist}(x, L(t_1))^{(p-2)} \\
- P_{L(t_2)}(x) P_{L(t_2)}(x)^T \| \text{dist}(x, L(t_2))^{(p-2)} \|_F \\
\leq \| P_{L(t_1)}(x) \| \text{dist}(x, L(t_1))^{(p-2)} \| \| P_{L(t_1)}(x) - P_{L(t_2)}(x) \| \\
+ \| P_{L(t_2)}(x) \| \| P_{L(t_1)}(x) \| \text{dist}(x, L(t_1))^{(p-2)} - P_{L(t_2)}(x) \| \text{dist}(x, L(t_2))^{(p-2)} \| \\
\leq \| P_{L(t_1)}(x) - P_{L(t_2)}(x) \| \\
+ \| P_{L(t_1)}(x) \| \text{dist}(x, L(t_1))^{(p-2)} - P_{L(t_2)}(x) \| \text{dist}(x, L(t_2))^{(p-2)} \| \\
\leq \begin{cases} 
(t_2 - t_1) + (p - 1) (t_2 - t_1), & \text{if } p \geq 2; \\
(t_2 - t_1) + 2^{3-p} (t_2 - t_1)^{p-1}, & \text{if } 1 < p < 2
\end{cases}
\] (142)

In view of (137), (141), (142), and our choice of \( f \), we define

\[
g(t) = \begin{cases} 
p t & \text{if } p \geq 2, \\
2^{4-p} \max((t_2 - t_1), t_2 - t_1) & \text{if } 1 < p < 2
\end{cases}
\] (143)

We note that its inverse function is

\[
g^{-1}(t) = \begin{cases} 
\frac{1}{p} t & \text{if } p \geq 2, \\
\min(2^{p-4} t, (2^{p-4} t)^{\frac{1}{p-1}}) & \text{if } 1 < p < 2.
\end{cases}
\] (144)
Applying Lemma A.2 with \( f \) and \( g \) as above and \( x_0 = 0 \), we prove (105) as follows. We write \( c_1 = \| E_\mu \left( \text{tr}(CVD_{L^*_1,x,p}U^T) \right) \|_F \). When \( p \geq 2 \), \( f'(x_0) = pc_1 \), and

\[
\zeta_1 \geq pc_1 \cdot \frac{pc_1}{p} - \frac{pc_1}{p} \frac{p}{x} dx = \frac{pc_1^2}{p} - \frac{pc_1^2}{2p} = \frac{p^2c_1^2}{2p} = \frac{p}{2} \| E_\mu \left( \text{tr}(CVD_{L^*_1,x,p}U^T) \right) \|_F^2.
\]

When \( 1 < p < 2 \), applying

\[
\text{tr}(CVD_{L^*_1,x,p}U^T) \leq \| C \|_F \| VD_{L^*_1,x,p}U^T \|_F = \| C \|_F \| D_{L^*_1,x,p} \|_F \leq \| C \|_F \| P_{L^*_1}(x) \| \| P_{L^*_1}^{-1}(x)^T \| \text{dist}(x, L(t)) \| (p-2) \| \leq 1,
\]

we conclude that \( c_1 \leq 1 \) and \( 2^{p-4} pc_1 \leq 2^{p-3} c_1 < 1 \). Therefore, \( g^{-1}(t) = (2^{p-4} t)^{\frac{1}{p-1}} = 2^{\frac{p-4}{p-1}} t^{\frac{1}{p-1}} \) for \( 0 \leq t \leq pc_1 \), and

\[
\frac{1}{p-1} 2^{\frac{p-4}{p-1}} (pc_1)^{\frac{1}{p-1}} - \frac{1}{2^{p-1}} 2^{\frac{p-4}{p-1}} p^{\frac{p}{p-1}} dx = (p-1)p^{\frac{1}{p-1}} 2^{\frac{p-4}{p-1}} \| E_\mu \left( \text{tr}(CVD_{L^*_1,x,p}U^T) \right) \|_F^2.
\]

