Modeling Data by Multiple Subspaces: Theory and Algorithms

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Teng Zhang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Gilad Lerman

August, 2011

Acknowledgements

The research works in this thesis would not have been possible without the support and encouragement of many individuals, and I would like to express my sincere thanks and appreciation to all of them.

Professor Lerman, my PhD advisor, has been offering abundant help, suggestions, guidance and financial support. He replies email fast and meet with me timely. He has spent a huge amount of time, energy and dedication into this project and showed me how to think, work and communicate as a true mathematician.

I am thankful to Arthur Szlam, who offered me great ideas and directions to begin my research, which is part of this work and is also the motivation of another part of the work.

I am very grateful for the financial and academic support from the School of Mathematics at the University of Minnesota. I am also thankful for my committee members, Fadil Santosa, Andrew Odlyzko, and Hui Zou for their helpful feedbacks during my research.

I appreciated the help from other people in Professor Lerman's research group, including Guangliang Chen, Jonathan Tyler Whitehouse and Yi Wang.

Finally, I would like to express my appreciation to my family for their love and support during my study. Without their support I would not be where I am today.

Abstract

We study the problem of modeling data by several affine subspaces, which generalizes the common modeling by a single subspace. It arises, for example, in object tracking and structure from motion. One of the simplest methods for such modeling is based on energy minimization, where the energy involves p-th powers of distances of data points from subspaces. We first analyze under certain assumptions (e.g., spherically symmetric outliers) the effectiveness of such energy minimization for recovering all subspaces simultaneously and also recovering the most significant subspace. We reveal the following phase transition in our setting: when $p \leq 1$ the underlying subspaces can be recovered by such energy minimization; whereas when $p > 1$ the underlying subspaces are sufficiently far from the global minimizer. Nevertheless, for more general settings (i.e., outliers which are not spherically symmetric) we can point at some disadvantages of the energy minimization strategy. In order to practically solve the problem, we present two simple and fast geometric methods for multiple subspaces modeling. One of them minimize energy by gradient descent, and another forms a collection of local best fit affine subspaces, where the size of the local neighborhoods is determined automatically by the Peter Jones beta numbers. This collection of subspaces can then be further processed in various ways. For example, greedy selection procedure according to an appropriate energy or a spectral method to generate the final model. We demonstrate the state of the art accuracy and speed of the suggested procedure on applications for several applications.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the last decade, many algorithms have been developed to model data by multiple subspaces. Such modeling is often referred to as hybrid linear modeling (HLM). It was motivated by concrete problems in computer vision as well as the large effort of nonlinear dimensionality reduction. HLM is the simplest geometric framework for nonlinear dimensionality reduction. Nevertheless, only little theory has been developed to justify the performance of existing methods for such modeling.

There are two main contributions in this thesis: fist, we show the effectiveness of energy minimization methods in HLM. Second, we present two new algorithms for HLM. In the first contribution, we study on two problems: single subspace recovery and multiple subspaces recovery.

Our first study is motivated by the problem of sequential recovery of multiple subspaces buried in outliers, 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.

Principal Component Analysis (PCA) is the most common tool in the recovery of a single d-subspace. 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 $\mathcal{X} \subset \mathbb{R}^D$ and minimizes among all d-dimensional subspaces, L, the quantity:

$$e_{l_p}(\mathcal{X}, \mathrm{L}) = \sum_{\mathbf{x} \in \mathcal{X}} \mathrm{dist}(\mathbf{x}, \mathrm{L})^p, \tag{1.1}$$

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 (1.1) a *global l_p subspace*. We sometimes also refer to this optimization as *geometric l_p minimization*.

In our second study, we analyze the recovery of multiple subspaces by energy minimization. For a fixed parameter $p > 0$ and a data set $\mathcal{X}$, one can model $\mathcal{X}$ with K subspaces obtained by minimizing the following energy over the subspaces $\mathrm{L}_1, \cdots, \mathrm{L}_K$:

$$e_{l_p}(\mathcal{X}, \mathrm{L}_1, \cdots, \mathrm{L}_K) = \sum_{\mathbf{x} \in \mathcal{X}} \mathrm{dist}^p(\mathbf{x}, \cup_{i=1}^K \mathrm{L}_i), \tag{1.2}$$

where $\mathrm{dist}(\cdot, \cdot)$ denotes the Euclidean distance. For simplicity, we assume that $L_1, \ldots, L_K$ are d-subspaces (generalizations are discussed in Sections 5.6 and 5.7). In order to study the recovery problem, we suppose that the data $\mathcal{X}$ was independently and identically sampled from a mixture probability distribution μ of K spherically symmetric distributions along distinct d-subspaces and a rather general distribution of outliers. This way the energy (1.2) is the expectation of the function $\mathrm{dist}^p(\mathbf{x}, \cup_{i=1}^K \mathrm{L}_i)$ according to the empirical measure associated with i.i.d. samples from μ, while being multiplied by the number of samples. The recovery problem asks whether with overwhelming probability the minimization of (1.2) recovers the underlying subspaces. We show here that when $p \leq 1$ the answer to this problem is positive, whereas when $p > 1$ it is negative (here the assumption of spherical symmetry is unnecessary). Those results extend to homoscedastic noise around the subspaces, while replacing exact recovery with near recovery.

Recovery problems are common (e.g., recovery of a single subspace as in least squares type problems or recovery of multiple centers as in K-means). However, our current setting is rather recent and requires novel developments. One of the issues is the strong geometric nature of the problem, resulting from an optimization on a product space of Grassmannians. The other one is the difficulty to approximate the problem by convex optimization (as we clarify in 5.2). Thus even though it is a very elementary, though unexplored, problem in learning, it requires the development of techniques which are currently not widely common in learning.

In our algorithm part, we develop algorithms for HLM based on energy minimization. Indeed, many algorithms have been developed for hybrid linear modeling [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]. They find diverse applications in several areas, such as motion segmentation in computer vision, hybrid linear representation of images, classification of face images and temporal segmentation of video sequences (see e.g., [10, 13, 19]).

A sequential HLM algorithm was suggested by Yang et al. [12] using the Random Sample Consensus (RANSAC) [20] 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 [21, 22] 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.

The K-subspaces algorithm [1, 5, 6, 9] is the most basic heuristic for HLM. It is based on an iterative process aimed at minimizing the energy (1.2) with $p = 2$. Practically, it iterates the following two stages: 1) clustering points according to proximity to the given subspaces; 2) computing the underlying l_2 subspace of each given cluster (according to principal component analysis). The initial subspaces (or clusters) are either chosen randomly or by an output of another algorithm. Numerical experiments by Zhang et al. [17] have shown that this procedure is in general not robust to outliers; whereas a different method for minimizing (1.2) with $p = 1$ seems to be robust to outliers. We note that the K-means algorithm is the special case of the K-subspaces algorithm for 0-dimensional affine subspaces, i.e., points.

The rest of the thesis is organized as follows. In Chapter 2 we review some background on HLM modeling and l_1 minimization, and introduce the main results of this thesis. In Chapter 3 we present the theoretical analysis to the energy minimization problems and prove the main theorems. Chapter 4 presents two practical algorithms for HLM problem based on our analysis and compares them with other competing methods in artificial data sets and real-world applications. Chapter 5 concludes this paper and discusses some immediate extensions of its results as well as open directions.

Chapter 2

Theory for Recovery Subspaces by l_p minimization

2.1 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) [23].

The l_1 norm has also been successfully applied to robust regression [24, 25, 26, 27]. Furthermore, Basis pursuit [28] 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 [29, 30, 31, 32]. Candès et al. [33] 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 [34, 35, 36, 37].

Geometric l_p minimization (as in (1.1)) has been proposed by Guy David and Stephen Semmes [38] for $p \geq 1$ in a pure analytic setting (free of outliers in the context of "Ahlfors

regular measures"). Ding et al. [39] used the geometric l_1 minimization as a robust alternative for principal component analysis, though lacking any mathematical support. Very Recently, Xu et al. [40] have suggested the combination of the norm used in the geometric l_1 minimization with the nuclear norm (similar to Candès et al. [33]) 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).

