l_p-Recovery of the Most Significant Subspace among Multiple Subspaces with Outliers *

Gilad Lerman and Teng Zhang

Department of Mathematics, University of Minnesota
127 Vincent Hall, 206 Church Street SE, Minneapolis, MN 55455
e-mail: lerman@umn.edu, zhang620@umn.edu

Abstract: We assume data sampled from a mixture of d-dimensional linear subspaces with outliers distributed symmetrically around the origin. We study the recovery of the global l_0 subspace (i.e., with largest number of points) 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 non-convex for all $p > 0$ and thus requires different methods for its analysis. We show that if $0 < p \leq 1$, then the global l_0 subspace can be recovered by l_p minimization with overwhelming probability. Moreover, when adding homoscedastic noise around the underlying subspaces, the generalized l_0 subspace (with largest number of points "around it") can be nearly recovered by l_p minimization with an error proportional to the noise level. On the other hand, if $p > 1$ and there is more than one underlying subspace, then the global l_0 subspace cannot be recovered and the generalized one cannot even be nearly recovered. The results of this paper clarify the effect of using variants of l_p minimizations in RANSAC-type strategies for single subspace recovery.

AMS 2000 subject classifications: Primary 68Q32, 62G35, 60D05; secondary 62-07, 68T10.
Keywords and phrases: Best approximating subspace, l_p minimization as relaxation for l_0 minimization, robust statistics, sequential hybrid linear modeling, optimization on the Grassmannian, principal angles and vectors, geometric probability, high-dimensional data..

1. Introduction

Principal Component Analysis (PCA) is the most common tool in high-dimensional data analysis. It approximates a given data set by a low-dimensional affine subspace minimizing an l_2 sum of distances. Such minimization is not robust to outliers. Here we study the robustness to outliers of a generalized version of this minimization using an l_p sum for all $p > 0$ under particular assumptions.

This l_p optimization problem takes place over a data set $\mathcal{X} \in \mathbb{R}^D$ and minimizes among all d-dimensional subspaces, L, the quantity:

$$e_{l_p}(\mathcal{X}, \mathrm{L}) = \sum_{\mathbf{x} \in \mathcal{X}} \mathrm{dist}(\mathbf{x}, \mathrm{L})^p, \tag{1}$$

where $\mathrm{dist}(\mathbf{x}, \mathrm{L})$ denotes the Euclidean distance between a data point $\mathbf{x}$ and the subspace L. We call any of the global minimizers of (1) a *global l_p subspace*. We sometimes also refer to this optimization as *geometric l_p minimization*. Our study is motivated by the problem of sequential recovery of multiple subspaces buried in outliers,

*This work was supported by NSF grants DMS-09-15064 and DMS-09-56072

or in short, sequential Hybrid Linear Modeling (HLM). That is, recovering the most significant subspace among those subspaces, then removing the points along it (or in a strip around it) from the given data and repeating this procedure according to the given number of subspaces. It is common in HLM to assume only d-dimensional linear subspaces (as opposed to affine ones and with mixed dimensions), which we refer to as d-subspaces. Therefore our underlying model assumes multiple d-subspaces buried in outliers, while we investigate the recovery of a single d-subspace.

1.1. Background and Related Work

The l_1 norm has been widely used to form robust statistics. For example, the geometric median is the point in a data set minimizing the sum of distances from the rest of data points, i.e., the l_1-averaged distance. For points on the real axis, it coincides with the usual median. Its robustness is most commonly quantified by showing that it has a breakdown point of 0.5 (i.e., the estimator will obtain arbitrarily large values only when the proportion of large observations is at least a half) [20].

The l_1 norm has also been successfully applied to robust regression [17, 16, 23, 21]. Furthermore, Basis pursuit [5] uses l_1 minimization to search for the sparsest solutions (i.e., solutions minimizing the l_0 norm) of an undercomplete system of linear equations. This strategy was only recently fully justified [3, 10, 9, 4]. Candès et al. [2] proposed and analyzed the principal component pursuit algorithm for robust PCA, which minimizes a weighted combination of the nuclear norm and a different l_1 norm among all decompositions matching the available data. A simpler use of the l_1 norm between given data points and representative points in a lower dimensional model (though without using the nuclear norm term to infer this model) has appeared in several other works [1, 14, 19, 18].

Geometric l_p minimization (as in (1)) has been proposed by Guy David and Stephen Semmes [7] for $p \geq 1$ in a pure analytic setting (free of outliers in the context of "Ahlfors regular measures"). Ding et al. [8] used the geometric l_1 minimization as a robust alternative for principal component analysis, though lacking any mathematical support. Very Recently, Xu et al. [29] have suggested the combination of the norm used in the geometric l_1 minimization with the nuclear norm (similar to Candès et al. [2]) to obtain the outlier pursuit algorithm which is convex and robust to outliers and estimates the intrinsic dimension without prior knowledge. Nevertheless, it depends on a tuning parameter, which is used to weigh the two norms, and it also cannot use the true dimension (if known) or any other information on the underlying subspace (e.g., an initial guess).

A sequential HLM algorithm was suggested by Yang et al. [30] using the Random Sample Consensus (RANSAC) [13] heuristic to find a single subspace iteratively. This 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 ϵ around the d-subspace spanned by those d vectors (both ϵ 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 [26, 27] have suggested a RANSAC-type strategy which minimizes a 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.

1.2. Basic Conventions and Notation

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 C is a constant independent of N (we will also use "w.p." as a shorthand for "with probability").

We denote by $\mathrm{G}(D, d)$ the Grassmannian space, i.e., the set of all d-subspaces of $\mathbb{R}^D$ with a manifold structure. Following [22, Section 3.9], we denote by $\gamma_{D,d}$ the "uniform probability measure on $\mathrm{G}(D, d)$". We will measure distances between F and G in $\mathrm{G}(D, d)$ by the metric

$$\mathrm{dist}_{\mathrm{G}}(\mathrm{F}, \mathrm{G}) = \sqrt{\sum_{i=1}^{d} \theta_i^2}, \tag{2}$$

where $\{\theta_i\}_{i=1}^d$ are the principal angles between F and G (we explain the choice of this metric at the end of Section 3.2.1). We designate a ball in $\mathrm{G}(D, d)$ by $\mathrm{B}_{\mathrm{G}}(\mathrm{L}, r)$ as opposed to a Euclidean ball in $\mathbb{R}^D$, $\mathrm{B}(\mathbf{x}, r)$.

1.3. Precise Formulation of the Problem

We assume an underlying data set $\mathcal{X} \subseteq \mathbb{R}^D$ of N points identically and independently sampled from the following kind of a mixture measure representing a spherically symmetric setting of HLM with outliers:

Definition 1.1. *We say that a probability measure μ on $\mathbb{R}^D$ is a* spherically symmetric HLM measure *(equivalently, spherically symmetric HLM measure with no noise or with noise level $\epsilon = 0$) if $\mu = \sum_{i=0}^K \alpha_i \mu_i$, where $\alpha_0 \geq 0$, $\alpha_i > 0$, $i = 1, \ldots, K$, and $\sum_{i=0}^K \alpha_i = 1$, μ_0 is a probability measure spherically symmetric around the origin with bounded support (it represents outliers) and $\{\mu_i\}_{i=1}^K$ are probability measures supported within distinct d-subspaces , $\{\mathrm{L}_i\}_{i=1}^K$, respectively and created by an appropriate rotation of the same probability measure, which is spherically symmetric within a d-subspace and has a bounded and nontrivial support (i.e., its support is not a singleton).*

For $\epsilon > 0$, we say that μ_ϵ is a spherically symmetric HLM measure with noise level ϵ *if $\mu_\epsilon = \alpha_0 \mu_0 + \sum_{i=1}^K \alpha_i \mu_{i,\epsilon}$, where $\mu_{i,\epsilon} = \mu_i \times \nu_{i,\epsilon}$, $i = 1, \ldots, K$, $\{\alpha_i\}_{i=0}^K$, $\{\mathrm{L}_i\}_{i=1}^K$ and $\{\mu_i\}_{i=0}^K$ are the same as above and $\{\nu_{i,\epsilon}\}_{i=0}^K$ are probability measures with bound support in $\{\mathrm{L}_i^\perp\}_{i=1}^K$, first moments smaller than ϵ and p-th moments smaller than ϵ^p for $p < 1$.*

In Section 4.1 we discuss more general settings for an underlying HLM measure and the required modifications in our theory.

Throughout this paper we assume the condition

$$\alpha_1 > \sum_{i=2}^{K} \alpha_i \tag{3}$$

and consequently say that L_1 is the *most significant subspace*. For the noiseless case of $\epsilon = 0$, the most significant subspace coincides with the global l_0 subspace (i.e., the subspace containing the largest number of points) w.o.p. Indeed, the subspace with largest mixture component in the model is the subspace with largest number of sampled points w.o.p. For the noisy case of $\epsilon > 0$, we view condition (3) as a generalized notion of global l_0 subspace w.o.p., that is having the highest fraction of points "around" that subspace w.o.p.

1.4. Main Theorems

In the noiseless case and $0 < p \leq 1$, we can exactly recover the global l_0 subspace by l_p minimization as follows.

Theorem 1.1. *If μ is a spherically symmetric HLM measure on $\mathbb{R}^D$ with K d-subspaces $\{\mathrm{L}_i\}_{i=1}^K \subseteq \mathbb{R}^D$ and mixture coefficients $\{\alpha_i\}_{i=0}^K$ satisfying (3), $\mathcal{X}$ is a data set of N points identically and independently sampled from μ and $0 < p \leq 1$, then the probability that L_1 is a global l_p subspace is at least $1 - C\exp(-N/C)$, where C is a constant depending on D, d, K, p, α_0, α_1, μ_0, μ_1 and $\min_{2\leq i\leq K}(\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1, \mathrm{L}_i))$.*

In the noisy case, we extend the above formulation to near recovery. For this purpose we use the function

$$\psi_\mu(t) = \max_{\|\mathbf{v}\|=1} \left(\mu(\mathbf{x} \in \mathbb{R}^D : -t < |\mathbf{x}^T\mathbf{v}| < t)\right),$$

which we estimate for a special case in Appendix A.1.

Theorem 1.2. *If $\epsilon > 0$, μ_ϵ is a spherically symmetric HLM measure on $\mathbb{R}^D$ of noise level ϵ with K d-subspaces $\{\mathrm{L}_i\}_{i=1}^K \subseteq \mathbb{R}^D$ and mixture coefficients $\{\alpha_i\}_{i=0}^K$ satisfying (3), $\mathcal{X}$ is a data set of N points sampled identically and independently from μ_ϵ and $0 < p \leq 1$, then the global l_p subspace for μ_ϵ is in the ball $\mathrm{B}_{\mathrm{G}}(\mathrm{L}_1, f)$, where*

$$f \equiv f(\epsilon, K, d, p, \alpha_0, \alpha_1) = \frac{\pi\sqrt{d}\ \psi_{\mu_1}^{-1}\left(\frac{1+\mu_1(\{\mathbf{0}\})}{2}\right)\epsilon}{\left(\alpha_1 - \sum_{i=2}^K \alpha_i\right)^{\frac{1}{p}} (1-\mu_1(\{\mathbf{0}\}))^{\frac{1}{p}}\, 2^{\frac{p-3}{p}}}, \qquad (4)$$

w.p. at least $1 - C\exp(-N/C)$, where $C = C(\epsilon, p, d, D, \mu_1, \alpha_0, \alpha_1, \min_{2\leq i\leq K}(\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1, \mathrm{L}_i))$.

If $K = 1$, then the above statement extends for $1 < p < \infty$ with

$$f \equiv f(\epsilon, K, d, p, \alpha_0, \alpha_1) = \frac{\pi\sqrt{d}\ \psi_{\mu_1}^{-1}\left(\frac{1+\mu_1(\{\mathbf{0}\})}{2}\right) p^{\frac{1}{p}} \epsilon^{\frac{1}{p}}}{\alpha_1^{\frac{1}{p}} (1-\mu_1(\{\mathbf{0}\}))^{\frac{1}{p}}\, 2^{\frac{p-3}{p}}}.$$

In Section 3.5.3 we show that Theorem 1.2 is only relevant for sufficiently small ϵ.

At last, we formulate the impossibility of l_p recovery when $p > 1$ and $K > 1$ and thus demonstrate a phase transition at $p = 1$ when $K > 1$.

Theorem 1.3. *Assume that $\{\mathrm{L}_i\}_{i=1}^K$ 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 $\{\mathrm{L}_i\}_{i=1}^K$, let μ_ϵ be a spherically symmetric HLM measure on $\mathbb{R}^D$ of noise level ϵ w.r.t. $\{\mathrm{L}_i\}_{i=1}^K \subseteq \mathbb{R}^D$ and let $\mathcal{X}$ be a data set of N points sampled identically and independently from μ_ϵ. If $K > 1$ and $p > 1$, then for almost every $\{\mathrm{L}_i\}_{i=1}^K$ (w.r.t. $\gamma_{D,d}^K$), there exist positive constants δ_0 and κ_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 $\mathrm{B_G}(\mathrm{L}_1, \kappa_0)$ with overwhelming probability.*

We remark on the size of δ_0 and κ_0 in Section 3.6.5

1.5. Implications of the Theory to Subspace Modeling

Theorems 1.1, 1.2 and 1.3 provide some insights on the effectiveness of recovering the global l_0 d-subspace (or global l_0 strip of width ϵ as searched by RANSAC [13]) in a spherically symmetric HLM setting by minimizing l_p distances in the spirit of [26, 27]. In particular, they imply that if $K > 1$ then only l_p distances with $0 < p \leq 1$ should be considered. Even distances that coincide with the l_2 distance for sufficiently small values, such as [26, 27] or Huber's loss function [17], will not recover the underlying subspaces as the proofs of those theorems show. On the other hand, for a single underlying subspace with spherically symmetric outliers and possibly additive noise, l_p recovery should succeed in theory for any $0 < p < \infty$, though the bounding constants worsen as p increases. The idea of [26, 27] making the loss function constant for large values is expected to help with significantly far and nonsymmetric outliers (not covered by our model). Such outliers are discussed e.g., in Section 2.1.

In the setting of spherically symmetric HLM measure with no noise, Theorem 1.1 can be repetitively applied to justify sequential HLM using l_p minimization with $0 < p \leq 1$. Rigorous application of Theorem 1.2 for sequential HLM in the noisy case requires its extension to more general scenarios; such an extension depends on the precise way of removing the part of the data around a subspace. It also requires estimates of the local noise level (see e.g., [33, 32] and [6]).

1.6. Additional Results and Structure of the Paper

Section 2 reviews additional theory. In particular, Section 2.1 demonstrates natural instances, distinct from the case of spherically symmetric outliers, where the global l_0 subspace is neither a local l_p subspace (even for $p = 1$) nor global one (even for $0 < p < 1$); Section 2.2 studies the case of a single underlying subspace and asks when the global l_0 subspace is local minimum of the geometric l_p optimization. It establishes some necessary and sufficient conditions to solve such a problem. Unlike the rest of our theory, these conditions are model-independent and deterministic (i.e., not probabilistic); Section 2.3 uses those conditions to show that if one samples N_0 outliers and N_1 inliers from a spherically symmetric HLM model with $K = 1$ and if both $N_0 = o(N_1^2)$ and $p = 1$ or both $N_0 = \Omega(1)$ and $0 < p < 1$, then the global l_0 subspace is a local l_1 minimum. We separately include all mathematical details verifying

the theory of this paper in Section 3, while leaving some auxiliary verifications to the appendix. Section 4 concludes this paper and discusses some immediate extensions of its results as well as open directions.

2. Additional Theory

2.1. Counterexamples for Robustness of Best l_p Subspaces

We show here that there are many natural situations, though different than our underlying model of spherically symmetric outliers, 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.

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

If $p = 1$, then the global l_0 d-subspace in this example is still a local l_1 subspace. Nevertheless, if the outlier is located instead on a unit vector having elevation angle with the original d-subspace less than $\pi/2$, then ϵ can be chosen so that the global l_0 subspace is neither a local nor global l_1 subspace. However, if $0 < p < 1$, then the global l_0 subspace is still a local l_p subspace in both examples as well as almost any other scenario (see e.g., Proposition 2.1 below).

Similarly, it is not hard to produce an example of data points on the unit sphere of $\mathbb{R}^D$ where the global l_0 subspace is still not a global l_1 subspace. This is in contrast to the case of sparse representation of signals, where normalization of the column vectors of a matrix representing an undercomplete linear system of equations ensures that the solution minimizing the l_1 norm is also the sparsest solution as long as it is sufficiently sparse [11, Theorem 2]). For simplicity we give a counterexample for $d = 2$ by letting N_1 data points be uniformly sampled along an arc of length ϵ of a great circle of the sphere $\mathrm{S}^2 \subseteq \mathbb{R}^3$. We then place an outlier on another great circle, which passes through the center of the ϵ-arc and has a small angle with it. Taking ϵ sufficiently small and the outlier furthest from the intersection of the two great circles, we obtain that the global l_0 subspace is not a local l_1 subspace and consequently not a global one. We remark that in this example the assumption of bounded spherically symmetric outliers used throughout this paper is not satisfied.