References

1. Arias-Castro, E., Donoho, D.L., Huo, X., Tovey, C.A.: Connect the dots: how many random points can a regular curve pass through? Adv. Appl. Probab. 37(3), 571–603 (2005)
2. Arias-Castro, E., Chen, G., Lerman, G.: Spectral clustering based on local linear approximations. Electron. J. Stat. 5, 1537–1587 (2011)
3. Bargiela, A., Hartley, J.K.: Orthogonal linear regression algorithm based on augmented matrix formulation. Comput. Oper. Res. 20, 829–836 (1993)
4. Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? J. ACM 58(3), 11 (2011)
5. Candès, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory 52(2), 489–509 (2006)
6. Candès, E.J., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. 59(8), 1207–1223 (2006)
7. David, G., Semmes, S.: Singular integrals and rectifiable sets in \( \mathbb{R}^n \): au-delà des graphes Lipschitziens. Astérisque 193, 1–145 (1991)
8. Deming, W.E., Colcord, C.G.: The minimum in the gamma function. Nature 135(3422), 917 (1935)
9. Ding, C., Zhou, D., He, X., Zha, H.: 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, pp. 281–288. ACM, New York (2006)
10. Dodge, Y.: An introduction to $L_1$-norm based statistical data analysis. Comput. Stat. Data Anal. 5(4), 239–253 (1987)
11. Donoho, D.L.: For most large underdetermined systems of equations, the minimal $l_1$-norm near-solution approximates the sparsest near-solution. Commun. Pure Appl. Math. 59(7), 907–934 (2006)
12. Donoho, D.L.: For most large underdetermined systems of linear equations the minimal $l_1$-norm solution is also the sparsest solution. Commun. Pure Appl. Math. 59(6), 797–829 (2006)
13. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. 20(2), 303–353 (electronic) (1999)
14. Elhamifar, E., Vidal, R.: Sparse subspace clustering: algorithm, theory, and applications. IEEE Trans. Pattern Anal. Mach. Intell. 35(11), 2765–2781 (2013)
15. Fischler, M., Bolles, R.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24(6), 381–395 (1981)
16. Golub, G., Loan, C.V.: Matrix Computations. John Hopkins University Press, Baltimore (1996)
17. Hardt, M., Moitra, A.: Can we reconcile robustness and efficiency in unsupervised learning? In: Proceedings of the Twenty-Sixth Annual Conference on Learning Theory (COLT 2013) (2013)
18. Harter, H.L.: The method of least squares and some alternatives: Part II. Int. Stat. Rev. 42(3), 235–264+282 (1974)
19. Harter, H.L.: The method of least squares and some alternatives. I. Int. Stat. Rev. 42, 147–174 (1974)
20. Huber, P.J., Ronchetti, E.: Robust Statistics. Wiley Series in Probability and Mathematical Statistics. Probability and Mathematical Statistics. Wiley, New York (2009)
21. Lerman, G., Maunu, T.: Fast algorithm for robust subspace recovery (2014). arXiv:1406.6145
22. Lerman, G., Zhang, T.: Robust recovery of multiple subspaces by geometric $l_p$ minimization. Ann. Stat. 39(5), 2686–2715 (2011)
23. Lerman, G., McCoy, M., Tropp, J.A., Zhang, T.: Robust computation of linear models, or how to find a needle in a haystack. Found. Math. Comput (2014) (to appear)
24. Locantore, N., Marron, J., Simpson, D., Tripoli, N., Zhang, J., Cohen, K., Boente, G., Fraiman, R., Brumback, B., Croux, C., Fan, J., Kneip, A., Marden, J., Robust, D.P.: Principal component analysis for functional data. TEST 8(1), 1–73 (1999)
25. Maronna, R.A., Martin, R.D., Yohai, V.J.: Robust Statistics: Theory and Methods. Wiley Series in Probability and Statistics. Wiley, Chichester (2006)