2.2 Basic Conventions and Notation

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

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. We will measure distances between F and G in $\mathrm{G}(D,d)$ by the metric

$$\mathrm{dist}_{\mathrm{G}}(\mathrm{F},\mathrm{G})=\sqrt{\sum_{i=1}^{d}\theta_i^2}, \tag{2.1}$$

where $\{\theta_i\}_{i=1}^d$ are the principal angles between F and G. We use this distance since there is a simple formula for the geodesic lines on the Grassmannian equipped with this distance [41, equation 12], which is implicitly used in this paper. We designate an open ball in $\mathrm{G}(D,d)$ by $\mathrm{B}_{\mathrm{G}}(\mathrm{L},r)$ as opposed to the Euclidean open ball in $\mathbb{R}^D$, $\mathrm{B}(\mathbf{x},r)$. Following [42, Section 3.9], we denote by $\gamma_{D,d}$ the "uniform probability measure on $\mathrm{G}(D,d)$".

In formulating our theorems, we use the constant

$$\tau_0:=\frac{(1-\mu_1(\mathbf{0}))\cdot 2^{p-1}}{(\pi\sqrt{d})^p\cdot\psi_{\mu_1}^{-1}\left(\frac{1+(2K-1)\mu_1(\{\mathbf{0}\})}{2K}\right)^{\bar{p}}}, \tag{2.2}$$

where

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

(an estimate of the function ψ_μ for a uniform distribution on a d-dimensional ball appears in Appendix A.1).

We denote by $a \vee b$ and $a \wedge b$ the maximum and minimum of a and b respectively. 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.

2.3 Setting of This Paper

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.

Definition 2.3.1. *A probability measure μ on $\mathbb{R}^D$ is a* spherically symmetric HLM measure *(equivalently, spherically symmetric HLM measure with noise level $\epsilon = 0$) if $\mu = \sum_{i=0}^K \alpha_i \mu_i$, where $\alpha_0 \geq 0$, $\alpha_i > 0$, $\forall 1 \leq i \leq K$, $\sum_{i=0}^K \alpha_i = 1$, μ_0 is a spherically symmetric Borel probability measure with bounded support (it represents outliers) and $\{\mu_i\}_{i=1}^K$ are Borel 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 $\mu_{i,\epsilon} = \mu_i \times \nu_{i,\epsilon}$, $\forall 1 \leq i \leq K$, $\{\alpha_i\}_{i=0}^K$, $\{\mathrm{L}_i^*\}_{i=1}^K$ and $\{\mu_i\}_{i=0}^K$ are the same as above and $\{\nu_{i,\epsilon}\}_{i=1}^K$ are Borel probability measures with bounded support in the corresponding orthogonal complements of $\{\mathrm{L}_i^*\}_{i=1}^K$ and with p-th moments are smaller than ϵ^p for all $p \leq 1$.*

Some of the theory developed here even applies to the following model:

Definition 2.3.2. *For $\epsilon \geq 0$, we say that a probability measure μ_ϵ on $\mathbb{R}^D$ is a* weak HLM measure with noise level ϵ *if $\mu_\epsilon = \alpha_0\mu_0 + \sum_{i=1}^K \alpha_i \mu_i \times \nu_{i,\epsilon}$, with the following components: $\{\alpha_i\}_{i=0}^K$, $\{\mathrm{L}_i^*\}_{i=1}^K$ and $\{\nu_{i,\epsilon}\}_{i=1}^K$ are the same as in Definition 2.3.1; $\{\mu_i\}_{i=1}^K$ are non-degenerate Borel probability measures on $\{\mathrm{L}_i^*\}_{i=1}^K$ respectively with bounded support; and μ_0 is a Borel probability measure on $\mathbb{R}^D$ with bounded support and there exists $r > 0$ such that the D-dimensional*

Lebesgue measure on $\mathrm{B}(\mathbf{0}, r)$ *is absolutely continuous with respect to the restriction of* μ_0 *to* $\mathrm{B}(\mathbf{0}, r)$.

In order to simplify our estimates we further assume that the supports of all underlying distributions (e.g., $\{\mu_i\}_{i=0}^K$, μ and μ_ϵ) lie in the unit ball (instead of arbitrarily bounded support).

In the minimization of (1.1) we also assume that

$$\alpha_1 > \sum_{i=2}^{K} \alpha_i \tag{2.3}$$

and consequently say that L_1^* is the *most significant subspace*.

The minimization of (1.2) is to recover (or nearly recover when $\epsilon > 0$) the underlying subspaces $\mathrm{L}_1^*, \ldots, \mathrm{L}_K^*$ from the sampled data $\mathcal{X}$ by the l_p minimization of (1.2).

2.4 Main Theorems of Single Subspace Recovery

In this section, we introduce the main results in minimizing (1.1). 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 2.4.1. *If* μ *is a spherically symmetric HLM measure on* $\mathbb{R}^D$ *with* K *d-subspaces* $\{\mathrm{L}_i\}_{i=1}^K \subset \mathrm{G}(D, d)$ *and mixture coefficients* $\{\alpha_i\}_{i=0}^K$ *satisfying* (2.3)*,* $\mathcal{X}$ *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* $\min_{2\leq i\leq K}(\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1, \mathrm{L}_i))$.

Theorem 2.4.2. *If* $\epsilon > 0$*,* μ_ϵ *is a spherically symmetric HLM measure on* $\mathbb{R}^D$ *of noise level* ϵ *with* K *d-subspaces* $\{\mathrm{L}_i\}_{i=1}^K \subset \mathrm{G}(D, d)$ *and mixture coefficients* $\{\alpha_i\}_{i=0}^K$ *satisfying* (2.3)*,* $\mathcal{X}$ *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*

$$f \equiv f(\epsilon, K, d, p, \alpha_0, \alpha_1) = \frac{\pi\sqrt{d}\ \psi_{\mu_1}^{-1}\left(\frac{1+\mu_1(\{\mathbf{0}\})}{2}\right)\epsilon}{\left(\alpha_1 - \sum_{i=2}^K \alpha_i\right)^{\frac{1}{p}} (1-\mu_1(\{\mathbf{0}\}))^{\frac{1}{p}}\, 2^{\frac{p-3}{p}}}, \tag{2.4}$$

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

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

In Section 3.7 we show that Theorem 2.4.2 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 2.4.3. *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. $\{\mathrm{L}_i\}_{i=1}^K \subset \mathrm{G}(D,d)$ 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 $\mathrm{B}_{\mathrm{G}}(\mathrm{L}_1, \kappa_0)$ with overwhelming probability.*

We remark on the size of δ_0 and κ_0 in Section 3.5

Theorems 2.4.1, 2.4.2 and 2.4.3 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 [20]) in a spherically symmetric HLM setting by minimizing l_p distances in the spirit of [21, 22]. 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 [21, 22] or Huber's loss function [24], 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 [21, 22] 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 3.1.

In the setting of spherically symmetric HLM measure with no noise, Theorem 2.4.1 can be repetitively applied to justify sequential HLM using l_p minimization with $0 < p \leq 1$. Rigorous application of Theorem 2.4.2 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., [18, 19] and [43]).

2.5 Main Theorems of Multiple Subspaces Recovery

In this section, we introduce the main result in minimizing (1.2). We first formulate the exact l_p recovery (w.o.p.) of the underlying subspaces whenever $0 < p \leq 1$. For this purpose, we use the constant τ_0 of (2.2).

Theorem 2.5.1. *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 $\{\alpha_i\}_{i=0}^K$, $\mathcal{X}$ is a data set identically and independently sampled from μ and $0 < p \leq 1$, then whenever*

$$\alpha_0 < \tau_0 \cdot \min_{i=1,\cdots,K} \alpha_i \cdot \left(1 \wedge \min_{1\leq i,j\leq K} \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_i^*, \mathrm{L}_j^*)^p/2^p\right), \tag{2.5}$$

the set $\{\mathrm{L}_1^*, \cdots, \mathrm{L}_K^*\}$ minimizes the energy (1.2) among all d-subspaces in $\mathbb{R}^D$ with overwhelming probability.