2.2. Combinatorial Conditions for l_0 Subspaces being Local l_p Subspaces

2.2.1. Preliminary Notation

We denote the orthogonal group of $n \times n$ matrices by $\mathrm{O}(n)$ and the semigroup of $n \times n$ nonnegative scalar matrices by $\mathrm{S}_+(n)$. We designate the projection from $\mathbb{R}^D$ onto the d-subspace L by P_{L} and the corresponding orthogonal projection by $P_{\mathrm{L}}^{\perp}$. The nuclear

norm of $\mathbf{A}$ is denoted by $\|\mathbf{A}\|_*$. We define the scaled outlying "correlation" matrix $\mathbf{B}_{\mathrm{L},\mathcal{X}}$ of a data set $\mathcal{X}$ and a d-subspace L as follows

$$\mathbf{B}_{\mathrm{L},\mathcal{X}} = \sum_{\mathbf{x}\in\mathcal{X}\setminus\mathrm{L}} P_{\mathrm{L}}(\mathbf{x})P_{\mathrm{L}}^{\perp}(\mathbf{x})^T/\mathrm{dist}(\mathbf{x},\mathrm{L}). \tag{5}$$

Example 1. *Let $D = 2$, $d = 1$, $\mathcal{X} = \{(0,1),(1,1),(1,0)\}$ and* L *be the x-axis. Then*

$$\begin{aligned}\mathbf{B}_{\mathrm{L},\mathcal{X}} &= \sum_{\mathbf{x}\in\mathcal{X}\setminus\mathrm{L}} P_{\mathrm{L}}(\mathbf{x})\,P_{\mathrm{L}}^{\perp}(\mathbf{x})^T\mathrm{dist}(\mathbf{x},\mathrm{L})^{-1}\\ &= P_{\mathrm{L}}((0,1))\,P_{\mathrm{L}}^{\perp}((0,1))^T/\mathrm{dist}((0,1),\mathrm{L}) + P_{\mathrm{L}}((1,1))\,P_{\mathrm{L}}^{\perp}((1,1))^T/\mathrm{dist}((1,1),\mathrm{L})\\ &= (0,0)^T\,(0,1)/1 + (1,0)^T\,(0,1)/1 = \begin{pmatrix} 0 & 1\\ 1 & 0\end{pmatrix}.\end{aligned}$$

2.2.2. The Three Conditions

We formulate conditions for the global l_0 subspace to be a local l_p subspace, while distinguishing between three cases: $p = 1$, $0 < p < 1$ and $p > 1$. We prove these results in Section 3.2.

Theorem 2.1. *If $\mathrm{L}_1 \in \mathrm{G}(D,d)$, $\mathcal{X}_1 = \{\mathbf{x}_i\}_{i=1}^{N_1} \in \mathrm{L}_1$, $\mathcal{X}_0 = \{\mathbf{y_i}\}_{i=1}^{N_0} \in \mathbb{R}^D \setminus \mathrm{L}_1$ and $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1$, then a sufficient condition for L_1 to be a local minimum of $e_{l_1}(\mathcal{X},\mathrm{L})$ among all d-subspaces $\mathrm{L} \in \mathrm{G}(D,d)$ is that for any $\mathbf{V} \in \mathrm{O}(d)$ and $\mathbf{C} \in \mathrm{S}_+(d)$:*

$$\sum_{i=1}^{N_1} ||\mathbf{C}\mathbf{V}P_{\mathrm{L}_1}(\mathbf{x}_i)|| > ||\mathbf{C}\mathbf{V}\mathbf{B}_{\mathrm{L}_1,\mathcal{X}}\|_*\,. \tag{6}$$

Proposition 2.1. *If $\mathrm{L}_1 \in \mathrm{G}(D,d)$, $\mathcal{X}_1 = \{\mathbf{x}_i\}_{i=1}^{N_1} \in \mathrm{L}_1$, $\mathcal{X}_0 = \{\mathbf{y_i}\}_{i=1}^{N_0} \in \mathbb{R}^D \setminus \mathrm{L}_1$, $\mathrm{Sp}(\{\mathbf{x}_i\}_{i=1}^{N_1}) = \mathrm{L}_1$ and $p < 1$, then L_1 is a local minimum of $e_{l_p}(\mathcal{X},\mathrm{L})$ among all $\mathrm{L} \in \mathrm{G}(D,d)$.*

Proposition 2.2. *If $\mathrm{L}_1 \in \mathrm{G}(D,d)$, $\mathcal{X}_1 = \{\mathbf{x}_i\}_{i=1}^{N_1} \in \mathrm{L}_1$, $\mathcal{X}_0 = \{\mathbf{y_i}\}_{i=1}^{N_0} \in \mathbb{R}^D \setminus \mathrm{L}_1$ and $p > 1$, then a necessary condition for L_1 to be a local minimum of $e_{l_p}(\mathcal{X},\mathrm{L})$ among all $\mathrm{L} \in \mathrm{G}(D,d)$ is*

$$\sum_{i=1}^{N_0} P_{\mathrm{L}_1}(\mathbf{y}_i)P_{\mathrm{L}_1}^{\perp}(\mathbf{y}_i)^T\mathrm{dist}(\mathbf{y}_i,\mathrm{L}_1)^{p-2} = 0. \tag{7}$$

The above results manifest a phase transition phenomenon. Indeed, the global l_0 subspace is almost always a local l_p subspace for $p < 1$, whereas for $p > 1$ this is often not the case (except for an underlying measure which is spherically symmetric in the complement of L_1; for example, in the case of an underlying spherically symmetric HLM with $K = 1$, the global l_0 subspace is asymptotically a global l_p subspace for all $p > 0$). The combinatorial condition implying when it is a local l_1 subspace is more complicated and we exemplify its application throughout the paper.

2.3. Local or Global l_p Subspaces for Spherically Symmetric Sampling with a Single Subspace

We assume here the probabilistic setting of spherically symmetric HLM measure with a single underlying subspace L_1, i.e., $K = 1$. Clearly, L_1 is the global l_0 subspace for the sampled data w.o.p. For any $p > 0$, we ask whether L_1 is also a local or even global l_p subspace w.o.p. We prove the corresponding results described below in Section 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 sufficiently large. In order to simplify our estimates we assume that the support of the underlying distribution lies in the unit ball.

Theorem 2.2. *If $\mathrm{L}_1 \in \mathrm{G}(D, d)$ and $\mathcal{X}$ is a data set in $\mathbb{R}^D$ of $N_0 + N_1$ points, where N_0 of them are identically and independently sampled from a spherically symmetric distribution on $\mathrm{B}(\mathbf{0}, 1)$ and N_1 of them are identically and independently sampled from a spherically symmetric distribution on $\mathrm{L}_1 \cap \mathrm{B}(\mathbf{0}, 1)$ with nontrivial support; Then L_1 is a local l_1 subspace of $\mathcal{X}$ w.p. at least*

$$1 - 2d^2 \exp\left(-\frac{N_1\eta^2}{8d^2}\right) - 2dD \exp\left(-\frac{N_0\epsilon^2}{2d^2 D}\right), \quad \textit{where} \quad \eta + \frac{N_0}{N_1}\epsilon < \delta_*(\mu_1),$$

and $\delta_*(\mu_1)$ is a constant depending only on μ_1.

In particular, if $N_0 = o(N_1^2)$, then L_1 is a local l_1 subspace of $\mathcal{X}$ w.p. at least

$$1 - 2d^2 \exp\left(-\frac{\delta_*(\mu_1)^2 N_1}{72\, d^2}\right) - 2dD \exp\left(-\frac{\delta_*(\mu_1)^2 N_1^2}{8\, d^2 D N_0}\right). \tag{8}$$

In Appendix A.5 we establish the following expression for the constant $\delta_*(\mu_1)$ in the special case where μ_1 is the uniform distribution on $\mathrm{L}_1 \cap \mathrm{B}(\mathbf{0}, 1)$:

$$\delta_*(\mu_1) = 1/(d + 2). \tag{9}$$

For $0 < p < 1$, Proposition 2.1 implies that if $N_1 = \Omega(1)$ then L_1 is a local l_p subspace w.o.p. On the other hand if $p > 1$ and $N_1 = \Omega(1)$, then the following proposition shows that the subspace L_1 is a local l_p subspace w.p. 0.

Proposition 2.3. *Consider $\mathrm{L}_1 \in \mathrm{G}(D, d)$, μ_0 a spherically symmetric distribution on $\mathbb{R}^D$ with bounded support satisfying $\mu_0(\{\mathbf{0}\}) = 0$, μ_1 a spherically symmetric distribution on L_1 with bounded and nontrivial support, $\mu = \alpha_0\mu_0 + \alpha_1\mu_1$, where α_0, α_1 are nonnegative numbers summing to 1 and $\mathcal{X}$ is a data set sampled identically and independently from μ. If $p > 1$, then the probability that L_1 is a local l_p subspace of $\mathcal{X}$ is 0.*

The proof of this proposition is immediate. Indeed, denoting the i.i.d. outliers sampled from μ_0 by $\{\mathbf{y}_i\}_{i=1}^{N_0}$ and applying (82), the probability that $P_{\mathrm{L}_1}(\mathbf{y}_i)$ is a fixed number is zero. Therefore, the probability that $\sum_{i=1}^{N_0} P_{\mathrm{L}_1}(\mathbf{y}_i) P_{\mathrm{L}_1}^{\perp}(\mathbf{y}_i)^T \mathrm{dist}(\mathbf{y}_i, \mathrm{L}_1)^{p-2} = \mathbf{0}$ is also zero.

Another question is whether the global l_0 subspace is also the global l_p subspace. Proposition 2.3 and Theorem 1.1 already answered this question in our setting. Indeed,

if $p > 1$, then by Proposition 2.3 the global l_0 subspace is a global l_p subspace with probability 0; whereas if $0 < p \leq 1$, then Theorem 1.1 with $K = 1$ implies that for $N_0 = O(N_1)$ the global l_0 subspace is also the global l_p subspace w.o.p.

At last, we remark that the phase transition phenomenon demonstrated above at $p = 1$ is rather artificial in the current setting. Indeed, this phase transition is based on the fact that (7) holds w.p. 0 for $p > 1$ and any finite sample; however, the LHS of (7) divided by N is 0 w.p. 1 as N approaches infinity. Moreover, when $p > 1$ the positive distance between the global l_0 subspace and the global l_p subspace approaches 0 as N approaches infinity. We will show in Theorem 1.2 that this formal phase transition also breaks down with noise. Nevertheless, as we show in Theorem 1.3, there is a clear phase transition for a spherically symmetric HLM model with $K > 1$. This is rather intuitive since the underlying measure of the latter case is not spherically symmetric on the complement of L_1, unlike the case where $K = 1$.

3. Verification of Theory

We describe here the complete proofs of the various theorems and propositions of this paper. We start with preliminary notation and conventions as well statements of auxiliary lemmata and then prove the theory according to the following order of sections: 2.2, 2.3 and 1.

3.1. Preliminaries

3.1.1. Basic Notation and Conventions

All distributions in the statements of theorems have bounded supports. We assume WLOG that the support of these distributions is contained in $\mathrm{B}(\mathbf{0}, 1)$.

The Frobenius norm of $\mathbf{A}$ is denoted by $\|\mathbf{A}\|_{\mathrm{F}}$. The $n \times n$ identity matrix is written as $\mathbf{I}_n$. We denote the subset of $\mathrm{S}_+(n)$ with Frobenius norm 1 by $\mathrm{NS}_+(n)$. If $m > n$ we let $\mathrm{O}(m, n) = \{\mathbf{X} \in \mathbb{R}^{m \times n} : \mathbf{X}^T\mathbf{X} = \mathbf{I}_n\}$, whereas if $n > m$, $\mathrm{O}(m, n) = \{\mathbf{X} \in \mathbb{R}^{m \times n} : \mathbf{X}\mathbf{X}^T = \mathbf{I}_m\}$.

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

3.1.2. Auxiliary Lemmata

We formulate several technical lemmata, which will be proved in Appendices A.2-A.4.

Lemma 3.1. *If $\mathrm{L}_1, \hat{\mathrm{L}}_1 \in \mathrm{G}(D, d)$, $p > 0$, μ_1 is a spherically symmetric measure on L_1 with bounded and nontrivial support and $\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1, \hat{\mathrm{L}}_1) > \epsilon$, then*

$$E_{\mu_1}\left(e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1)\right) > \frac{(1 - \mu_1(\{\mathbf{0}\})) \cdot 2^{p-1} \cdot \epsilon^p}{\left(\pi\sqrt{d}\right)^p \cdot \left(\psi_{\mu_1}^{-1}\left((1 + \mu_1(\{\mathbf{0}\}))/2\right)\right)^p}.$$

Lemma 3.2. *For any* $\mathbf{x} \in \mathbb{R}^D$ *and* $\mathrm{L}_1, \mathrm{L}_2 \in \mathrm{G}(D,d)$*:*

$$|\mathrm{dist}(\mathbf{x}, \mathrm{L}_1) - \mathrm{dist}(\mathbf{x}, \mathrm{L}_2)| \leq \|\mathbf{x}\| \, \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1, \mathrm{L}_2).$$

Lemma 3.3. *If* $\mathrm{L}_1, \mathrm{L}_2 \in \mathrm{G}(D,d)$*,* μ_1 *and* μ_2 *are probability measures supported within* L_1 *and* L_2 *respectively and created by an appropriate rotation of the same probability measure, which is spherically symmetric within a d-subspace and has a bounded and nontrivial support (i.e., not a singleton), and* $p \leq 1$*, then for any* $\hat{\mathrm{L}} \in \mathrm{G}(D,d)$*:*

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}_1 \in \mu_1}(\mathrm{dist}(\mathbf{x}_1, \hat{\mathrm{L}})^p) + \mathbb{E}_{\mathbf{x}_2 \in \mu_2}(\mathrm{dist}(\mathbf{x}_2, \hat{\mathrm{L}})^p) \\ \geq &\mathbb{E}_{\mathbf{x}_1 \in \mu_1}(\mathrm{dist}(\mathbf{x}_1, \mathrm{L}_i)^p) + \mathbb{E}_{\mathbf{x}_2 \in \mu_2}(\mathrm{dist}(\mathbf{x}_2, \mathrm{L}_i)^p) \ \text{ for } i = 1, 2. \end{aligned} \tag{10}$$

3.2. Proofs for Theory of Section 2.2: 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 denote the principal angles [15] between two d-subspaces F and G by $\pi/2 \geq \theta_1 \geq \theta_2 \geq \cdots \geq \theta_d \geq 0$, where we order them decreasingly, unlike common notation. We denote by $k = k(\mathrm{F}, \mathrm{G})$ the largest number such that $\theta_k \neq 0$, so that $\theta_1 \geq \ldots \geq \theta_k > \theta_{k+1} = \ldots = \theta_d = 0$. We refer to this number as 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 $\{\mathbf{v}_i\}_{i=1}^d$ and $\{\mathbf{v}'_i\}_{i=1}^d$ of F and G respectively are two orthogonal bases for F and G satisfying

$$\langle \mathbf{v}_i, \mathbf{v}'_i \rangle = \cos(\theta_i), \quad \text{for } i = 1, \ldots, d,$$

and

$$\mathbf{v}_i \perp \mathbf{v}'_j, \quad \text{for all } 1 \leq i \neq j \leq k.$$

We define the complementary orthogonal system $\{\mathbf{u}_i\}_{i=1}^d$ for G with respect to F by the formula:

$$\begin{cases} \mathbf{v}'_i = \cos(\theta_i)\mathbf{v}_i + \sin(\theta_i)\mathbf{u}_i, & i = 1, 2, \cdots, k, \\ \mathbf{u}_i = \mathbf{v}_i, & i = k+1, \cdots, d. \end{cases} \tag{11}$$

We note that

$$\mathbf{u}_i \perp \mathbf{v}_j \ \text{ for all } \ 1 \leq i, j \leq k\,.$$

We note that the above vectors orthogonally decompose F+G into the 2-dimensional subspaces $\mathrm{Sp}(\mathbf{v}_i, \mathbf{u}_i)$, $i = 1, \ldots, k$, of mutually orthogonal systems and the residual subspace $\mathrm{F} \cap \mathrm{G}$. The interaction between F and G can then be described only within these subspaces via the principal angles. This idea is also motivated by purely geometric intuition in [28, Section 2].

We implicitly use principal vectors to represent $\mathrm{G}(D,d)$ by $\mathrm{O}(d) \times \mathrm{O}(d, D-d) \times \mathrm{S}_+(d)$. Indeed, we fix a d-subspace $\mathrm{L}_1 \in \mathrm{G}(D,d)$ and for any $\mathrm{L} \in \mathrm{G}(D,d)$ we form

the principal vectors $\{\mathbf{v}_i\}_{i=1}^d$ and $\{\mathbf{v}'_i\}_{i=1}^d$ for L_1 and L respectively; the projection of $\{\mathbf{v}_i\}_{i=1}^d$ onto L_1 corresponds to an element of $\mathrm{O}(d)$; the projection of $\{\mathbf{v}'_i\}_{i=1}^d$ (or the complementary vectors $\{\mathbf{u}_i\}_{i=1}^d$ of L w.r.t. L_1) onto $\mathrm{L}_1^\perp$ gives rise to an element of $\mathrm{O}(d, D-d)$; The principal angles in S_+ then relate elements projected onto $\mathrm{L}_1^\perp$ and L_1. Our representation is rather different than the common representation in numerical computation [12, Table 2.1], which uses either of the quotient spaces: $\mathrm{O}(D,d)/\mathrm{O}(d)$ or $\mathrm{O}(D)/(\mathrm{O}(d)\times\mathrm{O}(D-d))$.

It follows from [28, 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 [12, Theorem 2.3], we can parametrize this line from F to G by the following function L: [0,1]$\to \mathrm{G}(D,d)$, which is expressed in terms of the principal angles $\{\theta_i\}_{i=1}^d$ of F and G, the principal vectors $\{\mathbf{v}_i\}_{i=1}^d$ of F and the complementary orthogonal system $\{\mathbf{u}\}_{i=1}^d$ of G with respect to F:

$$\mathrm{L}(t) = \mathrm{Sp}(\{\cos(t\theta_i)\mathbf{v}_i + \sin(t\theta_i)\mathbf{u}_i\}_{i=1}^d). \tag{12}$$

We remark that this formula only holds when equipping the Grassmannian with the distance dist_G of (2) and this is the reason why we use this distance.

3.2.2. *Proof of Theorem 2.1*

In order to show that L_1 is a local minimum of $e_{l_1}(\mathcal{X}, \mathrm{L})$ among all d-subspaces in $\mathrm{G}(D,d)$, we arbitrarily fix a d-subspace $\hat{\mathrm{L}} \in \mathrm{B}_\mathrm{G}(\mathrm{L}_1, 1)$ and show that the derivative of the l_1 energy when restricted to the geodesic line from L_1 to an arbitrary subspace $\hat{\mathrm{L}}$ is positive at L_1.