26. Mattila, P.: Geometry of Sets and Measures in Euclidean Spaces. Cambridge University Press, Cambridge (1995)
27. McCoy, M., Tropp, J.: Two proposals for robust PCA using semidefinite programming. Electron. J. Stat. 5, 1123–1160 (2011)
28. Nyquist, H.: Least orthogonal absolute deviations. Comput. Stat. Data Anal. 6(4), 361–367 (1988)
29. Osborne, M.R., Watson, G.A.: An analysis of the total approximation problem in separable norms, and an algorithm for the total $l_1$ problem. SIAM J. Sci. Stat. Comput. 6(2), 410–424 (1985)
30. Qi, F., Guo, B.-N., Chen, C.-P.: The best bounds in Gautschi-Kershaw inequalities. Math. Inequal. Appl. 9(3), 427–436 (2006)
31. Rousseeuw, P.J., Leroy, A.M.: Robust Regression and Outlier Detection. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York (1987)
32. Soltanolkotabi, M., Candés, E.J.: A geometric analysis of subspace clustering with outliers. Ann. Stat. 40(4), 2195–2238 (2012)
33. Soltanolkotabi, M., Elhamifar, E., Candés, E.J.: Robust subspace clustering. CoRR (2013). arXiv:1301.2603
34. Späth, H., Watson, G.A.: On orthogonal linear approximation. Numer. Math. 51, 531–543 (1987)
35. Szarek, S.J.: The finite-dimensional basis problem with an appendix on nets of Grassmann manifolds. Acta Math. 151(3–4), 153–179 (1983)
36. Szarek, S.J.: Metric entropy of homogeneous spaces. In: Quantum Probability (Gdansk, 1997), vol. 43 of Banach Center Publication, pp. 395–410. Polish Academy Sciences, Warsaw (1998)
37. Torr, P.H.S., Zisserman, A.: Robust computation and parametrization of multiple view relations. In: ICCV ’98: Proceedings of the Sixth International Conference on Computer Vision, p. 727. IEEE Computer Society, Washington (1998)
38. Torr, P.H.S., Zisserman, A.: MLESAC: a new robust estimator with application to estimating image geometry. Comput. Vis. Image Understand. 78(1), 138–156 (2000)
39. Vershynin, R.: Introduction to the non-asymptotic analysis of random matrices. In: Eldar, Y.C., Kutyniok, G. (eds.) Compressed Sensing: Theory and Applications. Cambridge University Press (to appear)
40. Watson, G.A.: Some Problems in Orthogonal Distance and Non-Orthogonal Distance Regression. Defense Technical Information Center (2001)
41. Watson, G.A.: On the gauss-newton method for \(l_1\) orthogonal distance regression. IMA J. Numer. Anal. 22(3), 345–357 (2002)
42. Wong, Y.-C.: Differential geometry of Grassmann manifolds. Proc. Natl Acad. Sci. USA 57, 589–594 (1967)
43. Xu, H., Caramanis, C., Sanghavi, S.: Robust PCA via outlier pursuit. In: NIPS, pp. 2496–2504, 2010
44. Xu, H., Caramanis, C., Sanghavi, S.: Robust PCA via outlier pursuit. IEEE Trans. Inf. Theory 58(5), 3047–3064 (2012)
45. Yan, J., Pollefeys, M.: A general framework for motion segmentation: independent, articulated, rigid, non-rigid, degenerate and nondegenerate. ECCV 4, 94–106 (2006)
46. Zhang, T., Lerman, G.: A novel m-estimator for robust PCA. J. Mach. Learn. Res. 15, 749–808 (2014)
47. Zhang, T., Szlam, A., Lerman, G.: Median \(K\)-flats for hybrid linear modeling with many outliers. In: 2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops), pp. 234–241. Kyoto, Japan (2009)
48. Zhang, T., Szlam, A., Wang, Y., Lerman, G.: Randomized hybrid linear modeling by local best-fit flats. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1927–1934, June 2010
49. Zhang, T., Szlam, A., Wang, Y., Lerman, G.: Hybrid linear modeling via local best-fit flats. Int. J. Comput. Vis. 100, 217–240 (2012)