Theorem 2.5.1 extends to spherically symmetric HLM measures with restricted noise level in the following way:

Theorem 2.5.2. *Let $\epsilon > 0$, μ_ϵ a spherically symmetric HLM measure of noise level ϵ on $\mathbb{R}^D$ with K d-subspaces, $\{\mathrm{L}_i^*\}_{i=1}^K \subseteq \mathbb{R}^D$ and mixture coefficients $\{\alpha_i\}_{i=0}^K$ and let $\mathcal{X}$ be a data set sampled identically and independently from μ_ϵ. If $0 < p \leq 1$ and*

$$\epsilon < 3^{-\frac{1}{p}} \left(\tau_0 \cdot \min_{i=1,\cdots,K} \alpha_i \cdot \left(1 \wedge \min_{1\leq i,j\leq K} \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_i^*, \mathrm{L}_j^*)^p/2^p\right) - \alpha_0\right)^{\frac{1}{p}}, \tag{2.6}$$

then the minimizer of (1.2) in $\mathrm{G}(D,d)^K$ has a distance smaller than

$$f \equiv f(\epsilon, K, d, p, \{\alpha_i\}_{i=1}^K) = 3^{\frac{1}{p}} \cdot \left(\tau_0 \min_{1\leq j\leq K} \alpha_j - \alpha_0\right)^{-\frac{1}{p}} \cdot \epsilon \tag{2.7}$$

from one of the permutations of $(\mathrm{L}_1^*, \cdots, \mathrm{L}_K^*)$ with overwhelming probability.

If $p = 2$ and either there is a single underlying subspace (i.e., $K = 1$) or multiple centers (i.e., $d = 0$ and "affine 0-subspaces"), then despite the noise one can still recover exactly the underlying model as the number of data points approach infinity ([44, Section 11.6][45]). However, we show in Section 3.7 that in our general setting such results are impossible and thus Theorem 2.5.2 is still relevant, even though it is straightforward.

At last, we formulate the impossibility to recover the underlying d-subspaces by l_p minimization when $p > 1$ and consequently demonstrate a phase transition at $p = 1$ for multiple

subspaces recovery by l_p minimization. Here, we assume a weaker model, where spherical symmetry along the d-subspaces is unnecessary.

Theorem 2.5.3. *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 weak HLM measure with noise level ϵ and let $\mathcal{X}$ be a data set of N points sampled identically and independently from μ_ϵ. If $p > 1$ and $K > 1$, 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 $\epsilon < \delta_0$ the minimizer of (1.2), $\hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K$, satisfies w.o.p.:*

$$\mathrm{dist}_{\mathrm{G^K}}((\hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K), (\mathrm{L}_1^*, \ldots, \mathrm{L}_K^*)) > \kappa_0 \,.$$

We bound the constants δ_0 and κ_0 in Section 3.8.5

The above theorems have direct implications for HLM with spherically symmetric sampling along the subspaces. Theorems 2.5.1 and 2.5.2 clarify to some extent the robustness of two algorithms for HLM: MKF and LBF (using the l_1 energy (1.2)), which we will present in Chapter 4. Theorem 2.5.3 explains why common strategies that use the l_2 energy (1.2) (e.g., K-subspaces) are generally not robust to outliers.

Chapter 3

Verification of Theory

Section 3.1 reviews additional theory. In particular, Section 3.1 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 $0 < p < 1$); 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 3.1 also 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 main theory of this paper in Sections 3.3 to 3.7, while leaving some auxiliary verifications to the appendix.

3.1 Additional Theory

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 3.1.1 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 [46, 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.

Combinatorial Conditions for l_0 Subspaces being Local l_p Subspaces

Preliminary Notation

We denote the orthogonal group of $n \times n$ matrices by $\mathrm{O}(n)$ and the semigroup of $n \times n$ nonnegative diagonal 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

$$\mathbf{B}_{\mathrm{L},\mathcal{X}} = \sum_{\mathbf{x} \in \mathcal{X} \setminus \mathrm{L}} P_{\mathrm{L}}(\mathbf{x}) P_{\mathrm{L}}^{\perp}(\mathbf{x})^T / \mathrm{dist}(\mathbf{x}, \mathrm{L}). \tag{3.1}$$

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

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

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 3.1.1. *If $\mathrm{L}_1 \in \mathrm{G}(D,d)$, $\mathcal{X}_1 = \{\mathbf{x}_i\}_{i=1}^{N_1} \subset \mathrm{L}_1$, $\mathcal{X}_0 = \{\mathbf{y_i}\}_{i=1}^{N_0} \subset \mathbb{R}^D \setminus \mathrm{L}_1$ 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 $\mathbf{C} \in \mathrm{S}_+(d)$:*

$$\sum_{i=1}^{N_1} ||\mathbf{CV}P_{\mathrm{L}_1}(\mathbf{x}_i)|| > ||\mathbf{CVB}_{\mathrm{L}_1,\mathcal{X}}||_* \,. \tag{3.2}$$

Proposition 3.1.1. *If $\mathrm{L}_1 \in \mathrm{G}(D,d)$, $\mathcal{X}_1 = \{\mathbf{x}_i\}_{i=1}^{N_1} \subset \mathrm{L}_1$, $\mathcal{X}_0 = \{\mathbf{y_i}\}_{i=1}^{N_0} \subset \mathbb{R}^D \setminus \mathrm{L}_1$, $\mathrm{Sp}(\{\mathbf{x}_i\}_{i=1}^{N_1}) = \mathrm{L}_1$ 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 3.1.2. *If $\mathrm{L}_1 \in \mathrm{G}(D,d)$, $\mathcal{X}_1 = \{\mathbf{x}_i\}_{i=1}^{N_1} \subset \mathrm{L}_1$, $\mathcal{X}_0 = \{\mathbf{y_i}\}_{i=1}^{N_0} \subset \mathbb{R}^D \setminus \mathrm{L}_1$ 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*

$$\sum_{i=1}^{N_0} P_{\mathrm{L}_1}(\mathbf{y}_i) P_{\mathrm{L}_1}^{\perp}(\mathbf{y}_i)^T \mathrm{dist}(\mathbf{y}_i,\mathrm{L}_1)^{p-2} = 0. \tag{3.3}$$

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.

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 3.2.

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 3.1.2. *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*

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

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

$$1 - 2d^2 \exp\left(-\frac{\delta_*(\mu_1)^2 N_1}{72\, d^2}\right) - 2dD \exp\left(-\frac{\delta_*(\mu_1)^2 N_1^2}{8\, d^2 D N_0}\right). \tag{3.4}$$

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 $\mathrm{L}_1 \cap \mathrm{B}(\mathbf{0}, 1)$:

$$\delta_*(\mu_1) = 1/(d + 2). \tag{3.5}$$

For $0 < p < 1$, Proposition 3.1.1 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 3.1.3. *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 (A.4), 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 3.1.3 and Theorem 2.4.1 already answered this question in our setting. Indeed, if $p > 1$, then by Proposition 3.1.3 the global l_0 subspace is a global l_p subspace with probability 0; whereas if $0 < p \leq 1$, then Theorem 2.4.1 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 (3.96) holds w.p. 0 for $p > 1$ and any finite sample; however, the LHS of (3.96) 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 2.4.2 that this formal phase transition also breaks down with noise. Nevertheless, as we show in Theorem 2.4.3, 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 $K = 1$.

3.2 Verification of the Additional Theory

We describe here the complete proofs of the various theorems and propositions in Section 3.1.