The restriction of $\hat{\mathrm{L}}$ to $\mathrm{B}_\mathrm{G}(\mathrm{L}_1, 1)$ implies that $\theta_1 \leq 1$ and thus by [28, Theorem 9] this geodesic line (connecting L_1 and $\hat{\mathrm{L}}$) is unique. We parametrize it by the function L: [0,1]$\to \mathrm{G}(D,d)$ of (12), where here $\{\theta_i\}_{i=1}^d$ are the principal angles between L_1 and $\hat{\mathrm{L}}$, $\{\mathbf{v}_i\}_{i=1}^d$ are the principal vectors of L_1 and $\{\mathbf{u}\}_{i=1}^d$ are the complementary orthogonal system for $\hat{\mathrm{L}}$ with respect to L_1. Using this parametrization we need to prove that the function $e_{l_1}(\mathcal{X}, \mathrm{L}(t))$: [0,1]$\to \mathbb{R}$ has a positive derivative at $t = 0$.

We follow by simplifying the expression for the function $e_{l_1}(\mathcal{X}, \mathrm{L}(t))$ and its derivative according to t. We denote the projection from $\mathbb{R}^D$ onto $\mathrm{Sp}(\mathbf{v}_j, \mathbf{u}_j)$, where $1 \leq j \leq d$, by P_j and the projection from $\mathbb{R}^D$ onto $(\mathrm{L}_1 + \hat{\mathrm{L}})^\perp$ by $P^\perp$ and use this notation to express the following components of the function $e_{l_1}(\mathcal{X}, \mathrm{L}(t))$ for $i = 1, \ldots, N_1$:

$$\mathrm{dist}(\mathbf{y}_i, \mathrm{L}(t)) = \sqrt{\sum_{j=1}^d \mathrm{dist}^2(P_j(\mathbf{y}_i), \mathrm{L}(t)) + \mathrm{dist}^2(P^\perp(\mathbf{y}_i), \mathrm{L}(t))}. \tag{13}$$

For $1 \leq j \leq d$, we let $\phi_j \in [0, 2\pi]$ denote the angle such that $P_j(\mathbf{y}_i) = ||P_j(\mathbf{y}_i)||$ $(\cos(\phi_j)\mathbf{v}_j + \sin(\phi_j)\mathbf{u}_j)$ and consequently express each term of the sum in (13) as follows:

$$\mathrm{dist}^2(P_j(\mathbf{y}_i), \mathrm{L}(t)) = ||P_j(\mathbf{y}_i)||^2 \sin^2(\phi_j - t\theta_j), \quad j = 1, \ldots, d\,. \tag{14}$$

Applying (14) in (13) and differentiating, we obtain the following expression for the derivative of $\text{dist}(\mathbf{y}_i, \mathrm{L}(t))$ for all $1 \le i \le N_0$:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\text{dist}(\mathbf{y}_i, \mathrm{L}(t))\right) &= -\frac{\sum_{j=1}^{d} \theta_i ||P_j(\mathbf{y}_i)||^2 \sin(\phi_j - t\theta_j)\cos(\phi_j - t\theta_j)}{\text{dist}(\mathbf{y}_i, \mathrm{L}(t))} \\ &= -\frac{\sum_{j=1}^{d} \theta_j \left((\cos(t\theta_j)\mathbf{v}_j + \sin(t\theta_j)\mathbf{u}_j)\cdot \mathbf{y}_i\right)\left((-\sin(t\theta_j)\mathbf{v}_j + \cos(t\theta_j)\mathbf{u}_j)\cdot \mathbf{y}_i\right)}{\text{dist}(\mathbf{y}_i, \mathrm{L}(t))}.\end{aligned} \tag{15}$$

At $t = 0$ it becomes

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\text{dist}(\mathbf{y}_i, \mathrm{L}(t))\right)\Big|_{t=0} &= -\frac{\sum_{j=1}^{d} \theta_j (\mathbf{v}_j \cdot \mathbf{y}_i)(\mathbf{u}_j \cdot \mathbf{y}_i)}{\text{dist}(\mathbf{y}_i, \mathrm{L}(0))} \\ &= -\frac{\sum_{j=1}^{k} \theta_j (\mathbf{v}_j \cdot \mathbf{y}_i)(\mathbf{u}_j \cdot \mathbf{y}_i)}{\text{dist}(\mathbf{y}_i, \mathrm{L}(0))},\end{aligned} \tag{16}$$

where the interaction dimension $k = k(\mathrm{L}_1, \hat{\mathrm{L}})$ has been introduced in Section 3.2.1.

We form the following matrices: $\mathbf{C} = \mathrm{diag}(\theta_1, \theta_2, \cdots, \theta_d)$, $\mathbf{V} \in \mathrm{O}(d, D)$ with j-th row $\mathbf{v}_j^T$ and $\mathbf{U} \in \mathrm{O}(k, D)$ with j-th row $\mathbf{u}_j^T$. We then reformulate (16) using these matrices as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\text{dist}(\mathbf{y}_i, \mathrm{L}(t))\right)\Big|_{t=0} = -\frac{\mathrm{tr}_k(\mathbf{C}\mathbf{V}\mathbf{y}_i\mathbf{y}_i^T\mathbf{U}^T)}{\text{dist}(\mathbf{y}_i, \mathrm{L}_1)}, \tag{17}$$

where tr_k denotes the trace of the first k rows of the corresponding $d \times k$ matrix, whose last $d - k$ rows are zeros. Similarly, for all $\mathbf{x}_i \in \mathrm{L}_1$, $i = 1, 2, \cdots, N_1$,

$$\text{dist}(\mathbf{x}_i, \mathrm{L}(t)) = \sqrt{\sum_{j=1}^{d} |(\mathbf{v}_j \cdot \mathbf{x}_i)|^2 \sin^2(t\theta_j)},$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\text{dist}(\mathbf{x}_i, \mathrm{L}(t))\right) = \frac{\sum_{j=1}^{d} \theta_j |\mathbf{v}_j \cdot \mathbf{x}_i|^2 \sin(t\theta_j)\cos(t\theta_j)}{\text{dist}(\mathbf{x}_i, \mathrm{L}(t))}. \tag{18}$$

At $t = 0$, this derivative becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\text{dist}(\mathbf{x}_i, \mathrm{L}(t))\right)\Big|_{t=0} = \sqrt{\sum_{j=1}^{d} |(\mathbf{v}_j \cdot \mathbf{x}_i)|^2 \theta_j^2} = ||\mathbf{C}\mathbf{V}\mathbf{x}_i||. \tag{19}$$

Combining (17) and (19) and using

$$\mathbf{A} := \sum_{i=1}^{N_0} \mathbf{y}_i^T \mathbf{y}_i / \text{dist}(\mathbf{y}_i, \mathrm{L}_1),$$

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

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e_{l_1}(\mathcal{X},\mathrm{L}(t))\right)\Big|_{t=0} = \sum_{i=1}^{N_1} ||\mathbf{C}\mathbf{V}\mathbf{x}_i|| - \mathrm{tr}_k(\mathbf{C}\mathbf{V}\mathbf{A}\mathbf{U}^T). \tag{20}$$

Since $\mathbf{V}$ is a projection onto L_1 and $\mathbf{U}$ is a projection onto $\mathrm{L}_1^{\perp}$, we may rewrite this expression by the matrix $\hat{\mathbf{V}} \in \mathrm{O}(d)$, whose j-th row is $P_{\mathrm{L}_1}(\mathbf{v}_j)^T$ and the matrix $\hat{\mathbf{U}} \in \mathrm{O}(k, D-d)$, whose j-th row is $P_{\mathrm{L}_1}^{\perp}(\mathbf{v}_j)^T$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e_{l_1}(\mathcal{X},\mathrm{L}(t))\right)\Big|_{t=0} = \sum_{i=1}^{N_1} ||\mathbf{C}\hat{\mathbf{V}}P_{\mathrm{L}_1}\mathbf{x}_i|| - \mathrm{tr}_k(\mathbf{C}\hat{\mathbf{V}}\mathbf{B}_{\mathrm{L}_1,\mathcal{X}}\hat{\mathbf{U}}^T). \tag{21}$$

At last, we note that

$$\max_{\hat{\mathbf{U}}^T}(\mathrm{tr}_k(\mathbf{C}\hat{\mathbf{V}}\mathbf{B}_{\mathrm{L}_1,\mathcal{X}}\hat{\mathbf{U}}^T)) = ||\mathbf{C}\hat{\mathbf{V}}\mathbf{B}_{\mathrm{L}_1,\mathcal{X}}||_*. \tag{22}$$

Indeed, denoting the SVD decomposition of $\mathbf{C}\hat{\mathbf{V}}\mathbf{B}_{\mathrm{L}_1,\mathcal{X}}$ by $\mathbf{U}_0\mathbf{\Sigma}_0\mathbf{V}_0^T$ we have that

$$\begin{aligned}\mathrm{tr}_k(\mathbf{C}\hat{\mathbf{V}}\mathbf{B}_{\mathrm{L}_1,\mathcal{X}}\hat{\mathbf{U}}^T) &= \mathrm{tr}_k(\mathbf{U}_0\mathbf{\Sigma_0}\mathbf{V}_0^T\hat{\mathbf{U}}^T) = \mathrm{tr}_k(\mathbf{\Sigma_0}\mathbf{V}_0^T\hat{\mathbf{U}}^T\mathbf{U}_0) \leq \sum(\mathrm{diag}(\mathbf{\Sigma_0})) \\ &= ||\mathbf{C}\hat{\mathbf{V}}\mathbf{B}_{\mathrm{L}_1,\mathcal{X}}||_*\end{aligned}$$

and this equality can be achieved when $\hat{\mathbf{U}}^T$ consists of the first k columns of $\mathbf{V}_0\mathbf{U}_0^T$. The theorem is thus concluded by combing (21) and (22).

3.2.3. Simultaneous Proof for Both Propositions 2.1 and 2.2

For the d-subspace L_1 and an arbitrary d-subspace $\hat{\mathrm{L}} \in \mathrm{B_G}(\mathrm{L}_1, 1)$, we form the geodesic line parametrization $\mathrm{L}(t)$ and the corresponding matrices $\mathbf{C}$, $\mathbf{V}$, $\mathbf{U}$, $\hat{\mathbf{V}}$ and $\hat{\mathbf{U}}$ as in the proof of Theorem 2.1. Similarly to verifying (17) and (19) in the latter proof, we obtain that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{dist}(\mathbf{y}_i,\mathrm{L}(t))^p\right)\Big|_{t=0} = -p\,\mathrm{dist}(\mathbf{y}_i,\mathrm{L}_1)^{p-2}\,\mathrm{tr}_k(\mathbf{C}\mathbf{V}\mathbf{y}_i\mathbf{y}_i^T\mathbf{U}^T) \tag{23}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{dist}(\mathbf{x}_i,\mathrm{L}(t))^p\right)\Big|_{t=0} = p\,\mathrm{dist}(\mathbf{x}_i,\mathrm{L}_1)^{p-1}||\mathbf{C}\mathbf{V}\mathbf{x}_i||. \tag{24}$$

Consequently

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e_{l_p}(\mathcal{X},\mathrm{L}(t))\right)\Big|_{t=0} = p\sum_{i=1}^{N_1}\mathrm{dist}(\mathbf{x}_i,\mathrm{L}_1)^{p-1}||\mathbf{C}\mathbf{V}\mathbf{x}_i|| \tag{25}$$

$$-\,p\sum_{i=1}^{N_0}\mathrm{dist}(\mathbf{y}_i,\mathrm{L}_1)^{p-2}\,\mathrm{tr}_k(\mathbf{C}\mathbf{V}\mathbf{y}_i\mathbf{y}_i^T\mathbf{U}^T) = p\sum_{i=1}^{N_1}\mathrm{dist}(\mathbf{x}_i,\mathrm{L}_1)^{p-1}||\mathbf{C}\hat{\mathbf{V}}P_{\mathrm{L}_1}(\mathbf{x}_i)||$$

$$- p \sum_{i=1}^{N_0} \mathrm{dist}(\mathbf{y}_i, \mathrm{L}_1)^{p-2} \,\mathrm{tr}_k(\mathbf{C}\hat{\mathbf{V}} P_{\mathrm{L}_1}(\mathbf{y}_i) P_{\mathrm{L}_1}^{\perp}(\mathbf{y}_i)^T \hat{\mathbf{U}}^T).$$

Assume first that $p < 1$. Then

$$\frac{\mathrm{d}}{\mathrm{d}t^p} \left(e_{l_p}(\mathcal{X}, \mathrm{L}(t)) \right) \Big|_{t=0} = p\, t^{1-p} \sum_{i=1}^{N_1} \mathrm{dist}(\mathbf{x}_i, \mathrm{L}_1)^{p-1} ||\mathbf{C}\hat{\mathbf{V}} P_{\mathrm{L}_1}(\mathbf{x}_i)|| \tag{26}$$

$$- p\, t^{1-p} \sum_{i=1}^{N_0} \mathrm{dist}(\mathbf{y}_i, \mathrm{L}_1)^{p-2} \,\mathrm{tr}_k(\mathbf{C}\hat{\mathbf{V}} P_{\mathrm{L}_1}(\mathbf{y}_i) P_{\mathrm{L}_1}^{\perp}(\mathbf{y}_i)^T \hat{\mathbf{U}}^T)$$

$$= p \sum_{i=1}^{N_1} \left(\lim_{t \to 0} \mathrm{dist}(\mathbf{x}_i, \mathrm{L}(t))/t \right)^{p-1} ||\mathbf{C}\hat{\mathbf{V}} P_{\mathrm{L}_1}(\mathbf{x}_i)|| = \sum_{i=1}^{N_0} ||\mathbf{C}\hat{\mathbf{V}} P_{\mathrm{L}_1}(\mathbf{x}_i)||^p.$$

It follows immediately from the definitions of $\mathbf{C}$ and $\mathbf{V}$ that

$$||\mathbf{C}\mathbf{V}\mathbf{x}_i|| \geq \theta_1 \, ||\mathbf{v}_1^T \mathbf{x}_i||. \tag{27}$$

Now, the assumption $\mathrm{Sp}(\{\mathbf{x}_i\}_{i=1}^{N_1}) = \mathrm{L}_1$ implies that there exists $1 \leq j \leq N_1$ such that $\mathbf{v}_1^T \mathbf{x}_j \neq 0$ and thus $||\mathbf{C}\hat{\mathbf{V}} P_{\mathrm{L}_1}(\mathbf{x}_i)|| = ||\mathbf{C}\mathbf{V}\mathbf{x}_i|| > 0$. Therefore, (26) is positive, L_1 is a local minimum of $e_{l_p}(\mathcal{X}, \mathrm{L}(t))$ and Proposition 2.1 is proved.

Next, assume that $p > 1$ and note that

$$p \sum_{i=1}^{N_1} \mathrm{dist}(\mathbf{x}_i, \mathrm{L}_1)^{p-1} \|\mathbf{C}\hat{\mathbf{V}} P_{\mathrm{L}_1} \mathbf{x}_i\| = 0. \tag{28}$$

Since L_1 is a local minimum of $e_{l_p}(\mathcal{X}, \mathrm{L})$, the derivative in (25) is nonnegative and in view of (28), the subtracted term in (25) is thus nonpositive. Now, for a subspace $\hat{\mathrm{L}} \in \mathrm{G}(D, d)$ such that $\mathbf{C} = \hat{\mathbf{V}} = \mathbf{I}_d$ we obtain that

$$0 \geq \max_{\hat{\mathbf{U}}} p \sum_{i=1}^{N_0} \mathrm{dist}(\mathbf{y}_i, \mathrm{L}_1)^{p-2} \,\mathrm{tr}_k(P_{\mathrm{L}_1}(\mathbf{y}_i) P_{\mathrm{L}_1}^{\perp}(\mathbf{y}_i)^T \hat{\mathbf{U}}^T)$$

$$= p \left\| \sum_{i=1}^{N_0} \mathrm{dist}(\mathbf{y}_i, \mathrm{L}_1)^{p-2} P_{\mathrm{L}_1}(\mathbf{y}_i) P_{\mathrm{L}_1}^{\perp}(\mathbf{y}_i)^T \right\|_*,$$

where the last equality follows from (22). Therefore, (7) holds and Proposition 2.2 is thus proved.

3.3. Proof of Theorem 2.2: Combination of Combinatorial Estimates (Section 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 (6) for arbitrary $\hat{\mathrm{L}} \in \mathrm{B}_{\mathrm{G}}(\mathrm{L}_1, 1)$. We use the similar notation as in the proof of Theorem 2.1, in particular, we denote the N_0 outliers and

N_1 inliers by $\{\mathbf{y_i}\}_{i=1}^{N_0}$ and $\{\mathbf{x_i}\}_{i=1}^{N_1}$ respectively. Due to the homogeneity of (6) in $\mathbf{C}$, we will assume WLOG that $||\mathbf{C}||_2 = 1$, i.e., $\theta_1 = 1$.