Preliminaries

Auxiliary Lemmata

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

Lemma 3.2.1. *Suppose that* $\mathrm{L}_1, \hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K \in \mathrm{G}(D, d)$*,* $p > 0$ *and* μ_1 *is a spherically symmetric distribution in* $\mathrm{B}(\mathbf{0}, 1) \cap \mathrm{L}_1$*. If* $\min_{1 \leq j \leq K} \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1, \hat{\mathrm{L}}_j) > \epsilon$*, then*

$$E_{\mu_1}\left(e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K)\right) > \tau_0\, \epsilon^p.$$

Lemma 3.2.2. *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 3.2.3. *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* $\hat{\mathrm{L}} \in \mathrm{G}(D,d)$*:*

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}_1 \in \mu_1}(\mathrm{dist}(\mathbf{x}_1, \hat{\mathrm{L}})^p) + \mathbb{E}_{\mathbf{x}_2 \in \mu_2}(\mathrm{dist}(\mathbf{x}_2, \hat{\mathrm{L}})^p) \\ \geq &\mathbb{E}_{\mathbf{x}_1 \in \mu_1}(\mathrm{dist}(\mathbf{x}_1, \mathrm{L}_i)^p) + \mathbb{E}_{\mathbf{x}_2 \in \mu_2}(\mathrm{dist}(\mathbf{x}_2, \mathrm{L}_i)^p) \ \text{ for } i = 1, 2. \end{aligned} \tag{3.6}$$

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

We denote the principal angles [47] 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

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

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:

$$\begin{cases} \mathbf{v}'_i = \cos(\theta_i)\mathbf{v}_i + \sin(\theta_i)\mathbf{u}_i, & i = 1, 2, \cdots, k, \\ \mathbf{u}_i = \mathbf{v}_i, & i = k+1, \cdots, d. \end{cases} \tag{3.7}$$

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 $\mathrm{F} + \mathrm{G}$ 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 [48, 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 [49, 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 [48, 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 [49, 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:

$$\mathrm{L}(t) = \mathrm{Sp}(\{\cos(t\theta_i)\mathbf{v}_i + \sin(t\theta_i)\mathbf{u}_i\}_{i=1}^d). \tag{3.8}$$

We remark that this formula only holds when equipping the Grassmannian with the distance dist_G of (2.1) and this is the reason why we use this distance.

Proof of Theorem 3.1.1

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 [48, 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 (3.8), 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 $i = 1, \ldots, N_1$:

$$\begin{aligned}\operatorname{dist}(\mathbf{y}_i, \mathrm{L}(t)) &= \sqrt{\sum_{j=1}^{d} \operatorname{dist}^2(P_j(\mathbf{y}_i), \mathrm{L}(t)) + \operatorname{dist}^2(P^\perp(\mathbf{y}_i), \mathrm{L}(t))} \\ &= \sqrt{\sum_{j=1}^{d} \left((-\sin(t\theta_j)\mathbf{v}_j + \cos(t\theta_j)\mathbf{u}_j) \cdot \mathbf{y}_i\right)^2 + \operatorname{dist}^2(P^\perp(\mathbf{y}_i), \mathrm{L}(t))}. \end{aligned} \tag{3.9}$$

Notice that $\operatorname{dist}^2(P^\perp(\mathbf{y}_i), \mathrm{L}(t))$ is independent of t, we obtain the following expression for the derivative of $\operatorname{dist}(\mathbf{y}_i, \mathrm{L}(t))$ for all $1 \le i \le N_0$:

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\left(\operatorname{dist}(\mathbf{y}_i, \mathrm{L}(t))\right) \\ =& -\frac{\sum_{j=1}^{d} \theta_j \left((\cos(t\theta_j)\mathbf{v}_j + \sin(t\theta_j)\mathbf{u}_j) \cdot \mathbf{y}_i\right)\left((-\sin(t\theta_j)\mathbf{v}_j + \cos(t\theta_j)\mathbf{u}_j) \cdot \mathbf{y}_i\right)}{\operatorname{dist}(\mathbf{y}_i, \mathrm{L}(t))}. \end{aligned} \tag{3.10}$$

At $t = 0$ it becomes

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\operatorname{dist}(\mathbf{y}_i, \mathrm{L}(t))\right)\Big|_{t=0} &= -\frac{\sum_{j=1}^{d} \theta_j (\mathbf{v}_j \cdot \mathbf{y}_i)(\mathbf{u}_j \cdot \mathbf{y}_i)}{\operatorname{dist}(\mathbf{y}_i, \mathrm{L}(0))} \\ &= -\frac{\sum_{j=1}^{k} \theta_j (\mathbf{v}_j \cdot \mathbf{y}_i)(\mathbf{u}_j \cdot \mathbf{y}_i)}{\operatorname{dist}(\mathbf{y}_i, \mathrm{L}(0))}, \end{aligned} \tag{3.11}$$

where the interaction dimension $k = k(\mathrm{L}_1, \hat{\mathrm{L}})$ has been introduced in Section 3.2.

We form the following matrices: $\mathbf{C} = \operatorname{diag}(\theta_1, \theta_2, \cdots, \theta_d)$, $\mathbf{V} \in \mathrm{O}(d, D)$ 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 (3.11) using these matrices as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\operatorname{dist}(\mathbf{y}_i, \mathrm{L}(t))\right)\Big|_{t=0} = -\frac{\operatorname{tr}_k(\mathbf{C}\mathbf{V}\mathbf{y}_i\mathbf{y}_i^T\mathbf{U}^T)}{\operatorname{dist}(\mathbf{y}_i, \mathrm{L}_1)}, \tag{3.12}$$

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 $\mathbf{x}_i \in \mathrm{L}_1$, $i = 1, 2, \cdots, N_1$,

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

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\operatorname{dist}(\mathbf{x}_i, \mathrm{L}(t))\right) = \frac{\sum_{j=1}^{d} \theta_j |\mathbf{v}_j \cdot \mathbf{x}_i|^2 \sin(t\theta_j)\cos(t\theta_j)}{\operatorname{dist}(\mathbf{x}_i, \mathrm{L}(t))}. \tag{3.13}$$

At $t = 0$, this derivative becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{dist}(\mathbf{x}_i, \mathrm{L}(t))\right)\Big|_{t=0} = \sqrt{\sum_{j=1}^{d} |(\mathbf{v}_j \cdot \mathbf{x}_i)|^2 \, \theta_j^2} = ||\mathbf{C}\mathbf{V}\mathbf{x}_i||. \tag{3.14}$$

Combining (3.12) and (3.14) and using

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

we obtain the following expression for the derivative of the l_1 energy of (1.1):

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e_{l_1}(\mathcal{X}, \mathrm{L}(t))\right)\Big|_{t=0} = \sum_{i=1}^{N_1} ||\mathbf{C}\mathbf{V}\mathbf{x}_i|| - \mathrm{tr}_k(\mathbf{C}\mathbf{V}\mathbf{A}\mathbf{U}^T). \tag{3.15}$$

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 $P_{\mathrm{L}_1}^{\perp}(\mathbf{v}_j)^T$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e_{l_1}(\mathcal{X}, \mathrm{L}(t))\right)\Big|_{t=0} = \sum_{i=1}^{N_1} ||\mathbf{C}\hat{\mathbf{V}}P_{\mathrm{L}_1}\mathbf{x}_i|| - \mathrm{tr}_k(\mathbf{C}\hat{\mathbf{V}}\mathbf{B}_{\mathrm{L}_1,\mathcal{X}}\hat{\mathbf{U}}^T). \tag{3.16}$$

At last, we note that

$$\max_{\hat{\mathbf{U}}^T}(\mathrm{tr}_k(\mathbf{C}\hat{\mathbf{V}}\mathbf{B}_{\mathrm{L}_1,\mathcal{X}}\hat{\mathbf{U}}^T)) = ||\mathbf{C}\hat{\mathbf{V}}\mathbf{B}_{\mathrm{L}_1,\mathcal{X}}||_*. \tag{3.17}$$

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 (3.16) and (3.17).

Simultaneous Proof for Both Propositions 3.1.1 and 3.1.2

For the d-subspace L_1 and an arbitrary d-subspace $\hat{\mathrm{L}} \in \mathrm{B}_{\mathrm{G}}(\mathrm{L}_1, 1)$, 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 3.1.1. Similarly to verifying (3.12) and (3.14) in the latter proof, we obtain that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{dist}(\mathbf{y}_i, \mathrm{L}(t))^p\right)\Big|_{t=0} = -p\,\mathrm{dist}(\mathbf{y}_i, \mathrm{L}_1)^{p-2}\,\mathrm{tr}_k(\mathbf{C}\mathbf{V}\mathbf{y}_i\mathbf{y}_i^T\mathbf{U}^T) \tag{3.18}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{dist}(\mathbf{x}_i,\mathrm{L}(t))^p\right)\Big|_{t=0} = p\,\mathrm{dist}(\mathbf{x}_i,\mathrm{L}_1)^{p-1}||\mathbf{C}\mathbf{V}\mathbf{x}_i||. \tag{3.19}$$

Consequently

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(e_{l_p}(\mathcal{X},\mathrm{L}(t))\right)\Big|_{t=0} &= p\sum_{i=1}^{N_1}\mathrm{dist}(\mathbf{x}_i,\mathrm{L}_1)^{p-1}||\mathbf{C}\mathbf{V}\mathbf{x}_i|| \qquad (3.20)\\ -p\sum_{i=1}^{N_0}\mathrm{dist}(\mathbf{y}_i,\mathrm{L}_1)^{p-2}\,\mathrm{tr}_k(\mathbf{C}\mathbf{V}\mathbf{y}_i\mathbf{y}_i^T\mathbf{U}^T) &= p\sum_{i=1}^{N_1}\mathrm{dist}(\mathbf{x}_i,\mathrm{L}_1)^{p-1}||\mathbf{C}\hat{\mathbf{V}}P_{\mathrm{L}_1}(\mathbf{x}_i)||\\ -p\sum_{i=1}^{N_0}\mathrm{dist}(\mathbf{y}_i,\mathrm{L}_1)^{p-2}\,\mathrm{tr}_k(\mathbf{C}\hat{\mathbf{V}}P_{\mathrm{L}_1}(\mathbf{y}_i)P_{\mathrm{L}_1}^{\perp}(\mathbf{y}_i)^T\hat{\mathbf{U}}^T).&\end{aligned}$$

Assume first that $p<1$. Then

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t^p}\left(e_{l_p}(\mathcal{X},\mathrm{L}(t))\right)\Big|_{t=0} &= p\,t^{1-p}\sum_{i=1}^{N_1}\mathrm{dist}(\mathbf{x}_i,\mathrm{L}_1)^{p-1}||\mathbf{C}\hat{\mathbf{V}}P_{\mathrm{L}_1}(\mathbf{x}_i)|| \qquad (3.21)\\ -p\,t^{1-p}\sum_{i=1}^{N_0}\mathrm{dist}(\mathbf{y}_i,\mathrm{L}_1)^{p-2}\,\mathrm{tr}_k(\mathbf{C}\hat{\mathbf{V}}P_{\mathrm{L}_1}(\mathbf{y}_i)P_{\mathrm{L}_1}^{\perp}(\mathbf{y}_i)^T\hat{\mathbf{U}}^T)&\\ = p\sum_{i=1}^{N_1}\left(\lim_{t\to 0}\mathrm{dist}(\mathbf{x}_i,\mathrm{L}(t))/t\right)^{p-1}||\mathbf{C}\hat{\mathbf{V}}P_{\mathrm{L}_1}(\mathbf{x}_i)|| &= \sum_{i=1}^{N_0}||\mathbf{C}\hat{\mathbf{V}}P_{\mathrm{L}_1}(\mathbf{x}_i)||^p.\end{aligned}$$

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

$$||\mathbf{C}\mathbf{V}\mathbf{x}_i|| \geq \theta_1\,||\mathbf{v}_1^T\mathbf{x}_i||. \tag{3.22}$$

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, (3.21) is positive, L_1 is a local minimum of $e_{l_p}(\mathcal{X},\mathrm{L}(t))$ and Proposition 3.1.1 is proved.

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

$$p\sum_{i=1}^{N_1}\mathrm{dist}(\mathbf{x}_i,\mathrm{L}_1)^{p-1}\|\mathbf{C}\hat{\mathbf{V}}P_{\mathrm{L}_1}\mathbf{x}_i\| = 0. \tag{3.23}$$

Since L_1 is a local minimum of $e_{l_p}(\mathcal{X},\mathrm{L})$, the derivative in (3.20) is nonnegative and in view of (3.23), the subtracted term in (3.20) 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

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

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

where the last equality follows from (3.17). Therefore, (3.96) holds and Proposition 3.1.2 is thus proved.

Proof of Theorem 3.1.2: Combination of Combinatorial Estimates 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 (3.2) for arbitrary $\hat{\mathrm{L}} \in \mathrm{B_G}(\mathrm{L}_1, 1)$. We use the similar notation as in the proof of Theorem 3.1.1, 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 (3.2) 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 (3.2) 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

$$\Pr\left(\frac{\max_{p,l} |(\mathbf{B}_{\mathrm{L}_1,\mathcal{X}})_{p,l}|}{N_0} < \frac{\epsilon}{d\sqrt{D}}\right) \geq 1 - 2dD \exp\left(-\frac{N_0 \epsilon^2}{2d^2 D}\right). \tag{3.24}$$

Next, we estimate the probability that the LHS of (3.2) 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

$$\sum_{i=1}^{N_1} ||\mathbf{CV} P_{\mathrm{L}_1}(\mathbf{x}_i)|| \geq \sum_{i=1}^{N_1} |\theta_1 \mathbf{v}_1^T P_{\mathrm{L}_1}(\mathbf{x}_i)| = \sum_{i=1}^{N_1} |\mathbf{v}_1^T P_{\mathrm{L}_1}(\mathbf{x}_i)|$$
$$\geq \sqrt{\sum_{i=1}^{N_1} |\mathbf{v}_1^T P_{\mathrm{L}_1}(\mathbf{x}_i)|^2} \geq \min_t \sigma_t \left(\sum_{i=1}^{N_1} P_{\mathrm{L}_1}(\mathbf{x}_i) P_{\mathrm{L}_1}(\mathbf{x}_i)^T \right). \quad (3.25)$$

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

$$E_{\mu_1}(P_{\mathrm{L}_1}(\mathbf{x}) P_{\mathrm{L}_1}(\mathbf{x})^T) = \delta_* \mathbf{I}_d, \text{ where } \delta_* = \delta_*(\mu_1) \text{ depends on } \mu_1. \quad (3.26)$$

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

$$\text{If } \max_t \sigma_t \left(\sum_{i=1}^{N_1} P_{\mathrm{L}_1}(\mathbf{x}_i) P_{\mathrm{L}_1}(\mathbf{x}_i)^T - \delta_* \mathbf{I}_d \right) < \eta,$$
$$\text{then } \min_t \sigma_t \left(\sum_{i=1}^{N_1} P_{\mathrm{L}_1}(\mathbf{x}_i) P_{\mathrm{L}_1}(\mathbf{x}_i)^T \right) > \delta_* - \eta. \quad (3.27)$$

We combine (3.25)-(3.27) and Hoeffding's inequality to obtain the following probabilistic estimate for the LHS of (3.2):

$$\Pr\left(\frac{\sum_{i=1}^{N_1} ||\mathbf{CV} P_{\mathrm{L}_1}(\mathbf{x}_i)||}{N_1} > \delta_* - \eta \right) \quad (3.28)$$
$$\geq \Pr\left(\min_t \sigma_t \left(\frac{\sum_{i=1}^{N_1} P_{\mathrm{L}_1}(\mathbf{x}_i) P_{\mathrm{L}_1}(\mathbf{x}_i)^T}{N_1} \right) > \delta_* - \eta \right)$$
$$\geq \Pr\left(\max_t \sigma_t \left(\frac{\sum_{i=1}^{N_1} P_{\mathrm{L}_1}(\mathbf{x}_i) P_{\mathrm{L}_1}(\mathbf{x}_i)^T}{N_1} - \delta_* \mathbf{I}_d \right) < \eta \right)$$
$$\geq \Pr\left(\left\| \frac{\sum_{i=1}^{N_1} P_{\mathrm{L}_1}(\mathbf{x}_i) P_{\mathrm{L}_1}(\mathbf{x}_i)^T}{N_1} - \delta_* \mathbf{I}_d \right\|_F < \eta \right)$$
$$\geq \Pr\left(\max_{p,l} \left| \frac{\sum_{i=1}^{N_1} P_{\mathrm{L}_1}(\mathbf{x}_i) P_{\mathrm{L}_1}(\mathbf{x}_i)^T}{N_1} - \delta_* \mathbf{I}_d \right|_{p,l} < \frac{\eta}{d} \right) \geq 1 - 2d^2 \exp\left(-\frac{N_1 \eta^2}{2d^2} \right).$$