We start with estimating the probability that the RHS of (6) is small. Applying the above assumption that $||\mathbf{C}||_2 = 1$ we have that

$$||\mathbf{CVB}_{\mathrm{L}_1,\mathcal{X}}||_F \leq ||\mathbf{VB}_{\mathrm{L}_1,\mathcal{X}}||_F = ||\mathbf{B}_{\mathrm{L}_1,\mathcal{X}}||_F$$

and consequently

$$\Pr\left(\frac{||\mathbf{CVB}_{\mathrm{L}_1,\mathcal{X}}||_*}{N_0} < \epsilon\right) \geq \Pr\left(\frac{||\mathbf{CVB}_{\mathrm{L}_1,\mathcal{X}}||_F}{N_0} < \frac{\epsilon}{\sqrt{d}}\right)$$
$$\geq \Pr\left(\frac{||\mathbf{B}_{\mathrm{L}_1,\mathcal{X}}||_F}{N_0} < \frac{\epsilon}{\sqrt{d}}\right) \geq \Pr\left(\frac{\max_{p,l}|(\mathbf{B}_{\mathrm{L}_1,\mathcal{X}})_{p,l}|}{N_0} < \frac{\epsilon}{d\sqrt{D}}\right).$$

f We further estimate this probability by Hoeffding's inequality as follows: we view the matrix $\mathbf{B}_{\mathrm{L}_1,\mathcal{X}}$ as the sum of random variables $P_{\mathrm{L}_1}(\mathbf{y}_i)P_{\mathrm{L}_1}^{\perp}(\mathbf{y}_i)^T/||P_{\mathrm{L}_1}^{\perp}(\mathbf{y}_i)||$, $i = 1,\ldots,N_0$. Since the distribution of outliers is spherically symmetric in $\mathrm{B}(\mathbf{0},1)$, the coordinates of both $P_{\mathrm{L}_1}(\mathbf{y}_i)$ and $P_{\mathrm{L}_1}^{\perp}(\mathbf{y}_i)^T/||P_{\mathrm{L}_1}^{\perp}(\mathbf{y}_i)||$ have expectations 0 and take values in [-1,1]. We can thus apply Hoeffding's inequality to the sum defining $\mathbf{B}_{\mathrm{L}_1,\mathcal{X}}$ and consequently obtain that

$$\Pr\left(\frac{\max_{p,l}|(\mathbf{B}_{\mathrm{L}_1,\mathcal{X}})_{p,l}|}{N_0} < \frac{\epsilon}{d\sqrt{D}}\right) \geq 1 - 2dD\exp\left(-\frac{N_0\epsilon^2}{2d^2D}\right). \tag{29}$$

Next, we estimate the probability that the LHS of (6) is sufficiently large. Unlike the rest of the paper where we often represent P_{L_1} by a $D \times D$ projection matrix (of rank d), it will be convenient here to represent it as a $D \times d$ matrix of projection. We first note that

$$\sum_{i=1}^{N_1}||\mathbf{CV}P_{\mathrm{L}_1}(\mathbf{x}_i)|| \geq \sum_{i=1}^{N_1}|\theta_1\mathbf{v}_1^TP_{\mathrm{L}_1}(\mathbf{x}_i)| = \sum_{i=1}^{N_1}|\mathbf{v}_1^TP_{\mathrm{L}_1}(\mathbf{x}_i)|$$
$$\geq \sqrt{\sum_{i=1}^{N_1}|\mathbf{v}_1^TP_{\mathrm{L}_1}(\mathbf{x}_i)|^2} \geq \min_t \sigma_t\left(\sum_{i=1}^{N_1}P_{\mathrm{L}_1}(\mathbf{x}_i)P_{\mathrm{L}_1}(\mathbf{x}_i)^T\right). \tag{30}$$

Second of all, since μ_1 is spherically symmetric distribution in $\mathrm{L}_1 \cap \mathrm{B}(\mathbf{0},1)$ and given the representation of P_{L_1} by a $D \times d$ matrix, we have

$$E_{\mu_1}(P_{\mathrm{L}_1}(\mathbf{x})P_{\mathrm{L}_1}(\mathbf{x})^T) = \delta_*\mathbf{I}_d, \text{ where } \delta_* = \delta_*(\mu_1) \text{ depends on } \mu_1. \tag{31}$$

We will prove in Appendix A.6 the following statement:

$$\text{If } \max_t \sigma_t\left(\sum_{i=1}^{N_1}P_{\mathrm{L}_1}(\mathbf{x}_i)P_{\mathrm{L}_1}(\mathbf{x}_i)^T - \delta_*\mathbf{I}_d\right) < \eta,$$
$$\text{then } \min_t \sigma_t\left(\sum_{i=1}^{N_1}P_{\mathrm{L}_1}(\mathbf{x}_i)P_{\mathrm{L}_1}(\mathbf{x}_i)^T\right) > \delta_* - \eta. \tag{32}$$

We combine (30)-(32) and Hoeffding's inequality to obtain the following probabilistic estimate for the LHS of (6):

$$\begin{aligned}
&\Pr\left(\frac{\sum_{i=1}^{N_1}||\mathbf{CV}P_{\mathrm{L}_1}(\mathbf{x}_i)||}{N_1} > \delta_* - \eta\right) \qquad (33)\\
&\geq \Pr\left(\min_t \sigma_t\left(\frac{\sum_{i=1}^{N_1} P_{\mathrm{L}_1}(\mathbf{x}_i)P_{\mathrm{L}_1}(\mathbf{x}_i)^T}{N_1}\right) > \delta_* - \eta\right)\\
&\geq \Pr\left(\max_t \sigma_t\left(\frac{\sum_{i=1}^{N_1} P_{\mathrm{L}_1}(\mathbf{x}_i)P_{\mathrm{L}_1}(\mathbf{x}_i)^T}{N_1} - \delta_*\mathbf{I}_d\right) < \eta\right)\\
&\geq \Pr\left(\left\|\frac{\sum_{i=1}^{N_1} P_{\mathrm{L}_1}(\mathbf{x}_i)P_{\mathrm{L}_1}(\mathbf{x}_i)^T}{N_1} - \delta_*\mathbf{I}_d\right\|_F < \eta\right)\\
&\geq \Pr\left(\max_{p,l}\left|\frac{\sum_{i=1}^{N_1} P_{\mathrm{L}_1}(\mathbf{x}_i)P_{\mathrm{L}_1}(\mathbf{x}_i)^T}{N_1} - \delta_*\mathbf{I}_d\right|_{p,l} < \frac{\eta}{d}\right) \geq 1 - 2d^2\exp\left(-\frac{N_1\eta^2}{2d^2}\right).
\end{aligned}$$

From (29) and (33), (6) is valid with probability at least

$$1 - 2d^2\exp\left(-\frac{N_1\eta^2}{2d^2}\right) - 2dD\exp\left(-\frac{N_0\epsilon^2}{2d^2D}\right) \quad \forall\,\epsilon,\ \eta \text{ s.t. } \eta + \frac{N_0}{N_1}\epsilon < \delta_*(\mu_1). \qquad (34)$$

We can choose $\epsilon = N_1\delta_*(\mu_1)/(2N_0) = N_1/(2N_0(d+2))$, $\eta = 1/(3(d+2))$ and obtain that if $N_0 = o(N_1^2)$ then (6) is valid with the probability specified in (8).

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

3.4.1. Proof of the Special Case: $K = 1$

Part I: L_1 *is a Global* l_p *Subspace in* $\mathrm{B_G}(\mathrm{L}_1, \gamma_1)$

We assume here that there is only one underlying subspace, L_1, since it is easier to follow our proof in this case. We prove in this part that there exists a constant $\gamma_1 > 0$ such that w.o.p. L_1 is the global l_p subspace in $\mathrm{B_G}(\mathrm{L}_1, \gamma_1)$. We arbitrarily choose $\hat{\mathrm{L}} \in \mathrm{G}(D, d)$ such that $\mathrm{dist_G}(\hat{\mathrm{L}}, \mathrm{L}_1) = 1$ and parameterize a geodesic line from L_1 to $\hat{\mathrm{L}}$ by a function L: [0,1] $\to \mathrm{G}(D, d)$, where $\mathrm{L}(0) = \mathrm{L}_1$ and $\mathrm{L}(1) = \hat{\mathrm{L}}$. We then observe that there exists $\gamma_1 > 0$ such that the function $e_{l_1}(\mathcal{X}, \mathrm{L}(t))$: [0,1] $\to \mathbb{R}$ of (1) has a positive derivative w.o.p. at any $t \in [0, \gamma_1]$, that is,

$$\frac{\mathrm{d}}{\mathrm{d}t^p}\left(\frac{\sum_{\mathbf{x}\in\mathcal{X}}\mathrm{dist}(\mathbf{x}, \mathrm{L}(t))^p}{N}\right) > 0 \ \text{ for all } t \in [0, \gamma_1] \ \text{ w.o.p.} \qquad (35)$$

We will deduce (35) from the following two equations:

$$\frac{\mathrm{d}}{\mathrm{d}t^p}\left(\frac{\sum_{\mathbf{x}\in\mathcal{X}}\mathrm{dist}(\mathbf{x}, \mathrm{L}(t))^p}{N}\right)\Bigg|_{t=0} > \gamma_2 \text{ w.o.p. for some } \gamma_2 > 0. \qquad (36)$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t^p}\left(\frac{\sum_{\mathbf{x}\in\mathcal{X}}\mathrm{dist}(\mathbf{x},\mathrm{L}(t))^p}{N}\right)\Big|_{t=0} - \frac{\mathrm{d}}{\mathrm{d}t^p}\left(\frac{\sum_{\mathbf{x}\in\mathcal{X}}\mathrm{dist}(\mathbf{x},\mathrm{L}(t))^p}{N}\right)\Big|_{t=t_0} < \frac{\gamma_2}{2}, \tag{37}$$

$\forall t_0 \in [0,\gamma_1]$ w.o.p.

When $p = 1$, equation (36) practically follows from the proof of Theorem 2.2 by arbitrarily fixing ϵ and η such that $\epsilon\alpha_0/\alpha_1+\eta+\gamma_2/\alpha_1 < \delta_*$ and noting that when sampling from the mixture measure specified in the current theorem (unlike Theorem 2.2) the ratio of sampled outliers to inliers, N_0/N_1, goes w.o.p. to α_0/α_1. When $p < 1$, equation (36) follows from (26). We also observe that $\gamma_2 \equiv \gamma(\alpha_0,\alpha_1,d,\mu_1,p)$.

We first verify (37) for the sum of elements in $\mathcal{X}_1 = \mathcal{X}\cap \mathrm{L}_1$. In view of (18), for any $\mathbf{x}\in\mathcal{X}_1$ the single term in that sum (i.e., $\mathrm{dist}(\mathbf{x},\mathrm{L}(t))^p$) has a bounded second derivative with respect to t; hence, we can find constants γ_1 and γ_2 satisfying

$$\frac{\mathrm{d}}{\mathrm{d}t^p}\left(\frac{\sum_{\mathbf{x}\in\mathcal{X}_1}\mathrm{dist}(\mathbf{x},\mathrm{L}(t))^p}{N}\right)\Big|_{t=0} - \frac{\mathrm{d}}{\mathrm{d}t^p}\left(\frac{\sum_{\mathbf{x}\in\mathcal{X}_1}\mathrm{dist}(\mathbf{x},\mathrm{L}(t))^p}{N}\right)\Big|_{t=t_0} < \frac{\gamma_2}{6} \tag{38}$$

$\forall t_0 \in [0,\gamma_1]$.

We derive a similar estimate by replacing the summation of $\mathbf{x}\in\mathcal{X}_1$ by the summation of $\mathbf{x}\in\mathcal{X}\setminus\mathcal{X}_1$. Using the constant γ_3, which we clarify later, we separate the latter sum into two components: $\hat{\mathcal{X}} := \{\mathbf{x}\in\mathcal{X}\setminus\mathcal{X}_1 : \mathrm{dist}(\mathbf{x},\mathrm{L}_1)\le 2\,\gamma_3\}$ and $(\mathcal{X}\setminus\mathcal{X}_1)\setminus\hat{\mathcal{X}}$.

In order to deal with the first sum, we define

$$\gamma_4 := \mu(\mathbf{x} : 0 < \mathrm{dist}(\mathbf{x},\mathrm{L}_1)\le 2\,\gamma_3)$$

and note that we can choose $\gamma_3 \equiv \gamma_3(D,\gamma_2,\mu_0) \equiv \gamma_3(D,d,\alpha_0,\alpha_1,\mu_0,\mu_1,p)$ sufficiently small such that $\gamma_4 \equiv \gamma_4(d,\alpha_0,\alpha_1,\mu_0)$ is arbitrarily small. We use γ_4 to bound the ratio of sampled points from $\hat{\mathcal{X}}$ and $\mathcal{X}$ as follows:

$$\frac{\#(\hat{\mathcal{X}})}{\#(\mathcal{X})} \le 2\gamma_4 \ \text{ w.o.p.} \tag{39}$$

Indeed, we note that $\#(\hat{\mathcal{X}}) = \sum_{\mathbf{x}\in\mathcal{X}} I_{\hat{\mathcal{X}}}(\mathbf{x})$, $E(I_{\hat{\mathcal{X}}}(\mathbf{x})) = \mu(\mathbf{x}:\mathbf{x}\in\hat{\mathcal{X}}) = \gamma_4$ and $I_{\hat{\mathcal{X}}}(\mathbf{x})$ takes values in $[0,1]$, therefore by applying Hoeffding's inequality to $I_{\hat{\mathcal{X}}}(\mathbf{x})$, where $\mathbf{x}\in\mathcal{X}$, we conclude (39).

Now for $\mathbf{y}_i\in\hat{\mathcal{X}}$, the derivatives expressed in (15) and (26) are bounded by 1 since the support of μ_0 is contained in $\mathrm{B}(\mathbf{0},1)$. Thus, by combining this observation with (39) we obtain that there exists γ_3 and γ_4 such that for any $t_0\in[0,\gamma_1]$:

$$\frac{\mathrm{d}}{\mathrm{d}t^p}\left(\frac{\sum_{\mathbf{x}\in\hat{\mathcal{X}}}\mathrm{dist}(\mathbf{x},\mathrm{L}(t))^p}{N}\right)\Big|_{t=0} - \frac{\mathrm{d}}{\mathrm{d}t^p}\left(\frac{\sum_{\mathbf{x}\in\hat{\mathcal{X}}}\mathrm{dist}(\mathbf{x},\mathrm{L}(t))^p}{N}\right)\Big|_{t=t_0} < \frac{\gamma_2}{6} \tag{40}$$

w.o.p.

Differentiating (15) and (26) one more time, we obtain that for $\mathbf{x} \in (\mathcal{X} \setminus \mathcal{X}_1) \setminus \hat{\mathcal{X}}$, the second derivative of $\text{dist}(\mathbf{x}, \mathrm{L}(t))$ with respect to t^p is bounded by $C(d)/\gamma_3^3$. Thus we can choose $\gamma_1 \equiv \gamma_1(\gamma_2, \gamma_3, d) \equiv \gamma_1(\alpha_0, \alpha_1, \mu_0, \mu_1, d, D, p)$ sufficiently small such that for any $t_0 \in [0, \gamma_1]$:

$$\left.\frac{\mathrm{d}}{\mathrm{d}t^p}\frac{\sum_{\mathbf{x}\in(\mathcal{X}\setminus\mathcal{X}_1)\setminus\hat{\mathcal{X}}}\text{dist}(\mathbf{x},\mathrm{L}(t))^p}{N}\right|_{t=0} - \left.\frac{\mathrm{d}}{\mathrm{d}t^p}\frac{\sum_{\mathbf{x}\in(\mathcal{X}\setminus\mathcal{X}_1)\setminus\hat{\mathcal{X}}}\text{dist}(\mathbf{x},\mathrm{L}(t))^p}{N}\right|_{t=t_0} < \frac{\gamma_2}{6}. \tag{41}$$

Equation (37) and consequently (35) are thus verified by combing (38), (40) and (41). That is, we showed that L_1 is the global l_p subspace in $\mathrm{B_G}(\mathrm{L}_1, \gamma_1)$ for sufficiently small γ_1.

Part II: L_1 *is a Global* l_p *Subspace in* $\mathrm{G}(D, d)$

We will first show that for all $\mathrm{L} \in \mathrm{G}(D, d) \setminus \mathrm{B_G}(\mathrm{L}_1, \gamma_1)$ and any fixed $p \leq 1$, there exists some $\gamma_7 > 0$ such that

$$e_{l_p}(\mathcal{X}, \mathrm{L}) - e_{l_p}(\mathcal{X}, \mathrm{L}_1) > \gamma_7 N, \quad \text{w.o.p.} \tag{42}$$

Indeed, we first conclude from Lemma 3.1 that

$$\begin{aligned} &E_\mu\left(e_{l_p}(\mathbf{x}, \mathrm{L})\right) - E_\mu\left(e_{l_p}(\mathbf{x}, \mathrm{L}_1)\right) > \alpha_0\left(E_{\mu_0}\left(e_{l_p}(\mathbf{x}, \mathrm{L})\right) - E_{\mu_0}\left(e_{l_p}(\mathbf{x}, \mathrm{L}_1)\right)\right) \\ &+ \alpha_1\left(E_{\mu_1}\left(e_{l_p}(\mathbf{x}, \mathrm{L})\right) - E_{\mu_1}\left(e_{l_p}(\mathbf{x}, \mathrm{L}_1)\right)\right) \geq \frac{\alpha_1(1-\mu_1(\{\mathbf{0}\}))2^{p-1}\gamma_1^p}{(\pi\sqrt{d})^p\left(\psi_{\mu_1}^{-1}\left(\frac{1+\mu_1(\{\mathbf{0}\})}{2}\right)\right)^p}. \end{aligned} \tag{43}$$

Setting $\gamma_7 = \frac{\alpha_1(1-\mu_1(\{\mathbf{0}\}))2^p\gamma_1^p}{(\pi\sqrt{d})^p\left(\psi_{\mu_1}^{-1}\left(\frac{1+\mu_1(\{\mathbf{0}\})}{2}\right)\right)^p}$ and combining (43) with Hoeffding's inequality, we obtain (42).

Now, (42) extends for a small neighborhood of L. That is, for any $\mathrm{L} \in \mathrm{G}(D, d)$ we can find a ball $\mathrm{B_G}(\mathrm{L}, t)$ for some $t > 0$ such that w.o.p. the subspace L_1 is a better l_p subspace than any of the subspaces in that ball. By covering the compact space $\mathrm{G}(D, d) \setminus \mathrm{B_G}(\mathrm{L}_1, \gamma_1)$ with finite number of such balls we obtain that w.o.p. L_1 is the global l_p subspace in $\mathrm{G}(D, d) \setminus \mathrm{B_G}(\mathrm{L}_1, \gamma_1)$. Combining this observation with part I, we conclude that w.o.p. L_1 is the global l_p subspace in $\mathrm{G}(D, d)$.

3.4.2. Extension of the Proof to $K > 1$

Part I: L_1 *is a Global* l_p *Subspace in* $\mathrm{B_G}(\mathrm{L}_1, \gamma_1)$

We maintain the same notation of Section 3.4.1, especially for similar constants. We will show in this part that w.o.p. L_1 is a global l_p subspace in the ball $\mathrm{B_G}(\mathrm{L}_1, \gamma_1)$, where γ_1 is a sufficiently small constant different than the one of Section 3.4.1.

In order to do so, we arbitrarily fix $\hat{\mathrm{L}} \in \mathrm{G}(D, d)$ such that $\text{dist}_{\mathrm{G}}(\hat{\mathrm{L}}, \mathrm{L}_1) = 1$ (so that $\mathbf{C} \in \mathrm{NS}_+(d)$) and parameterize a geodesic line from L_1 to $\hat{\mathrm{L}}$ by a function L: [0,1]

$\to \mathrm{G}(D,d)$, where $\mathrm{L}(0)=\mathrm{L}_1$ and $\mathrm{L}(1)=\hat{\mathrm{L}}$. We will then estimate the probability that for any such $\hat{\mathrm{L}}$ the function $e_{l_p}(\mathcal{X},\mathrm{L}(t))$: [0,1] $\to \mathbb{R}$ has a positive derivative at any $t\in(0,\gamma_1)$, that is

$$\frac{\mathrm{d}}{\mathrm{d}t^p}\left(\frac{\sum_{\mathbf{x}\in\mathcal{X}}\mathrm{dist}(\mathbf{x},\mathrm{L}(t))^p}{N}\right)>0 \quad \text{for all } t\in(0,\gamma_1). \tag{44}$$

First of all, we prove that the LHS of (44) is larger than some constant $\gamma_2>0$ at $t=0$ w.o.p., that is:

$$\frac{\mathrm{d}}{\mathrm{d}t^p}\left(\frac{\sum_{\mathbf{x}\in\mathcal{X}}\mathrm{dist}(\mathbf{x},\mathrm{L}(t))^p}{N}\right)\Bigg|_{t=0}>\gamma_2 \quad \text{w.o.p.} \tag{45}$$

When $0<p<1$, it follows from (26) and Hoeffding's inequality that (45) is valid w.p. $1-\exp(-2N\gamma_2^2)$ for $\gamma_2=\alpha_1\, E_{\mu_0}(\|\mathbf{C}_0\mathbf{V}_0P_{\mathrm{L}_1}(\mathbf{x})\|^p)/2$. When $p=1$, it follows from (6) that this probability is the same as the probability of the event

$$\frac{\sum_{\mathbf{x}\in\mathcal{X}_1}||\mathbf{C}\mathbf{V}P_{\mathrm{L}_1}(\mathbf{x})||-||\mathbf{C}\mathbf{V}\mathbf{B}_{\mathrm{L}_1,\mathcal{X}\setminus\mathcal{X}_1}||}{N}>\gamma_2 \tag{46}$$
$$\forall\mathbf{C}\in\mathrm{NS}_+(d) \text{ and } \mathbf{V}\in\mathrm{O}(d).$$

Applying the spherical symmetry of μ_0, we have that for all $\mathbf{C}\in\mathrm{NS}_+(d)$ and $\mathbf{V}\in\mathrm{O}(d)$:

$$\begin{aligned}
&||\mathbf{C}\mathbf{V}\mathbf{B}_{\mathrm{L}_1,\mathcal{X}\setminus\mathcal{X}_1}||_*=||\mathbf{C}\mathbf{V}\sum_{\mathbf{x}\in\mathcal{X}\setminus\mathcal{X}_1}P_{\mathrm{L}_1}(\mathbf{x})P_{\mathrm{L}_1}^{\perp}(\mathbf{x})^T/\mathrm{dist}(\mathbf{x},\mathrm{L}_1)||_*\\
&=||\mathbf{C}\mathbf{V}\sum_{\mathbf{x}\in\mathcal{X}\setminus\{\mathcal{X}_1\cup\mathcal{X}_0\}}P_{\mathrm{L}_1}(\mathbf{x})P_{\mathrm{L}_1}^{\perp}(\mathbf{x})^T/\mathrm{dist}(\mathbf{x},\mathrm{L}_1)||_*\\
&\le\sum_{\mathbf{x}\in\mathcal{X}\setminus\{\mathcal{X}_1\cup\mathcal{X}_0\}}||\mathbf{C}\mathbf{V}P_{\mathrm{L}_1}(\mathbf{x})P_{\mathrm{L}_1}^{\perp}(\mathbf{x})^T/\|P_{\mathrm{L}_1}^{\perp}(\mathbf{x})\|\,||_*\le\sum_{\mathbf{x}\in\mathcal{X}\setminus\{\mathcal{X}_1\cup\mathcal{X}_0\}}||\mathbf{C}\mathbf{V}P_{\mathrm{L}_1}(\mathbf{x})||.
\end{aligned}$$

Consequently, in order to estimate the probability of (46) it is sufficient to estimate the probability that

$$\frac{\sum_{\mathbf{x}\in\mathcal{X}_1}||\mathbf{C}\mathbf{V}P_{\mathrm{L}_1}(\mathbf{x})||-\sum_{\mathbf{x}\in\mathcal{X}\setminus\{\mathcal{X}_1\cup\mathcal{X}_0\}}||\mathbf{C}\mathbf{V}P_{\mathrm{L}_1}(\mathbf{x})||}{N}>\gamma_2 \tag{47}$$
$$\forall\mathbf{C}\in\mathrm{NS}_+(d) \text{ and } \mathbf{V}\in\mathrm{O}(d).$$

We arbitrarily fix $\mathbf{C}_0\in\mathrm{NS}_+(d)$, $\mathbf{V}_0\in\mathrm{O}(d)$ and verify (47) by Hoeffding's inequality in the following way. We define the random variable $J(\mathbf{x})=(2I(\mathbf{x}\in\mathcal{X}_1)-1)||\mathbf{C}_0\mathbf{V}_0P_{\mathrm{L}_1}(\mathbf{x})||$ and using the spherical symmetry of $\{\mu_i\}_{i=1}^K$, we have

$$E_\mu(J(\mathbf{x}))=E_{\mu^N}\left(\frac{\sum_{\mathbf{x}\in\mathcal{X}_1}||\mathbf{C}_0\mathbf{V}_0P_{\mathrm{L}_1}(\mathbf{x})||-\sum_{\mathbf{x}\in\mathcal{X}\setminus\{\mathcal{X}_1\cup\mathcal{X}_0\}}||\mathbf{C}_0\mathbf{V}_0P_{\mathrm{L}_1}(\mathbf{x})||}{N}\right) \tag{48}$$

$$
\begin{aligned}
&= \alpha_1 E_{\mu_1} ||\mathbf{C}_0\mathbf{V}_0 P_{\mathrm{L}_1}(\mathbf{x})|| - \sum_{j=2}^{K} \alpha_j E_{\mu_j} ||\mathbf{C}_0\mathbf{V}_0 P_{\mathrm{L}_1}(\mathbf{x})|| \\
&\geq \alpha_1 E_{\mu_1} ||\mathbf{C}_0\mathbf{V}_0 P_{\mathrm{L}_1}(\mathbf{x})|| - \sum_{j=2}^{K} \alpha_j E_{\mu_1} ||\mathbf{C}_0\mathbf{V}_0 P_{\mathrm{L}_1}(\mathbf{x})|| \\
&= \beta_0 E_{\mu_1} ||\mathbf{C}_0\mathbf{V}_0 P_{\mathrm{L}_1}(\mathbf{x})||,
\end{aligned}
$$

where $\beta_0 = \alpha_1 - \sum_{j=2}^{K} \alpha_j$.

Now, let $\gamma_2 := \beta_0 E_{\mu_1} ||\mathbf{C}_0\mathbf{V}_0 P_{\mathrm{L}_1}(\mathbf{x})||/4$. We note that the random variable $J(\mathbf{x})$ has expectation larger than $4\gamma_2$ and it takes values in $[-1, 1]$; thus by Hoeffding's inequality:

$$
\begin{aligned}
&\frac{\sum_{\mathbf{x}\in\mathcal{X}_1} ||\mathbf{C}_0\mathbf{V}_0 P_{\mathrm{L}_1}(\mathbf{x})|| - \sum_{\mathbf{x}\in\mathcal{X}\setminus\mathcal{X}_1} ||\mathbf{C}_0\mathbf{V}_0 P_{\mathrm{L}_1}(\mathbf{x})||}{N} > 2\gamma_2 \qquad (49) \\
&\text{w.p.} \geq 1 - \exp(-2N\gamma_2^2).
\end{aligned}
$$

We have thus proved that (45) is valid with sufficiently high probability for fixed matrices $\mathbf{C}_0 \in \mathrm{NS}_+(d)$ and $\mathbf{V}_0 \in \mathrm{O}(d)$. Next we estimate the probability of (45) for all matrices $\mathbf{C} \in \mathrm{NS}_+(d)$ and $\mathbf{V} \in \mathrm{O}(d)$, when restricted to a ball with sufficiently small radius. We let

$$
\mathrm{dist}_{(\mathrm{NS}_+(d),\mathrm{O}(d))}((\mathbf{C}_1, \mathbf{V}_1), (\mathbf{C}_2, \mathbf{V}_2)) := \max(||\mathbf{C}_1 - \mathbf{C}_2||_2, ||\mathbf{V}_1 - \mathbf{V}_2||_2) \qquad (50)
$$

and note that whenever $\mathrm{dist}_{(\mathrm{NS}_+(d),\mathrm{O}(d))}((\mathbf{C}_1, \mathbf{V}_1), (\mathbf{C}_2, \mathbf{V}_2)) < \gamma_2/2$ and $\mathbf{x} \in \mathrm{B}(\mathbf{0}, 1)$ we have that

$$
\begin{aligned}
&||\mathbf{C}_1\mathbf{V}_1 P_{\mathrm{L}_1}(\mathbf{x})|| - ||\mathbf{C}_2\mathbf{V}_2 P_{\mathrm{L}_1}(\mathbf{x})|| \\
&= (||\mathbf{C}_1\mathbf{V}_1 P_{\mathrm{L}_1}(\mathbf{x})|| - ||\mathbf{C}_2\mathbf{V}_1 P_{\mathrm{L}_1}(\mathbf{x})||) + (||\mathbf{C}_2\mathbf{V}_1 P_{\mathrm{L}_1}(\mathbf{x})|| - ||\mathbf{C}_2\mathbf{V}_2 P_{\mathrm{L}_1}(\mathbf{x})||) \\
&\leq ||\mathbf{C}_1 - \mathbf{C}_2||_2 + ||\mathbf{C}_2||_2 ||\mathbf{V}_1 - \mathbf{V}_2||_2 \leq \gamma_2. \qquad (51)
\end{aligned}
$$

Combining (49) and (51) we obtain that for any ball in $\mathrm{G}(D, d)$ of radius $\gamma_2/2$ and center $(\mathbf{C}_0, \mathbf{V}_0)$:

$$
\frac{\sum_{\mathbf{x}\in\mathcal{X}_1} ||\mathbf{C}\mathbf{V} P_{\mathrm{L}_1}(\mathbf{x})|| - \sum_{\mathbf{x}\in\mathcal{X}\setminus\mathcal{X}_1} ||\mathbf{C}\mathbf{V} P_{\mathrm{L}_1}(\mathbf{x})||}{N} > \gamma_2 \ \text{ w.p.} \geq 1 - \exp(-2N\gamma_2^2). \qquad (52)
$$

We easily extend (52) for all pairs of matrices $(\mathbf{C}, \mathbf{V})$ in the compact space $\mathrm{NS}_+(d) \times \mathrm{O}(d)$ (with the distance specified in (50)). Indeed, it follows from [25] together with some basic estimates that the latter space can be covered by $C_1^{d(d+1)/2}/(\gamma_2/2)^{d(d+1)/2}$ balls of radius $\gamma_2/2$. Therefore,

(45) is valid for any $\mathbf{C} \in \mathrm{NS}_+(d)$ and $\mathbf{V} \in \mathrm{O}(d)$

$$
\text{w.p. } 1 - C_1^{2d} \exp(-2N\gamma_2^2)/(\gamma_2/2)^{2d-1}. \qquad (53)
$$

Equation (44) follows w.o.p. from (45) in exactly the same way of deriving (35) from (36) and (37). We remark that (37), which is deterministic, easily extends to the

current case. While we did not estimate the overwhelming probability for (35), it is easy to show that in the current case, (45) implies (44) w.p. $1-\exp(-N\gamma_8)/\gamma_8$. Carrying this analysis, one notices that both γ_1 and γ_8 depend on d, K, α_0, α_1, μ_0, μ_1, p and $\min_{2\le i\le K}(\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1,\mathrm{L}_i))$. Combining this with (53), we obtain that

$$\mathrm{L}_1 \text{ is a global } l_p \text{ subspace in } \mathrm{B}_{\mathrm{G}}(\mathrm{L}_1,\gamma_1) \quad \text{w.p. } 1-C_1^{2d}\exp(-2N\gamma_2^2)/(\gamma_2/2)^{2d-1}-\exp(-N\gamma_4)/\gamma_4. \tag{54}$$

Part II: L_1 *is a Global* l_p *Subspace in* $\mathrm{G}(D,d)$

We will first prove that L_1 is a global l_p subspace w.o.p. in $\mathrm{G}(D,d)\setminus \mathrm{B}_{\mathrm{G}}(\mathrm{L}_1,\gamma_1)$. Applying Lemma 3.3 we obtain that for all $2\le i\le K$:

$$E_{\mu_1}\left(\mathrm{dist}(\mathbf{x},\mathrm{L})^p-\mathrm{dist}(\mathbf{x},\mathrm{L}_1)^p\right)+E_{\mu_i}(\mathrm{dist}(\mathbf{x},\mathrm{L})^p-\mathrm{dist}(\mathbf{x},\mathrm{L}_1)^p)\ge 0. \tag{55}$$

Further application of Lemma 3.1 with $\mathrm{L}\in \mathrm{G}(D,d)\setminus \mathrm{B}_{\mathrm{G}}(\mathrm{L}_1,\gamma_1)$ results in the inequality:

$$E_{\mu_1}(\mathrm{dist}(\mathbf{x},\mathrm{L}))>\frac{(1-\mu_1(\{\mathbf{0}\}))\cdot 2^{p-1}\cdot\gamma_1^p}{\left(\pi\sqrt{d}\right)^p\cdot\left(\psi_{\mu_1}^{-1}\left((1+\mu_1(\{\mathbf{0}\}))/2\right)\right)^p}. \tag{56}$$

Now, combining (55) and (56) we have that

$$\begin{aligned}
&E_{\mu}(\mathrm{dist}(\mathbf{x},\mathrm{L})^p-\mathrm{dist}(\mathbf{x},\mathrm{L}_1)^p)\\
=&\sum_{i=2}^{K}\alpha_i((E_{\mu_1}(\mathrm{dist}(\mathbf{x},\mathrm{L})^p-\mathrm{dist}(\mathbf{x},\mathrm{L}_1)^p)+E_{\mu_i}(\mathrm{dist}(\mathbf{x},\mathrm{L})^p-\mathrm{dist}(\mathbf{x},\mathrm{L}_1)^p))\\
&+\beta_0 E_{\mu_1}(\mathrm{dist}(\mathbf{x},\mathrm{L})^p-\mathrm{dist}(\mathbf{x},\mathrm{L}_1)^p)\\
\ge&\frac{\beta_0\cdot(1-\mu_1(\{\mathbf{0}\}))\cdot 2^{p-1}\cdot\gamma_1^p}{\left(\pi\sqrt{d}\right)^p\cdot\left(\psi_{\mu_1}^{-1}\left((1+\mu_1(\{\mathbf{0}\}))/2\right)\right)^p},
\end{aligned}$$

where γ_9 depends on d, K, μ_0, μ_1, α_0, α_1 and $\min_{2\le i\le K}(\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1,\mathrm{L}_i))$. Noting further that $\mathrm{dist}(\mathbf{x},\mathrm{L})-\mathrm{dist}(\mathbf{x},\mathrm{L}_1)$ takes bounded values and applying Hoeffding's inequality we obtain that for any $\mathrm{L}\in\mathrm{G}(D,d)\setminus\mathrm{B}_{\mathrm{G}}(\mathrm{L}_1,\gamma_1)$:

$$e_{l_p}(\mathcal{X},\mathrm{L})-e_{l_p}(\mathcal{X},\mathrm{L}_1)>\gamma_9 N/2 \text{ w.p. } \ge 1-\exp(-N\gamma_9^2/8). \tag{57}$$

By Lemma 3.2 we have that for any $\mathrm{L}'\in\mathrm{G}(D,d)$ satisfying $\mathrm{dist}_{\mathrm{G}}(\mathrm{L},\mathrm{L}')<(\gamma_9/4)^{1/p}$ and any $\mathbf{x}\in\mathrm{B}(\mathbf{0},1)$:

$$|\mathrm{dist}(\mathbf{x},\mathrm{L}')^p-\mathrm{dist}(\mathbf{x},\mathrm{L})^p|<\gamma_9/4.$$

Consequently, for any $\mathrm{L}\in\mathrm{G}(D,d)\setminus\mathrm{B}_{\mathrm{G}}(\mathrm{L}_1,\gamma_1)$ and all $\mathrm{L}'\in\mathrm{B}_{\mathrm{G}}(\mathrm{L},(\gamma_9/4)^{1/p})$:

$$e_{l_p}(\mathcal{X},\mathrm{L}')-e_{l_p}(\mathcal{X},\mathrm{L}_1)>0 \text{ w.p. } \ge 1-\exp(-N\gamma_9^2/8). \tag{58}$$

Following [24, Remark 8.4] we can cover $\mathrm{G}(D,d)\backslash \mathrm{B}_{\mathrm{G}}(\mathrm{L}_1,\gamma_1)$ by $C_2^{d(D-d)}/\gamma_9^{d(D-d)/p}$ balls of radius $(\gamma_9/4)^{1/p}$. Now, for each such ball we have that (57) is valid for its center w.p. $1-\exp(-N\gamma_9^2/8)$ and consequently (58) is valid for subspaces in that ball with the same probability. We thus conclude that (58) is valid for all $\mathrm{L}' \in \mathrm{G}(D,d)\setminus \mathrm{B}_{\mathrm{G}}(\mathrm{L}_1,\gamma_1)$ w.p. $1-\exp(-N\gamma_9^2/8)C_2^{d(D-d)/p}/\gamma_9^{d(D-d)}$. Combining this with (54), we obtain that the probability that L_1 is a global l_1 subspace in $\mathrm{G}(D,d)$ is

$$1-C_1^{2d}\exp(-2N\gamma_2^2)/(\gamma_2/2)^{2d-1}-\exp(-N\gamma_4)/\gamma_4-\exp(-N\gamma_9^2/8)C_2^{d(D-d)}/\gamma_9^{d(D-d)/p},$$

or equivalently, $1-C\exp(-N/C)$ for some C depending on D, d, K, μ_0, μ_1, α_0, α_1, p and $\min_{2\le i\le K}(\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1,\mathrm{L}_i))$.