From (3.24) and (3.28), (3.2) is valid with probability at least

$$1 - 2d^2 \exp\left(-\frac{N_1 \eta^2}{2d^2} \right) - 2dD \exp\left(-\frac{N_0 \epsilon^2}{2d^2 D} \right) \quad \forall\, \epsilon,\ \eta \text{ s.t. } \eta + \frac{N_0}{N_1} \epsilon < \delta_*(\mu_1). \quad (3.29)$$

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 (3.2) is valid with the probability specified in (3.4).

3.3 Proof of Theorem 2.4.1: From Local Probabilistic Estimates to Global Ones

Proof of the Special Case: $K = 1$

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

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 $\mathrm{B}_{\mathrm{G}}(\mathrm{L}_1, \gamma_1)$. We arbitrarily choose $\hat{\mathrm{L}} \in \mathrm{G}(D, d)$ such that $\mathrm{dist}_{\mathrm{G}}(\hat{\mathrm{L}}, \mathrm{L}_1) = 1$ and parameterize a geodesic line from L_1 to $\hat{\mathrm{L}}$ by a function L: [0,1] $\rightarrow$ $\mathrm{G}(D, d)$, 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] $\rightarrow \mathbb{R}$ of (1.1) has a positive derivative w.o.p. at any $t \in [0, \gamma_1]$, that is,

$$\frac{\mathrm{d}}{\mathrm{d}t^p}\left(\frac{\sum_{\mathbf{x}\in\mathcal{X}} \mathrm{dist}(\mathbf{x}, \mathrm{L}(t))^p}{N}\right) > 0 \text{ for all } t \in [0, \gamma_1] \text{ w.o.p.} \tag{3.30}$$

We will deduce (3.30) from the following two equations:

$$\frac{\mathrm{d}}{\mathrm{d}t^p}\left(\frac{\sum_{\mathbf{x}\in\mathcal{X}} \mathrm{dist}(\mathbf{x}, \mathrm{L}(t))^p}{N}\right)\Bigg|_{t=0} > \gamma_2 \text{ w.o.p. for some } \gamma_2 > 0. \tag{3.31}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t^p}\left(\frac{\sum_{\mathbf{x}\in\mathcal{X}} \mathrm{dist}(\mathbf{x}, \mathrm{L}(t))^p}{N}\right)\Bigg|_{t=0} - \frac{\mathrm{d}}{\mathrm{d}t^p}\left(\frac{\sum_{\mathbf{x}\in\mathcal{X}} \mathrm{dist}(\mathbf{x}, \mathrm{L}(t))^p}{N}\right)\Bigg|_{t=t_0} < \frac{\gamma_2}{2}, \tag{3.32}$$

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

When $p = 1$, equation (3.31) practically follows from the proof of Theorem 3.1.2 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 3.1.2) the ratio of sampled outliers to inliers, N_0/N_1, goes w.o.p. to α_0/α_1. When $p < 1$, equation (3.31) follows from (3.21). We also observe that $\gamma_2 \equiv \gamma(\alpha_0, \alpha_1, d, \mu_1, p)$.

We first verify (3.32) for the sum of elements in $\mathcal{X}_1 = \mathcal{X} \cap \mathrm{L}_1$. In view of (3.13), 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

$$\frac{\mathrm{d}}{\mathrm{d}t^p}\left(\frac{\sum_{\mathbf{x}\in\mathcal{X}_1} \mathrm{dist}(\mathbf{x}, \mathrm{L}(t))^p}{N}\right)\Bigg|_{t=0} - \frac{\mathrm{d}}{\mathrm{d}t^p}\left(\frac{\sum_{\mathbf{x}\in\mathcal{X}_1} \mathrm{dist}(\mathbf{x}, \mathrm{L}(t))^p}{N}\right)\Bigg|_{t=t_0} < \frac{\gamma_2}{6} \tag{3.33}$$

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

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: $\hat{\mathcal{X}} := \{\mathbf{x} \in \mathcal{X} \setminus \mathcal{X}_1 : \mathrm{dist}(\mathbf{x}, \mathrm{L}_1) \leq 2\,\gamma_3\}$ and $(\mathcal{X} \setminus \mathcal{X}_1) \setminus \hat{\mathcal{X}}$.

In order to deal with the first sum, we define

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

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:

$$\frac{\#(\hat{\mathcal{X}})}{\#(\mathcal{X})} \leq 2\gamma_4 \ \text{ w.o.p.} \tag{3.34}$$

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 (3.34).

Now for $\mathbf{y}_i \in \hat{\mathcal{X}}$, the derivatives expressed in (3.10) and (3.21) are bounded by 1 since the support of μ_0 is contained in $\mathrm{B}(\mathbf{0}, 1)$. Thus, by combining this observation with (3.34) we obtain that there exists γ_3 and γ_4 such that for any $t_0 \in [0, \gamma_1]$:

$$\frac{\mathrm{d}}{\mathrm{d}t^p}\left(\frac{\sum_{\mathbf{x} \in \hat{\mathcal{X}}} \mathrm{dist}(\mathbf{x}, \mathrm{L}(t))^p}{N}\right)\bigg|_{t=0} - \frac{\mathrm{d}}{\mathrm{d}t^p}\left(\frac{\sum_{\mathbf{x} \in \hat{\mathcal{X}}} \mathrm{dist}(\mathbf{x}, \mathrm{L}(t))^p}{N}\right)\bigg|_{t=t_0} < \frac{\gamma_2}{6} \tag{3.35}$$

w.o.p.

Differentiating (3.10) and (3.21) one more time, we obtain that for $\mathbf{x} \in (\mathcal{X} \setminus \mathcal{X}_1) \setminus \hat{\mathcal{X}}$, the second derivative of $\mathrm{dist}(\mathbf{x}, \mathrm{L}(t))$ 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 $t_0 \in [0, \gamma_1]$:

$$\frac{\mathrm{d}}{\mathrm{d}t^p}\frac{\sum_{\mathbf{x} \in (\mathcal{X} \setminus \mathcal{X}_1) \setminus \hat{\mathcal{X}}} \mathrm{dist}(\mathbf{x}, \mathrm{L}(t))^p}{N}\bigg|_{t=0} - \frac{\mathrm{d}}{\mathrm{d}t^p}\frac{\sum_{\mathbf{x} \in (\mathcal{X} \setminus \mathcal{X}_1) \setminus \hat{\mathcal{X}}} \mathrm{dist}(\mathbf{x}, \mathrm{L}(t))^p}{N}\bigg|_{t=t_0} < \frac{\gamma_2}{6}. \tag{3.36}$$

Equation (3.32) and consequently (3.30) are thus verified by combing (3.33), (3.35) and (3.36). That is, we showed that L_1 is the global l_p subspace in $\mathrm{B}_{\mathrm{G}}(\mathrm{L}_1, \gamma_1)$ 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 $\mathrm{L} \in \mathrm{G}(D,d) \setminus \mathrm{B}_{\mathrm{G}}(\mathrm{L}_1, \gamma_1)$ and any fixed $p \leq 1$, there exists some $\gamma_7 > 0$ such that

$$e_{l_p}(\mathcal{X}, \mathrm{L}) - e_{l_p}(\mathcal{X}, \mathrm{L}_1) > \gamma_7 N, \quad \text{w.o.p.} \tag{3.37}$$

Indeed, we first conclude from Lemma 3.2.1 that

$$\begin{aligned} &E_{\mu}\left(e_{l_p}(\mathbf{x}, \mathrm{L})\right) - E_{\mu}\left(e_{l_p}(\mathbf{x}, \mathrm{L}_1)\right) > \alpha_0 \left(E_{\mu_0}\left(e_{l_p}(\mathbf{x}, \mathrm{L})\right) - E_{\mu_0}\left(e_{l_p}(\mathbf{x}, \mathrm{L}_1)\right)\right) \\ &+ \alpha_1 \left(E_{\mu_1}\left(e_{l_p}(\mathbf{x}, \mathrm{L})\right) - E_{\mu_1}\left(e_{l_p}(\mathbf{x}, \mathrm{L}_1)\right)\right) \geq \frac{\alpha_1(1-\mu_1(\{\mathbf{0}\}))2^{p-1}\gamma_1^p}{(\pi\sqrt{d})^p \left(\psi_{\mu_1}^{-1}\left(\frac{1+\mu_1(\{\mathbf{0}\})}{2}\right)\right)^p}. \end{aligned} \tag{3.38}$$

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 (3.38) with Hoeffding's inequality, we obtain (3.37).

Now, (3.37) extends for a small neighborhood of L. That is, for any $\mathrm{L} \in \mathrm{G}(D,d)$ we can find a ball $\mathrm{B}_{\mathrm{G}}(\mathrm{L}, t)$ 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 $\mathrm{G}(D,d) \setminus \mathrm{B}_{\mathrm{G}}(\mathrm{L}_1, \gamma_1)$ with finite number of such balls we obtain that w.o.p. L_1 is the global l_p subspace in $\mathrm{G}(D,d) \setminus \mathrm{B}_{\mathrm{G}}(\mathrm{L}_1, \gamma_1)$. Combining this observation with part I, we conclude that w.o.p. L_1 is the global l_p subspace in $\mathrm{G}(D,d)$.

Extension of the Proof to $K > 1$

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

We maintain the same notation of Section 3.3, 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 $\mathrm{B}_{\mathrm{G}}(\mathrm{L}_1, \gamma_1)$, where γ_1 is a sufficiently small constant different than the one of Section 3.3.

In order to do so, we arbitrarily fix $\hat{\mathrm{L}} \in \mathrm{G}(D,d)$ such that $\mathrm{dist}_{\mathrm{G}}(\hat{\mathrm{L}}, \mathrm{L}_1) = 1$ (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

$$\frac{\mathrm{d}}{\mathrm{d}t^p}\left(\frac{\sum_{\mathbf{x}\in\mathcal{X}} \mathrm{dist}(\mathbf{x}, \mathrm{L}(t))^p}{N}\right) > 0 \quad \text{for all } t \in (0, \gamma_1). \tag{3.39}$$

First of all, we prove that the LHS of (3.39) is larger than some constant $\gamma_2 > 0$ at $t = 0$ w.o.p., that is:

$$\frac{\mathrm{d}}{\mathrm{d}t^p}\left(\frac{\sum_{\mathbf{x}\in\mathcal{X}}\mathrm{dist}(\mathbf{x},\mathrm{L}(t))^p}{N}\right)\Big|_{t=0} > \gamma_2 \quad \text{w.o.p.} \tag{3.40}$$

When $0 < p < 1$, it follows from (3.21) and Hoeffding's inequality that (3.40) is valid w.p. $1 - \exp(-2N\gamma_2^2)$ for $\gamma_2 = \alpha_1\, E_{\mu_0}(\|\mathbf{C}_0\mathbf{V}_0 P_{\mathrm{L}_1}(\mathbf{x})\|^p)/2$. When $p = 1$, it follows from (3.2) that this probability is the same as the probability of the event

$$\frac{\sum_{\mathbf{x}\in\mathcal{X}_1}||\mathbf{CV}P_{\mathrm{L}_1}(\mathbf{x})|| - ||\mathbf{CVB}_{\mathrm{L}_1,\mathcal{X}\setminus\mathcal{X}_1}||}{N} > \gamma_2 \tag{3.41}$$
$$\forall \mathbf{C} \in \mathrm{NS}_+(d) \text{ and } \mathbf{V} \in \mathrm{O}(d).$$

If we define the random variable $J(\mathbf{x}) = I(\mathbf{x} \in \mathrm{X}_0)\, P_{\mathrm{L}_1}(\mathbf{x})P_{\mathrm{L}_1}^{\perp}(\mathbf{x})^T/\mathrm{dist}(\mathbf{x},\mathrm{L}_1)$, then every element of $J(\mathbf{x})$ is bounded by $[-1,1]$, and using the symmetry of μ_0, we have

$$2E(J(\mathbf{x})) = E(J(\mathbf{x})) + E(J(P_{\mathrm{L}_1}(\mathbf{x}) - P_{\mathrm{L}_1}^{\perp})) = E(J(bx) + J(P_{\mathrm{L}_1}(\mathbf{x}) - P_{\mathrm{L}_1}^{\perp})) = \mathbf{0}.$$

Therefore, combining the fact that

$$\mathbf{D}_{i,j} = \mathbf{e}_i^T\mathbf{D}\mathbf{e}_j \le \max_{\mathbf{u},\mathbf{v}\in\mathbb{R}^D} \mathbf{u}^T\mathbf{D}\mathbf{v}/\|\mathbf{u}\|\|\mathbf{v}\| = \|\mathbf{D}\|_*,$$

for any $\mathbf{D} \in \mathbb{R}^{D\times D}$ and $1 \le i, j \le N$, and Hoeffding's inequality on random variable $J(\mathbf{x})$, we have

$$\begin{aligned}&\Pr\left(\|\sum_{\mathbf{x}\in\mathcal{X}_0} P_{\mathrm{L}_1}(\mathbf{x})P_{\mathrm{L}_1}^{\perp}(\mathbf{x})^T/\mathrm{dist}(\mathbf{x},\mathrm{L}_1)\|_*/N < 2\gamma_2\right)\\ \ge &\Pr\left(\|\sum_{\mathbf{x}\in\mathcal{X}_0} P_{\mathrm{L}_1}(\mathbf{x})P_{\mathrm{L}_1}^{\perp}(\mathbf{x})^T/\mathrm{dist}(\mathbf{x},\mathrm{L}_1)\|_\infty/N < 2\gamma_2\right) \ge 1 - 2\,D^2\exp(2\gamma_2^2\,N).\end{aligned} \tag{3.42}$$

For all $\mathbf{C} \in \mathrm{NS}_+(d)$ and $\mathbf{V} \in \mathrm{O}(d)$, we have that:

$$\begin{aligned}&||\mathbf{CVB}_{\mathrm{L}_1,\mathcal{X}\setminus\{\mathcal{X}_1\cup\mathcal{X}_0\}}\|_* = \|\mathbf{CV}\sum_{\mathbf{x}\in\mathcal{X}\setminus\{\mathcal{X}_1\cup\mathcal{X}_0\}} P_{\mathrm{L}_1}(\mathbf{x})P_{\mathrm{L}_1}^{\perp}(\mathbf{x})^T/\mathrm{dist}(\mathbf{x},\mathrm{L}_1)\|_*\\ \le &\sum_{\mathbf{x}\in\mathcal{X}\setminus\{\mathcal{X}_1\cup\mathcal{X}_0\}} ||\mathbf{CV}P_{\mathrm{L}_1}(\mathbf{x})P_{\mathrm{L}_1}^{\perp}(\mathbf{x})^T/\|P_{\mathrm{L}_1}^{\perp}(\mathbf{x})\|\,\|_* \le \sum_{\mathbf{x}\in\mathcal{X}\setminus\{\mathcal{X}_1\cup\mathcal{X}_0\}} ||\mathbf{CV}P_{\mathrm{L}_1}(\mathbf{x})||.\end{aligned} \tag{3.43}$$