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 $0<p\le 1$ 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 $0<p\le 1$, i.e., prove that the global minimum of $e_{l_p}(\mathcal{X},\mathrm{L})$ is in $\mathrm{B}_{\mathrm{G}}(\mathrm{L}_1,f)$ w.o.p., we only need to show that there exists a constant $\gamma_1>0$ such that for any $\mathrm{L}\notin \mathrm{B}_{\mathrm{G}}(\mathrm{L}_1,f)$:

$$E_{\mu_\epsilon}(e_{l_p}(\mathbf{x},\mathrm{L}))>E_{\mu_\epsilon}(e_{l_p}(\mathbf{x},\mathrm{L}_1))+\gamma_1. \tag{59}$$

Indeed, we cover the compact space $\mathrm{G}(D,d)\setminus \mathrm{B}_{\mathrm{G}}(\mathrm{L}_1,f)$ by small balls with radius $\gamma_1/2$. Then by using (59) and Hoeffding's inequality, we obtain that $e_{l_p}(\mathcal{X},\mathrm{L})>e_{l_p}(\mathcal{X},\mathrm{L}_1)$ for any L in each such ball w.o.p. Therefore, $e_{l_p}(\mathcal{X},\mathrm{L})>e_{l_p}(\mathcal{X},\mathrm{L}_1)$ for $\mathrm{L}\in \mathrm{G}(D,d)\setminus \mathrm{B}_{\mathrm{G}}(\mathrm{L}_1,f)$ w.o.p. Equivalently, $\mathrm{G}(D,d)\setminus \mathrm{B}_{\mathrm{G}}(\mathrm{L}_1,f)$ does not contain the global minimum of $e_{l_p}(\mathcal{X},\mathrm{L})$ w.o.p.

For $i=1,\ldots,K$, let $\tilde{\mu}_{i,\epsilon}$ be the measure obtained by projecting $\mu_{i,\epsilon}$ onto its corresponding subspace L_i (that is, for any set $E\subseteq \mathrm{B}(\mathbf{0},1)\cap \mathrm{L}_i$: $\tilde{\mu}_{i,\epsilon}(E)=\mu_{i,\epsilon}(P_{\mathrm{L}_i}^{-1}(E))$). We also let $\tilde{\mu}_\epsilon:=\alpha_0\mu_0+\sum_{i=1}^K \alpha_i\tilde{\mu}_{i,\epsilon}(E)$. By the triangle inequality and the definition of μ_ϵ:

$$|E_{\mu_\epsilon}(e_{l_p}(\mathbf{x},\mathrm{L}))-E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x},\mathrm{L}))|<\epsilon^p.$$

Hence, in order to prove (59) and thus Theorem 1.2 for $p\le 1$, the following equation is sufficient:

$$E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x},\mathrm{L}))>E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x},\mathrm{L}_1))+\gamma_1+2\epsilon^p,\ \text{ for any } \mathrm{L}\in \mathrm{G}(D,d)\setminus \mathrm{B}_{\mathrm{G}}(\mathrm{L}_1,f). \tag{60}$$

We can similarly reduce Theorem 1.2 when $K=1$ and $p>1$ to the following condition:

$$E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x},\mathrm{L}))>E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x},\mathrm{L}_1))+\gamma_1+2p\epsilon,\ \text{ for any } \mathrm{L}\in \mathrm{G}(D,d)\setminus \mathrm{B}_{\mathrm{G}}(\mathrm{L}_1,f). \tag{61}$$

This reduction follows the same arguments above combined with the following observation: For any $\mathbf{x}_1, \mathbf{x}_2 \in \mathrm{B}(\mathbf{0}, 1)$ with $\mathrm{dist}(\mathbf{x}_1, \mathbf{x}_2) < \eta < 1$ and any $\tilde{\mathrm{L}}_1, \tilde{\mathrm{L}}_2 \in \mathrm{G}(D, d)$ with $\mathrm{dist}_{\mathrm{G}}(\tilde{\mathrm{L}}_1, \tilde{\mathrm{L}}_2) < \eta$:

$$\mathrm{dist}(\mathbf{x}_1, \tilde{\mathrm{L}}_1)^p - \mathrm{dist}(\mathbf{x}_2, \tilde{\mathrm{L}}_1)^p < 1 - (1-\eta)^p < p\eta, \tag{62}$$

and

$$\mathrm{dist}(\mathbf{x}_1, \tilde{\mathrm{L}}_1)^p - \mathrm{dist}(\mathbf{x}_1, \tilde{\mathrm{L}}_2)^p < 1 - (1-\eta)^p < p\eta. \tag{63}$$

When $p = 1$, (62) follows from the triangle inequality and (63) follows from Lemma 3.2, whereas both equations extend to $p > 1$ by the following property of the p-th power: if $0 \leq y_1, y_2 \leq 1$, $y_1 - y_2 < \eta$ and $p > 1$, then $y_1^p - y_2^p < 1 - (1-\eta)^p$.

3.5.2. *Proof of* (60) *and* (61) *and Conclusion of Theorem 1.2*

We arbitrarily fix $\mathrm{L} \in \mathrm{G}(D, d) \setminus \mathrm{B}_{\mathrm{G}}(\mathrm{L}_1, f)$. We assume first that $0 < p \leq 1$ and apply Lemma 3.3 to obtain that

$$\begin{aligned} &E_{\tilde{\mu}_\epsilon - (\alpha_1 - \sum_{i=2}^K \alpha_i)\tilde{\mu}_{1,\epsilon}} e_{l_p}(\mathbf{x}, \mathrm{L}) - E_{\tilde{\mu}_\epsilon - (\alpha_1 - \sum_{i=2}^K \alpha_i)\tilde{\mu}_{1,\epsilon}} e_{l_p}(\mathbf{x}, \mathrm{L}_1) \\ &= \sum_{i=2}^K \alpha_i \left(E_{\tilde{\mu}_{1,\epsilon} + \tilde{\mu}_{i,\epsilon}} e_{l_p}(\mathbf{x}, \mathrm{L}) - E_{\tilde{\mu}_{1,\epsilon} + \tilde{\mu}_{i,\epsilon}} e_{l_p}(\mathbf{x}, \mathrm{L}_1) \right) \geq 0. \end{aligned}$$

Consequently, we prove (60) with $\gamma_1 := 2\epsilon^p$ as follows:

$$\begin{aligned} E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x}, \mathrm{L})) - E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x}, \mathrm{L}_1)) &\geq \left(\alpha_1 - \sum_{i=2}^K \alpha_i \right) E_{\tilde{\mu}_{1,\epsilon}}(e_{l_p}(\mathbf{x}, \mathrm{L})) \qquad (64) \\ \geq \frac{\left(\alpha_1 - \sum_{i=2}^K \alpha_i \right) (1 - \mu_1(\{\mathbf{0}\}))\, 2^{p-1} f^p}{(\pi\sqrt{d})^p \left(\psi_{\mu_1}^{-1} \left(\frac{1+\mu_1(\{\mathbf{0}\})}{2} \right) \right)^p} &= 4\epsilon^p, \end{aligned}$$

where the second inequality applies Lemma 3.1.

Equation (61) (with $p > 1$) follows from the same argument of (64), where ϵ^p is now replaced by $p\epsilon$.

3.5.3. *Remark on The Size of* ϵ

If $0 < p \leq 1$ and

$$\epsilon > \frac{\left(\alpha_1 - \sum_{i=2}^K \alpha_i \right)^{\frac{1}{p}} (1 - \mu_1(\{\mathbf{0}\}))^{\frac{1}{p}}}{2^{\frac{3}{p}}\, \psi_{\mu_1}^{-1}\left(\frac{1+\mu_1(\{\mathbf{0}\})}{2} \right)} \tag{65}$$

or $p > 1$, $K = 1$ and

$$\epsilon > \frac{\alpha_1 (1 - \mu_1(\{\mathbf{0}\}))\, 2^{p-3}}{p\, \pi^p d^{\frac{p}{2}}\, \psi_{\mu_1}^{-1}\left(\frac{1+\mu_1(\{\mathbf{0}\})}{2} \right)^p}, \tag{66}$$

then $f > \frac{\pi\sqrt{d}}{2}$, which implies that $\mathrm{B}_{\mathrm{G}}(\mathrm{L}_1, f) = \mathrm{G}(D, d)$ (since all principle angles are at most $\pi/2$). It thus makes sense to restrict the level of noise to be at least lower than the right hand sides of (65) or (66).

3.6. Proof of Theorem 1.3: Symmetry Arguments

3.6.1. First Reduction of Theorem 1.3

We use the same notation of Section 3.5.1, in particular, $\tilde{\mu}_\epsilon$. Theorem 1.3 states that the global l_p subspace is not in $\mathrm{B_G}(\mathrm{L}_1, \kappa_0)$ w.o.p. for almost every $\{\mathrm{L}_i\}_{i=1}^K \in \mathrm{G}(D,d)^K$. We claim that it reduces to the following simple equation:

$$\gamma_{D,d}^K \left(\{\mathrm{L}_i\}_{i=1}^K \subset \mathrm{G}(D,d) : \mathrm{L}_1 = \mathrm{argmin}_{\mathrm{L}} E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x}, \mathrm{L})) \right) = 0. \tag{67}$$

Indeed, if (67) is not satisfied, then for any K d-subspaces $\{\mathrm{L}_i\}_{i=1}^K$ in a subset of $\mathrm{G}(D,d)^K$ with nonzero $\gamma_{D,d}^K$ measure there exists $\mathrm{L}_0 \in \mathrm{G}(D,d)$ such that

$$\gamma_1 := E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x}, \mathrm{L}_1)) - E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x}, \mathrm{L}_0)) > 0.$$

Letting $\delta_0 = \kappa_0 = \gamma_1 / 4p\epsilon$, we obtain from (62) and (63) that for any $\mathrm{L}^* \in \mathrm{B_G}(\mathrm{L}_1, \kappa_0)$:

$$\begin{aligned} &E_{\mu_\epsilon}(e_{l_p}(\mathbf{x}, \mathrm{L}^*)) - E_{\mu_\epsilon}(e_{l_p}(\mathbf{x}, \mathrm{L}_0)) > E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x}, \mathrm{L}^*)) - E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x}, \mathrm{L}_0)) - 2\delta_0 p \\ &> E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x}, \mathrm{L}_1)) - E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x}, \mathrm{L}_0)) - 2\delta_0 p - \kappa_0 p = \frac{\gamma_1}{4}. \end{aligned}$$

Therefore, by Hoeffding's inequality:

$$e_{l_p}(\mathcal{X}, \mathrm{L}^*) - e_{l_p}(\mathcal{X}, \mathrm{L}_0) > \frac{\gamma_1 N}{8} \text{ w.o.p.}$$

In order to have

$$e_{l_p}(\mathcal{X}, \mathrm{L}^*) - e_{l_p}(\mathcal{X}, \mathrm{L}_0) > 0 \quad \text{for all } \mathrm{L}^* \in \mathrm{B_G}(\mathrm{L}_1, \kappa_0) \text{ w.o.p.},$$

we cover $\mathrm{B_G}(\mathrm{L}_1, \kappa_0)$ by small balls with radius $\gamma_1/16$, so that $e_{l_p}(\mathcal{X}, \mathrm{L}) > e_{l_p}(\mathcal{X}, \mathrm{L}_0)$ for all L in each such ball w.o.p. Therefore, $e_{l_p}(\mathcal{X}, \mathrm{L}) > e_{l_p}(\mathcal{X}, \mathrm{L}_0)$ for all $\mathrm{L} \in \mathrm{B_G}(\mathrm{L}_1, \kappa_0)$ w.o.p. Equivalently, $\mathrm{B_G}(\mathrm{L}_1, \kappa_0)$ will not contain the global minimum of $e_{l_p}(\mathcal{X}, \mathrm{L})$ w.o.p. This contradicts Theorem 1.3 and therefore (67) implies this theorem.

3.6.2. Second Reduction of Theorem 1.3

We define the operator

$$\mathbf{D}_{\mathrm{L},\mathbf{x},p} = P_{\mathrm{L}}(\mathbf{x}) P_{\mathrm{L}}^{\perp}(\mathbf{x})^T \mathrm{dist}(\mathbf{x}, \mathrm{L})^{(p-2)} \tag{68}$$

and the function

$$h(\mathrm{L}_1, \mathrm{L}_i) = E_{\tilde{\mu}_{i,\epsilon}}(\mathbf{D}_{\mathrm{L}_1,\mathbf{x},p}), \;\; 2 \le i \le K.$$

In view of Proposition 2.2, (67) follows from the condition:

$$\gamma_{D,d}^K \left(\{\mathrm{L}_i\}_{i=1}^K \subset \mathrm{G}(D,d) : E_{\tilde{\mu}_\epsilon}\left(\mathbf{D}_{\mathrm{L}_1,\mathbf{x},p}\right) = 0 \right) = 0, \tag{69}$$

which we rewrite as follows:

$$\gamma_{D,d}^K \left(\{\mathrm{L}_i\}_{i=1}^K \subset \mathrm{G}(D,d) : E_{\tilde{\mu}_\epsilon}\left(\mathbf{D}_{\mathrm{L}_1,\mathbf{x},p}\right) = 0 \right)$$

$$=\gamma_{D,d}^K\left(\{\mathrm{L}_i\}_{i=1}^K\subset \mathrm{G}(D,d): E_{\sum_{i=2}^K \alpha_i\tilde{\mu}_{i,\epsilon}}\left(\mathbf{D}_{\mathrm{L}_1,\mathbf{x},p}\right)=0\right)$$
$$=\gamma_{D,d}^K\left(\{\mathrm{L}_i\}_{i=1}^K\subset \mathrm{G}(D,d): \sum_{i=2}^K \alpha_i\, h(\mathrm{L}_1,\mathrm{L}_i)=0\right)=0. \tag{70}$$

Since $\{\mathrm{L}_i\}_{i=1}^K$ are identically and independently distributed according to $\gamma_{D,d}$, Fubini's Theorem implies that (70) follows from the equation:

$$\gamma_{D,d}\left(\mathrm{L}_2\in \mathrm{G}(D,d): h(\mathrm{L}_1,\mathrm{L}_2)=\mathbf{C}(\mathrm{L}_1,\mathrm{L}_3,\cdots,\mathrm{L}_K)\right)=0, \tag{71}$$

where $\mathbf{C}(\mathrm{L}_1,\mathrm{L}_3,\cdots,\mathrm{L}_K)=-\sum_{i=3}^K \alpha_i\, h(\mathrm{L}_1,\mathrm{L}_i)/\alpha_2$.

3.6.3. 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 $\{\hat{\mathbf{v}}_j\}_{j=1}^d$ and $\{\mathbf{v}_j\}_{j=1}^d$ respectively and the complementary orthogonal system for L_2 w.r.t. L_1 by $\{\mathbf{u}_j\}_{j=1}^d$. Note that $h(\mathrm{L}_1,\mathrm{L}_2)$, as a function of $\mathbf{x}$, maps $\mathrm{Sp}(\{\mathbf{u}_i\}_{i=1}^d)$ to $\mathrm{Sp}(\{\mathbf{v}_i\}_{i=1}^d)$. Now, transforming $\mathbf{x}\in \mathrm{L}_2\cap \mathrm{B}(\mathbf{0},1)$ to $\{a_i\}_{i=1}^d$ in a d-dimensional unit ball by $\mathbf{x}=\sum_{i=1}^d a_i\hat{\mathbf{v}}_i$, we have that for any $1\le i_1,i_2\le d$:

$$\mathbf{v}_{i_1}^T h(\mathrm{L}_1,\mathrm{L}_2)\mathbf{u}_{i_2}=E_{\mu_2}(\mathbf{v}_{i_1}^T P_{\mathrm{L}_1}(\mathbf{x})P_{\mathrm{L}_1}^{\perp}(\mathbf{x})^T\mathbf{u}_{i_2}\mathrm{dist}(\mathbf{x},\mathrm{L}_1)^{p-2})$$
$$=\int_{\sum_{i=1}^d {a_i}^2\le 1}\cos\theta_{i_1}a_{i_1}\sin\theta_{i_2}a_{i_2}\left(\sum_{i=1}^d a_i^2\sin^2\theta_i\right)^{\frac{p-2}{2}}\mathrm{d}V,$$

where $\mathrm{d}V$ denotes the scaled volume element on the d-dimensional ball $\sum_{i=1}^d {a_i}^2\le 1$.

For simplicity, we will assume till the rest of the proof that μ_2 is a uniform distribution on $\mathrm{B}(\mathbf{0},1)\cap \mathrm{L}_2$. Nevertheless, the proof can be easily generalized to any spherically symmetric distribution on L_2 with bounded support. When $i_1\neq i_2$, the function

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

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

$$\mathbf{v}_{i_1}^T h(\mathrm{L}_1,\mathrm{L}_2)\mathbf{u}_{i_2}=\int_{\sum_{i=1}^d {a_i}^2\le 1}\cos\theta_{i_1}a_{i_1}\sin\theta_{i_2}a_{i_2}\left(\sum_{i=1}^d a_i^2\sin^2\theta_i\right)^{\frac{p-2}{2}}\mathrm{d}V=0.$$

Therefore, when we form $\mathbf{V}$ and $\mathbf{U}$ as in (17), the $d\times d$ matrix $\mathbf{V}h(\mathrm{L}_1,\mathrm{L}_2)\mathbf{U}^T$ is diagonal with the elements

$$\int_{\sum_{i=1}^d {a_i}^2\le 1}\cos\theta_j\sin\theta_j a_j^2\left(\sum_{i=1}^d a_i^2\sin^2\theta_i\right)^{\frac{p-2}{2}}\mathrm{d}V,\quad j=1,\cdots,d.$$

Notice that $\mathbf{V}h(\mathrm{L}_1,\mathrm{L}_2) = h(\mathrm{L}_1,\mathrm{L}_2) = h(\mathrm{L}_1,\mathrm{L}_2)\mathbf{U}^T$, $h(\mathrm{L}_1,\mathrm{L}_2)$ has the following d singular values:

$$\lambda_j(h(\mathrm{L}_1,\mathrm{L}_2)) = \int_{\sum_{i=1}^d a_i 2\leq 1} \cos\theta_j \sin\theta_j a_j 2 \left(\sum_{i=1}^d a_i 2\sin 2\theta_i\right)^{\frac{p-2}{2}}, \quad j=1,\cdots,d.$$

We arbitrarily fix $\mathrm{L}_1, \mathrm{L}_3, \mathrm{L}_4, \cdots, \mathrm{L}_K$ and denote the singular values of $\mathbf{C} \equiv \mathbf{C}(\mathrm{L}_1, \mathrm{L}_3, \mathrm{L}_4, \cdots, \mathrm{L}_K)$ by $\{\sigma_i\}_{i=1}^D$ and observe that (71) is implied by the following equation:

$$\gamma_{D,d}\left(\mathrm{L}_2 \in \mathrm{G}(D,d) : \lambda_1(h(\mathrm{L}_1,\mathrm{L}_2)) \in \{\sigma_i\}_{i=1}^D\right) = 0, \tag{72}$$

which we express as:

$$\gamma_{D,d}\left(\int_{\sum_{i=1}^d {a_1}^2\leq 1} \cos\theta_1 \sin\theta_1 a_1^2 \left(\sum_{i=1}^d a_i^2 \sin^2\theta_i\right)^{\frac{p-2}{2}} \mathrm{d}V \in \{\sigma_i\}_{i=1}^D\right) = 0. \tag{73}$$

3.6.4. *Proof of* (73) *and Conclusion of Theorem 1.3*

We first conclude (73) when $p = 2$. In this case

$$\int_{\sum_{i=1}^d {a_1}^2\leq 1} \cos\theta_1 \sin\theta_1 a_1^2 \left(\sum_{i=1}^d a_i^2 \sin^2\theta_i\right)^{\frac{p-2}{2}} \mathrm{d}V \equiv \int_{\sum_{i=1}^d {a_1}^2\leq 1} \cos\theta_1 \sin\theta_1 a_1^2 \,\mathrm{d}V \tag{74}$$

is a monotone function of θ_1 on $[0,\pi/4]$ as well as $[\pi/4,\pi/2]$. That is, the requirement that $\lambda_1(h(\mathrm{L}_1,\mathrm{L}_2)) \in \{\sigma_i\}_{i=1}^D$ can occur only at discrete values of θ_1 and consequently has $\gamma_{D,d}$ measure 0, that is, (73) (and consequently (67)) is verified in this case.

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

$$\int_{\sum_{i=1}^d {a_1}^2\leq 1} \cos\theta_1 \sin\theta_1 a_1^2 \left(\sum_{i=1}^d a_i^2 \sin^2\theta_i\right)^{\frac{p-2}{2}} \mathrm{d}V \tag{75}$$

is a monotone function of θ_d. Following a similar argument, we obtain that

$$\gamma_{D,d}\left(h(\mathrm{L}_1,\mathrm{L}_2) \in \{\sigma_i\}_{i=1}^D | \{\theta_i\}_{i=1}^{d-1}\right) = 0. \tag{76}$$

Combining (76) and Fubini's Theorem, we conclude (73).

3.6.5. Remark on the Size of δ_0 and κ_0

The above constants δ_0 and κ_0 depend on other parameters of the underlying spherically symmetric HLM model in particular the underlying subspaces $\{\mathrm{L}_i\}_{i=1}^K$. For example, in the case of $p \geq 2$ one can estimate from below both κ_0 and δ_0 by the following number:

$$\frac{\| \sum_{i=2}^d \alpha_i E_{\tilde{\mu}_{i,\epsilon}}(\mathbf{D}_{\mathrm{L}_1,\mathbf{x},p})\|_2^2}{dD2^{p+5}},$$

where $\mathbf{D}_{\mathrm{L}_1,\mathbf{x},p}$ is defined in (68) and for any $i = 1, \ldots, K$, $\tilde{\mu}_{i,\epsilon}$ is obtained by projecting $\mu_{i,\epsilon}$ onto the subspace L_i (as in Section 3.5.1).

4. Discussion

We studied the effectiveness of l_p minimization for recovering and nearly recovering the most significant subspace w.o.p. Our setting assumed identical and independent sampling from a spherically symmetric HLM measure with noise level $\epsilon \geq 0$. A restricted setting like this is necessary and indeed we described some typical cases where global l_p subspaces are different than global l_0 subspaces for all $0 < p < \infty$. Our analysis has provided some guarantees for the robustness to bounded spherically symmetric outliers of the single subspace recovery advocated in [8] as well as sequential HLM as in [30] (while using l_p minimization with $0 < p \leq 1$ in the spirit of [26, 27]). We conclude with some possible extensions and open directions.

4.1. More General Distributions

The strict spherical symmetry of the distributions $\{\mu_i\}_{i=0}^K$ in Theorems 1.1 and 1.2 can be relaxed. Indeed, one can notice that our proofs extend with weaker bounds to *approximately* spherically symmetric distributions (with bounded support). By approximate spherically symmetric we mean that it is absolutely continuous with respect to a spherically symmetric distribution and with derivative bounded away from 0 and ∞. This weaker assumption requires an upper bound on α_0, i.e.,

$$\alpha_0 < C_\star(\mu_0, \mu_1), \tag{77}$$

and the condition

$$C_1(\mu_1)\alpha_1 > \sum_{i=2}^K C_i(\mu_i)\alpha_i + C_0(\mu_0)\alpha_0 \tag{78}$$

instead of (3). We also need to replace the corresponding part of the denominator of (4) by $(C_1(\mu_1)\alpha_1 - \sum_{i=2}^K C_i(\mu_i)\alpha_i - C_0(\mu_0)\alpha_0)^{\frac{1}{p}}$.

Similarly, one can relax Theorem 2.2 by assuming that both μ_0 and μ_1 are approximately spherically symmetric (with bounded support) as well conditions (77) and (78). This will imply though that the global l_0 subspace is a local l_p subspace only when $N_0 = o(N_1)$ (instead of $N_0 = o(N_1^2)$).

In Theorem 2.2 it is also possible to replace the spherical symmetry assumption on μ_0 by symmetry with respect to L_1, without changing the implication of that theorem. It is even possible to assume a slightly weaker assumption: $E_{\mu_0}(\mathbf{D}_{\mathrm{L}_1,\mathbf{x},p}) = 0$, where $\mathbf{D}_{\mathrm{L}_1,\mathbf{x},p}$ was defined in (68).

4.2. The Case of Affine Subspaces

The assumption of spherical symmetry is natural in the setting of linear subspaces, unlike affine subspaces. We can only formulate a weak theory for l_p-recovery of a single subspace among affine subspaces intersecting a fixed ball. For example, one can assume that the mixture distribution $\alpha_0\mu_0 + \sum_{i=2}^{K} \alpha_i\mu_i$ is approximately spherically symmetric with a bounded support and apply the theory developed in this paper to recover L_1 by l_p minimization. Strong restrictions on the sampling along affine subspaces are needed in order to avoid cases in the spirit of of Section 2.1. For example, points on a subspace, which is sufficiently far from the origin and sufficiently dense but not the global l_0 subspace, are outliers that can misguide the recovery of the global l_0 subspace by l_p minimization for all $p > 0$.

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.

4.3. Implementation and Relation to Other Algorithms

One can approximate the geometric l_1 minimizer by gradient descent or stochastic gradient descent (see e.g., [31]). However, since the underlying minimization is not convex such approximation will likely converge to a local minimum different than the global one. It will be interesting to suggest a convex strategy that is closely related to the geometric l_1 minimization without including an additional parameter.

Two convex strategies which include an additional parameter are the principal component pursuit [2] and the outlier pursuit [29]. It is possible that by carefully choosing the tuning parameter of [29], the rows of the low rank matrix obtained by [29] span the d-subspace that minimizes the l_1 energy in (1).

Acknowledgement

This work is inspired by our collaboration with Arthur Szlam on efficient and fast algorithms for hybrid linear modeling, which apply geometric l_1 minimization. The main impetus for this paper was a question by Arthur on the analog of [11, Theorem 2] for geometric l_1 approximation. We thank John Wright for referring us to [26, 27] and for relevant questions which we address in Section 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), in particular Doug Arnold and Fadil Santosa, for holding a workshop on multi-manifold modeling that GL co-organized and TZ participated in. This workshop broadened our perspective on the

relation between hybrid linear modeling and sparse approximation. GL thanks David Donoho for inviting him for a visit at Stanford University in Fall 2003 and for stimulating discussions at that time on the intellectual responsibilities of mathematicians analyzing massive and high-dimensional data as well as general advice. Those discussions effected GL's research program and his mentorship (TZ is a PhD candidate advised by GL). Both authors have been supported by NSF grants DMS-09-15064 and DMS-09-56072.

Appendix A: Supplementary Details

A.1. Upper Bound of ψ_μ for a Uniform Distribution in $\mathrm{B}(\mathbf{0},1)\cap \mathrm{L_1}$

We establish here the following upper bound on ψ_μ in the special case where μ is uniform on $\mathrm{B}(\mathbf{0},1)\cap \mathrm{L}_1$ and L_1 is a d-subspace in $\mathbb{R}^D$:

$$\psi_\mu(t) < \frac{2d}{\pi}\, t. \tag{79}$$

This implies a lower bound on ψ_μ^{-1}, which simplifies some of the estimates of this paper (involving ψ_μ^{-1}) in this special case.

Denoting the volume of d-dimensional unit ball by v_d and noticing that

$$\begin{aligned}&\{\mathbf{x}=(x_1,x_2,\cdots,x_d)\in \mathrm{B}(\mathbf{0},1)\cap \mathrm{L}_1 : |x_1|<t\}\\ \subset&\left\{\mathbf{x}=(x_1,x_2,\cdots,x_d)\in \mathrm{B}(\mathbf{0},1)\cap \mathrm{L}_1 : |x_1|<t, |x_2|\le 1, \sum_{i=3}^{d} x_i^2 \le 1\right\},\end{aligned}$$

we have that

$$\mathrm{Vol}\,\{\mathbf{x}:\mathbf{x}\in \mathrm{B}(\mathbf{0},1)\cap \mathrm{L}_1, |x_1|<t\} < 4\mathrm{v}_{d-2}t. \tag{80}$$

Combining (80) with the observation: $\mathrm{v}_d = \frac{2\pi}{d}\mathrm{v}_{d-2}$, we find the upper bound of $\psi_\mu(t)$:

$$\begin{aligned}&\psi_\mu(t) = \mathrm{Vol}\,\{\mathbf{x}\in \mathrm{B}(\mathbf{0},1)\cap \mathrm{L}_1 : |x_1|<t)\}\Big/\mathrm{Vol}\,\{\mathbf{x}\in \mathrm{B}(\mathbf{0},1)\cap \mathrm{L}_1\}\\ <&\frac{4\mathrm{v}_{d-2}t}{\mathrm{v}_d} = \frac{2d}{\pi}t.\end{aligned}$$

A.2. Proof of Lemma 3.1

We will use the following inequality, which we verify below in Section A.2.1:

$$\mu_1\left(\mathbf{x}\in \mathrm{B}(\mathbf{0},1)\cap \mathrm{L}_1 : \mathrm{dist}(\mathbf{x},\hat{\mathrm{L}}_1) < \beta\,\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1,\hat{\mathrm{L}}_1)\right) \le \psi_{\mu_1}(\frac{\pi\sqrt{d}}{2}\beta)\ \ \forall \beta>0. \tag{81}$$

We fix $\beta_1 = \frac{2}{\pi\sqrt{d}}\psi_{\mu_1}^{-1}\left(\frac{1+\mu_1(\{\mathbf{0}\})}{2}\right)$ and later prove the existence of this constant. Using the fact that $\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1,\hat{\mathrm{L}}_1)=\epsilon$ and applying (81), we obtain that

$$\mu_1\left(\mathbf{x}\in \mathrm{B}(\mathbf{0},1)\cap \mathrm{L}_1 : \mathrm{dist}(\mathbf{x},\hat{\mathrm{L}}_1) < \beta_1\,\epsilon\right)$$

$$= \mu_1\left(\mathbf{x} \in \mathrm{B}(\mathbf{0},1) \cap \mathrm{L}_1 : \mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}}_1) < \beta_1 \, \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1, \hat{\mathrm{L}}_1)\right) \leq (1 + \mu_1(\{\mathbf{0}\}))/2.$$

Consequently, we derive the following estimate

$$\mu_1\left(\mathbf{x} \in \mathrm{B}(\mathbf{0},1) \cap \mathrm{L}_1 : \mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}}_1) \geq \beta_1 \epsilon\right) \geq (1 - \mu_1(\{\mathbf{0}\}))/2,$$

and thus by Chebyshev's inequality the lemma is concluded as follows:

$$E_{\mu_1}\left(e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1)\right) \geq \beta_1^p \epsilon^p / 2 = \frac{(1 - \mu_1(\{\mathbf{0}\}))2^{p-1}\epsilon^p}{(\pi\sqrt{d})^p \psi_{\mu_1}^{-1}\left(\frac{1+\mu_1(\{\mathbf{0}\})}{2}\right)^p}.$$

The existence of $\psi_{\mu_1}^{-1}\left(\frac{1+\mu_1(\{\mathbf{0}\})}{\pi\sqrt{d}}\right)$ will follow from the following observation:

$$\begin{aligned} &\mu_1(\mathrm{L}) = 0 \ \text{ for any affine subspace } \mathrm{L} \subset \mathrm{L}_1, \\ &\mu_1(\mathrm{L}) = \mu_1(\{\mathbf{0}\}) \text{ for any linear subspace } \mathrm{L} \subsetneq \mathrm{L}_1, \end{aligned} \tag{82}$$

We prove it as follows: Assume that d_0 is the smallest dimension for which there exists a subspace L_0 such that (82) is not true, then we arbitrarily rotate L_0 with respect to the origin large number of times. Each of the rotated subspaces has the same positive measure as L_0, and the measure of the intersection between any such pair is $\mathbf{0}$ (since the intersection has a lower dimension than d_0), therefore the measure of the union of these rotated subspaces can be arbitrarily large, which contradicts $\mu_1(\mathbb{R}^D) = 1$. Then we proved (82), and from it we obtain that $\psi_{\mu_1}(0) = \mu_1(\{\mathbf{0}\})$, $\psi_{\mu_1}(1) = 1$, and $\psi_{\mu_1}(t)$ is continuous in the interval $[0,1]$. Therefore, the existence of $\psi_{\mu_1}^{-1}\left(\frac{1+\mu_1(\{\mathbf{0}\})}{2}\right)$ is concluded.

A.2.1. Proof of (81)

We denote the principal angles between L_1 and $\hat{\mathrm{L}}_1$ by $\{\theta_i\}_{i=1}^d$, the principle vectors of L_1 and $\hat{\mathrm{L}}_1$ by $\{\mathbf{v}_i\}_{i=1}^d$ and $\{\hat{\mathbf{v}}_i\}_{i=1}^d$ respectively, the interaction dimension by $k \equiv k(\mathrm{L}_1, \mathrm{L}_2)$ (see Section 3.2.1), the volume of the d-dimensional unit ball by v_d and

$$\gamma_i = \frac{\sin(\theta_i)^2}{\sum_{j=1}^k \sin(\theta_j)^2}, \ \ i = 1, \ldots, k.$$

Since $\sum_{i=1}^k \gamma_i = 1$, WLOG we assume that $\gamma_1 \geq 1/k \geq 1/d$. Expressing every point $\mathbf{x}$ in L_1 by $\mathbf{x} = (x_1, x_2, \cdots, x_d) = (\mathbf{v}_1^T\mathbf{x}, \mathbf{v}_2^T\mathbf{x}, \cdots, \mathbf{v}_d^T\mathbf{x})$, we obtain that

$$\begin{aligned} &\left\{\mathbf{x} \in \mathrm{L}_1 : \mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}}_1) < \beta \, \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1, \hat{\mathrm{L}}_1)\right\} \\ = &\left\{\mathbf{x} = (x_1, x_2, \cdots, x_d) \in \mathrm{L}_1 : \sqrt{\sum_{i=1}^d x_i^2 \sin^2\theta_i} < \beta \sqrt{\sum_{i=1}^d \theta_i^2}\right\} \end{aligned}$$

$$
\begin{aligned}
&\subset \left\{ \mathbf{x} = (x_1, x_2, \cdots, x_d) \in \mathrm{L}_1 : \sqrt{\sum_{i=1}^{d} x_i^2 \sin^2\theta_i} < \frac{\pi}{2}\beta \sqrt{\sum_{i=1}^{d} \sin^2\theta_i} \right\} \\
&= \left\{ \mathbf{x} = (x_1, x_2, \cdots, x_d) \in \mathrm{L}_1 : \sqrt{\sum_{i=1}^{k} \gamma_i\, x_i^2} < \frac{\pi}{2}\beta \right\} \\
&\subset \left\{ \mathbf{x} = (x_1, x_2, \cdots, x_d) \in \mathrm{L}_1 : |x_1| < \frac{\pi}{2\sqrt{\gamma_1}}\beta \right\} \\
&\subset \left\{ \mathbf{x} \in \mathrm{L}_1 : |\mathbf{v}_1^T \mathbf{x}| < \frac{\pi\sqrt{d}}{2}\beta \right\}.
\end{aligned}
$$

We prove (81), by combing the equation above and

$$
\mu_1\left(\left\{ \mathbf{x} \in \mathrm{L}_1 : |\mathbf{v}_1^T \mathbf{x}| < \frac{\pi\sqrt{d}}{2}\beta \right\} \right) = \psi_{\mu_1}(\frac{\pi\sqrt{d}}{2}\beta).
$$

A.3. Proof of Lemma 3.2

We denote the principal angles between the d-subspaces L_1, L_2 by $\theta_1 \geq \theta_2 \geq \theta_3 \geq \cdots \geq \theta_d$. Arbitrarily choosing $\mathbf{Q}_1$, $\mathbf{Q}_2 \in \mathrm{O}(D, d)$, representing L_1, L_2 respectively, we note that

$$
\begin{aligned}
&|\mathrm{dist}(\mathbf{x}, \mathrm{L}_1) - \mathrm{dist}(\mathbf{x}, \mathrm{L}_2)| = |\,||\mathbf{x} - \mathbf{x}\mathbf{Q}_1\mathbf{Q}_1^T|| - ||\mathbf{x} - \mathbf{x}\mathbf{Q}_2\mathbf{Q}_2^T||\,| \\
\leq &||\mathbf{x} - \mathbf{x}\mathbf{Q}_1\mathbf{Q}_1^T - \mathbf{x} + \mathbf{x}\mathbf{Q}_2\mathbf{Q}_2^T|| \leq ||\mathbf{x}||\, \left\|\mathbf{Q}_1\mathbf{Q}_1^T - \mathbf{Q}_2\mathbf{Q}_2^T\right\|_{\mathrm{F}} \\
= &||\mathbf{x}|| \sqrt{\sum_{i=1}^{d} \sin(\theta_i)^2} \leq ||\mathbf{x}|| \sqrt{\sum_{i=1}^{d} \theta_i^2} = ||\mathbf{x}||\, \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1, \mathrm{L}_2).
\end{aligned}
$$

A.4. Proof of Lemma 3.3

We assume WLOG that $i = 1$ in (10). We thus need to prove that for all $\hat{\mathrm{L}} \in \mathrm{G}(D, d)$:

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{x}_1 \in \mu_1}(\mathrm{dist}(\mathbf{x}_1, \hat{\mathrm{L}})^p) + \mathbb{E}_{\mathbf{x}_2 \in \mu_2}(\mathrm{dist}(\mathbf{x}_2, \hat{\mathrm{L}})^p) \\
\geq &\mathbb{E}_{\mathbf{x}_1 \in \mu_1}(\mathrm{dist}(\mathbf{x}_1, \mathrm{L}_1)^p) + \mathbb{E}_{\mathbf{x}_2 \in \mu_2}(\mathrm{dist}(\mathbf{x}_2, \mathrm{L}_1)^p).
\end{aligned} \tag{83}
$$

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 $\{\mathbf{v}_i\}_{i=1}^d$ and $\{\hat{\mathbf{v}}_i\}_{i=1}^d$ and the complementary orthogonal system for L_2 w.r.t. L_1 by $\{\mathbf{u}_i\}_{i=1}^d$.

We notice that we can restrict the set of subspaces $\hat{\mathrm{L}}$ satisfying (83). First of all, we only need to consider subspaces

$$
\hat{\mathrm{L}} \in \mathrm{L}_1 + \mathrm{L}_2\,. \tag{84}
$$

Indeed, the LHS of (83) is the same if we replace $\hat{\mathrm{L}}$ by $\hat{\mathrm{L}} \cap (\mathrm{L}_1 + \mathrm{L}_2)$.

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

$$\mathrm{Sp}(\hat{\mathbf{v}}_i, \mathbf{v}_i) \nsubseteq \hat{\mathrm{L}} \ \text{ for all } 1 \le i \le k. \tag{85}$$

Indeed, WLOG let $i = 1$ and suppose on the contrary to (85) that $\hat{\mathbf{v}}_1, \mathbf{v}_1 \in \hat{\mathrm{L}}$. Since $\hat{\mathrm{L}}$ is d-dimensional, there exists $2 \le j \le d$ (assume WLOG $j = 2$) such that it does not contain both $\hat{\mathbf{v}}_j$ and $\mathbf{v}_j$. For any pair of points $\mathbf{x} = \sum_{i=1}^d a_i \mathbf{v}_i \in \mathrm{L}_1$ and $\hat{\mathbf{x}} = \sum_{i=1}^d a_i \hat{\mathbf{v}}_i \in \mathrm{L}_2$:

$$\mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}}) = \sqrt{\sin(\theta_2)^2 a_2^2 + \nu_1^2} \ \text{ and } \ \mathrm{dist}(\hat{\mathbf{x}}, \hat{\mathrm{L}}) = \sqrt{\sin(\theta_1)^2 a_1^2 + \nu_2^2},$$

where

$$\nu_1 = \mathrm{dist}\left(\sum_{i=3}^d a_i \mathbf{v}_i, \hat{\mathrm{L}}\right) \ \text{ and } \ \nu_2 = \mathrm{dist}\left(\sum_{i=3}^d a_i \hat{\mathbf{v}}_i, \hat{\mathrm{L}}\right).$$

Now, for $\tilde{\mathrm{L}} = \mathrm{Sp}(\hat{\mathrm{L}} \setminus \{\mathbf{v}_1, \hat{\mathbf{v}}_1\}, \mathbf{v}_1, \mathbf{v}_2)$, we obtain that

$$\mathrm{dist}(\hat{\mathbf{x}}, \tilde{\mathrm{L}}) = \sqrt{\sin(\theta_1)^2 a_1^2 + \sin(\theta_2)^2 a_2^2 + \nu_2^2} \ \text{ and } \ \mathrm{dist}(\mathbf{x}, \tilde{\mathrm{L}}) = \nu_1.$$

Therefore

$$\mathrm{dist}(\mathbf{x}, \tilde{\mathrm{L}})^p + \mathrm{dist}(\hat{\mathbf{x}}, \tilde{\mathrm{L}})^p \le \mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}})^p + \mathrm{dist}(\hat{\mathbf{x}}, \hat{\mathrm{L}})^p$$

and by direct integration we have that

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}_1 \in \mu_1}(\mathrm{dist}(\mathbf{x}_1, \tilde{\mathrm{L}})^p) + \mathbb{E}_{\mathbf{x}_2 \in \mu_2}(\mathrm{dist}(\mathbf{x}_2, \tilde{\mathrm{L}})^p) \\ \le &\mathbb{E}_{\mathbf{x}_1 \in \mu_1}(\mathrm{dist}(\mathbf{x}_1, \hat{\mathrm{L}})^p) + \mathbb{E}_{\mathbf{x}_2 \in \mu_2}(\mathrm{dist}(\mathbf{x}_2, \hat{\mathrm{L}})^p). \end{aligned}$$

We can thus replace the subspace $\hat{\mathrm{L}}$ with the subspace $\tilde{\mathrm{L}}$, which satisfies (85) (for $i = 1$, but can similarly be changed for all $1 < i \le K$).

It follows from (84) and (85) that $\hat{\mathrm{L}}$ can be represented as follows:

$$\hat{\mathrm{L}} = \mathrm{Sp}(\mathbf{v}_1^*, \mathbf{v}_2^*, \cdots, \mathbf{v}_d^*),$$

where

$$\mathbf{v}_i^* = \cos\theta_i^* \mathbf{v}_i + \sin\theta_i^* \mathbf{u}_i.$$

Thus, for any pair of points $\mathbf{x} = \sum_{i=1}^d a_i \mathbf{v}_i \in \mathrm{L}_1$ and $\hat{\mathbf{x}} = \sum_{i=1}^d a_i \hat{\mathbf{v}}_i \in \mathrm{L}_2$:

$$\mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}}) = \sqrt{\sum_{i=1}^d \sin^2\theta_i^* a_i^2} \ \text{ and } \ \mathrm{dist}(\hat{\mathbf{x}}, \hat{\mathrm{L}}) = \sqrt{\sum_{i=1}^d \sin^2(\theta_i - \theta_i^*) a_i^2} \tag{86}$$

and

$$\mathrm{dist}(\mathbf{x}, \mathrm{L}_1) = 0 \ \text{ and } \ \mathrm{dist}(\hat{\mathbf{x}}, \mathrm{L}_1) = \sqrt{\sum_{i=1}^d \sin^2\theta_i a_i^2}. \tag{87}$$

Combining (86), (87), the triangle inequality (for "sine vectors" in $\mathbb{R}^d$) and the subadditivity of the sine function, we conclude that

$$\begin{aligned}\operatorname{dist}(\mathbf{x},\hat{\mathrm{L}})+\operatorname{dist}(\hat{\mathbf{x}},\hat{\mathrm{L}}) &\geq \sqrt{\sum_{i=1}^{d}\left(\sin\theta_i^*+\sin\left(\theta_i-\theta_i^*\right)\right)^2 a_i^2}\\ &\geq \sqrt{\sum_{i=1}^{d}\sin^2\theta_i a_i^2}=\operatorname{dist}(\hat{\mathbf{x}},\mathrm{L}_1)+\operatorname{dist}(\mathbf{x},\mathrm{L}_1).\end{aligned}$$

Since $p\leq 1$, this inequality extends to

$$\operatorname{dist}(\mathbf{x},\hat{\mathrm{L}})^p+\operatorname{dist}(\hat{\mathbf{x}},\hat{\mathrm{L}})^p\geq\operatorname{dist}(\hat{\mathbf{x}},\mathrm{L}_1)^p=\operatorname{dist}(\hat{\mathbf{x}},\mathrm{L}_1)^p+\operatorname{dist}(\mathbf{x},\mathrm{L}_1)^p. \tag{88}$$

Integrating (88) w.r.t. the uniform distribution we conclude (83) and thus prove the lemma.

A.5. Proof of (9)

The fact that $E_{\mu_1}(P_{\mathrm{L}_1}(\mathbf{x})P_{\mathrm{L}_1}(\mathbf{x})^T)$ is a scalar matrix follows from the symmetry of μ_1 on $\mathrm{L}_1\cup\mathrm{B}(\mathbf{0},1)$. We compute the underlying scalar, δ^*, as follows. We arbitrarily fix a vector $\mathbf{v}\in\mathbb{R}^d$ as well as a $(d-1)$-subspace $\hat{\mathrm{L}}_1\subseteq\mathrm{L}_1$ orthogonal to $\mathbf{v}$ and observe that

$$\delta_*=E_{\mu_1}\left((P_{\mathrm{L}_1}(\mathbf{x})^T\,\mathbf{v})^2\right)=E_{\mu_1}\left(\operatorname{dist}(\mathbf{x},\hat{\mathrm{L}})^2\right).$$

We further note that for any $0<r\leq 1$, the set $\{\mathbf{x}\in\mathrm{B}(\mathbf{0},1)\cap\mathrm{L}_1:\operatorname{dist}(\mathbf{x},\hat{\mathrm{L}})=r\}$ consists of two $(d-1)$-dimensional balls of radius $\sqrt{1-r^2}$. We consequently compute the constant δ_* using the beta function B and the Gamma function Γ in the following way:

$$\begin{aligned}\delta_*=E_{\mu_1}\left(\operatorname{dist}^2(\mathbf{x},\hat{\mathrm{L}})\right)&=\frac{\int_{r=0}^1 r^2(1-r^2)^{\frac{d-1}{2}}\,\mathrm{d}t}{\int_{r=0}^1(1-r^2)^{\frac{d-1}{2}}\,\mathrm{d}t}=\frac{\int_{\theta=0}^{\frac{\pi}{2}}\sin^2(\theta)\cos^{\frac{d+1}{2}}(\theta)\,\mathrm{d}\theta}{\int_{\theta=0}^{\frac{\pi}{2}}\cos^{\frac{d+1}{2}}(\theta)\,\mathrm{d}\theta}\\ &=\frac{B(\frac{3}{2},\frac{d+1}{2})}{B(\frac{1}{2},\frac{d+1}{2})}=\frac{\Gamma(\frac{3}{2})\,\Gamma(\frac{d+1}{2})\,\Gamma(\frac{d+2}{2})}{\Gamma(\frac{1}{2})\,\Gamma(\frac{d+1}{2})\,\Gamma(\frac{d+4}{2})}=\frac{1}{d+2}\,.\end{aligned}$$

A.6. Proof of (32)

For simplicity we denote $\mathbf{B}=\sum_{i=1}^{N_1}P_{\mathrm{L}_1}(\mathbf{x}_i)P_{\mathrm{L}_1}(\mathbf{x}_i)^T$. We note that if $\max_t\sigma_t\left(\mathbf{B}-\delta_*\mathbf{I}_d\right)<\eta$, then

$$\frac{\|\mathbf{B}\mathbf{v}-\delta_*\mathbf{v}\|}{\|\mathbf{v}\|}<\eta\ \text{ for all } v\in\mathbb{R}^d\setminus\{\mathbf{0}\},$$

and consequently

$$\delta_*-\eta<\frac{\|\mathbf{B}\mathbf{v}\|}{\|\mathbf{v}\|}\ \text{ for all } v\in\mathbb{R}^d\setminus\{\mathbf{0}\},$$

that is, $\min_t\sigma_t(\mathbf{B})>\delta_*-\eta$.

References

[1] A. Baccini, P. Besse, and A. de Falguerolles. A L_1-norm PCA and a heuristic approach. In E. Diday, Y. Lechevalier, and O. Opitz, editors, *Ordinal and symbolic data analysis*, pages 359–368, New York, 1996. Springer.

[2] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? Submitted, Dec. 2009, arXiv:0912.3599.

[3] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, 2006.

[4] E. J. Candès, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

[5] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61 (electronic), 1998.

[6] N. P. da Silva and J. P. Costeira. Subspace segmentation with outliers: A grassmannian approach to the maximum consensus subspace. In *CVPR*. IEEE Computer Society, 2008.

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

[8] C. Ding, D. Zhou, X. He, and H. Zha. R1-PCA: rotational invariant l1-norm principal component analysis for robust subspace factorization. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 281–288, New York, NY, USA, 2006. ACM.

[9] D. L. Donoho. For most large underdetermined systems of equations, the minimal l_1-norm near-solution approximates the sparsest near-solution. *Comm. Pure Appl. Math.*, 59(7):907–934, 2006.

[10] D. L. Donoho. For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.*, 59(6):797–829, 2006.

[11] D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization. *Proc. Natl. Acad. Sci. USA*, 100(5):2197–2202 (electronic), 2003.

[12] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353 (electronic), 1999.

[13] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. of the ACM*, 24(6):381–395, June 1981.

[14] J. Gao. Robust L_1 principal component analysis and its Bayesian variational inference. *Neural Comput.*, 20(2):555–572, 2008.

[15] G. Golub and C. V. Loan. *Matrix Computations*. John Hopkins University Press, Baltimore, Maryland, 1996.

[16] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions (Wiley Series in Probability and Statistics)*. Wiley-Interscience, New York, revised edition, April 2005.

[17] P. J. Huber. *Robust Statistics*. John Wiley & Sons Inc., New York, 1981. Wiley Series in Probability and Mathematical Statistics.

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

[19] N. Kwak. Principal component analysis based on L1-norm maximization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(9):1672–1680, 2008.

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

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

[22] P. Mattila. *Geometry of Sets and Measures in Euclidean Spaces*. Cambridge University Press, 1995.

[23] P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1987.

[24] S. J. Szarek. The finite-dimensional basis problem with an appendix on nets of Grassmann manifolds. *Acta Math.*, 151(3-4):153–179, 1983.

[25] S. J. Szarek. Metric entropy of homogeneous spaces. In *Quantum probability (Gdańsk, 1997)*, volume 43 of *Banach Center Publ.*, pages 395–410. Polish Acad. Sci., Warsaw, 1998.

[26] P. H. S. Torr and A. Zisserman. Robust computation and parametrization of multiple view relations. In *ICCV '98: Proceedings of the Sixth International Conference on Computer Vision*, page 727, Washington, DC, USA, 1998. IEEE Computer Society.

[27] P. H. S. Torr and A. Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 78(1):138–156, 2000.

[28] Y.-C. Wong. Differential geometry of Grassmann manifolds. *Proc. Nat. Acad. Sci. U.S.A.*, 57:589–594, 1967.

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

[30] A. Y. Yang, S. R. Rao, and Y. Ma. Robust statistical estimation and segmentation of multiple subspaces. In *CVPRW '06: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, page 99, Washington, DC, USA, 2006. IEEE Computer Society.

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

[32] T. Zhang, A. Szlam, Y. Wang, and G. Lerman. Hybrid linear modeling via local best-fit flats. Available at http://arxiv.org/abs/1010.3460, 2010.

[33] T. Zhang, A. Szlam, Y. Wang, and G. Lerman. Randomized hybrid linear model-

ing by local best-fit flats. pages 1927 –1934, jun. 2010.