Then we arbitrarily fix $\mathbf{C}_0 \in \mathrm{NS}_+(d)$, $\mathbf{V}_0 \in \mathrm{O}(d)$ and verify (3.40) by Hoeffding's inequality in the following way. We define the random variable $J(\mathbf{x}) = (I(\mathbf{x} \in \mathcal{X}_1) - I(\mathbf{x} \in \mathcal{X} \setminus \{\mathcal{X}_1 \cup \mathcal{X}_0\}))||\mathbf{C}_0\mathbf{V}_0 P_{\mathrm{L}_1}(\mathbf{x})||$ and using the spherical symmetry of $\{\mu_i\}_{i=1}^K$, we have

$$
\begin{aligned}
E_\mu(J(\mathbf{x})) &= E_{\mu^N}\left(\frac{\sum_{\mathbf{x}\in\mathcal{X}_1}||\mathbf{C}_0\mathbf{V}_0P_{\mathrm{L}_1}(\mathbf{x})|| - \sum_{\mathbf{x}\in\mathcal{X}\setminus\{\mathcal{X}_1\cup\mathcal{X}_0\}}||\mathbf{C}_0\mathbf{V}_0P_{\mathrm{L}_1}(\mathbf{x})||}{N}\right) \qquad (3.44)\\
&= \alpha_1 E_{\mu_1}||\mathbf{C}_0\mathbf{V}_0P_{\mathrm{L}_1}(\mathbf{x})|| - \sum_{j=2}^K \alpha_j E_{\mu_j}||\mathbf{C}_0\mathbf{V}_0P_{\mathrm{L}_1}(\mathbf{x})||\\
&\geq \alpha_1 E_{\mu_1}||\mathbf{C}_0\mathbf{V}_0P_{\mathrm{L}_1}(\mathbf{x})|| - \sum_{j=2}^K \alpha_j E_{\mu_1}||\mathbf{C}_0\mathbf{V}_0P_{\mathrm{L}_1}(\mathbf{x})||\\
&= \beta_0 E_{\mu_1}||\mathbf{C}_0\mathbf{V}_0P_{\mathrm{L}_1}(\mathbf{x})||,
\end{aligned}
$$

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

Now, let $\gamma_2 := \beta_0 E_{\mu_1}||\mathbf{C}_0\mathbf{V}_0P_{\mathrm{L}_1}(\mathbf{x})||/6$. We note that the random variable $J(\mathbf{x})$ has expectation larger than $6\gamma_2$ and it takes values in $[-1, 1]$; thus by Hoeffding's inequality:

$$
\begin{aligned}
&\frac{\sum_{\mathbf{x}\in\mathcal{X}_1}||\mathbf{C}_0\mathbf{V}_0P_{\mathrm{L}_1}(\mathbf{x})|| - \sum_{\mathbf{x}\in\mathcal{X}\setminus\{\mathcal{X}_1\cup\mathcal{X}_0\}}||\mathbf{C}_0\mathbf{V}_0P_{\mathrm{L}_1}(\mathbf{x})||}{N} > 4\gamma_2\\
&\text{w.p.} \geq 1 - \exp(-2N\gamma_2^2).
\end{aligned}
$$

Combining it with (3.42) and (3.43), we have

$$
\begin{aligned}
&\frac{\sum_{\mathbf{x}\in\mathcal{X}_1}||\mathbf{C}_0\mathbf{V}_0P_{\mathrm{L}_1}(\mathbf{x})|| - ||\mathbf{C}\mathbf{V}\mathbf{B}_{\mathrm{L}_1,\mathcal{X}\setminus\mathcal{X}_1}||_*}{N} > 2\gamma_2 \qquad (3.45)\\
&\text{w.p.} \geq 1 - (2D^2+1)\exp(-2N\gamma_2^2).
\end{aligned}
$$

We have thus proved that (3.40) 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 (3.40) 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

$$
\mathrm{dist}_{(\mathrm{NS}_+(d),\mathrm{O}(d))}((\mathbf{C}_1,\mathbf{V}_1),(\mathbf{C}_2,\mathbf{V}_2)) := \max(||\mathbf{C}_1-\mathbf{C}_2||_2, ||\mathbf{V}_1-\mathbf{V}_2||_2) \qquad (3.46)
$$

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

$$
||\mathbf{C}_1\mathbf{V}_1P_{\mathrm{L}_1}(\mathbf{x})|| - ||\mathbf{C}_2\mathbf{V}_2P_{\mathrm{L}_1}(\mathbf{x})||
$$

$$= (||\mathbf{C}_1\mathbf{V}_1 P_{\mathrm{L}_1}(\mathbf{x})|| - ||\mathbf{C}_2\mathbf{V}_1 P_{\mathrm{L}_1}(\mathbf{x})||) + (||\mathbf{C}_2\mathbf{V}_1 P_{\mathrm{L}_1}(\mathbf{x})|| - ||\mathbf{C}_2\mathbf{V}_2 P_{\mathrm{L}_1}(\mathbf{x})||)$$
$$\leq ||\mathbf{C}_1 - \mathbf{C}_2||_2 + ||\mathbf{C}_2||_2||\mathbf{V}_1 - \mathbf{V}_2||_2 \leq \gamma_2. \tag{3.47}$$

Combining (3.45) and (3.47) we obtain that for any ball in $\mathrm{G}(D,d)$ of radius $\gamma_2/2$ and center $(\mathbf{C}_0, \mathbf{V}_0)$:

$$\frac{\sum_{\mathbf{x}\in\mathcal{X}_1} ||\mathbf{C}\mathbf{V}P_{\mathrm{L}_1}(\mathbf{x})|| - \sum_{\mathbf{x}\in\mathcal{X}\setminus\mathcal{X}_1} ||\mathbf{C}\mathbf{V}P_{\mathrm{L}_1}(\mathbf{x})||}{N} > \gamma_2 \text{ w.p. } \geq 1 - \exp(-2N\gamma_2^2). \tag{3.48}$$

We easily extend (3.48) 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 (3.46)). Indeed, it follows from [50] 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,

(3.40) is valid for any $\mathbf{C} \in \mathrm{NS}_+(d)$ and $\mathbf{V} \in \mathrm{O}(d)$
$$\text{w.p. } 1 - C_1^{2d}\exp(-2N\gamma_2^2)/(\gamma_2/2)^{2d-1}. \tag{3.49}$$

Equation (3.39) follows w.o.p. from (3.40) in exactly the same way of deriving (3.30) from (3.31) and (3.32). We remark that (3.32), which is deterministic, easily extends to the current case. While we did not estimate the overwhelming probability for (3.30), it is easy to show that in the current case, (3.40) implies (3.39) w.p. $1 - \exp(-N\gamma_8)/\gamma_8$. Carrying this analysis, one notices that both γ_1 and γ_8 depend on $d, K, \alpha_0, \alpha_1, \mu_0, \mu_1, p$ and $\min_{2\leq i\leq K}(\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1, \mathrm{L}_i))$. Combining this with (3.49), we obtain that

L_1 is a global l_p subspace in $\mathrm{B}_{\mathrm{G}}(\mathrm{L}_1, \gamma_1)$
$$\text{w.p. } 1 - C_1^{2d}\exp(-2N\gamma_2^2)/(\gamma_2/2)^{2d-1} - \exp(-N\gamma_4)/\gamma_4. \tag{3.50}$$

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 3.2.3 we obtain that for all $2 \leq i \leq K$:

$$E_{\mu_1}\left(\mathrm{dist}(\mathbf{x}, \mathrm{L})^p - \mathrm{dist}(\mathbf{x}, \mathrm{L}_1)^p\right) + E_{\mu_i}(\mathrm{dist}(\mathbf{x}, \mathrm{L})^p - \mathrm{dist}(\mathbf{x}, \mathrm{L}_1)^p) \geq 0. \tag{3.51}$$

Further application of Lemma 3.2.1 with $\mathrm{L} \in \mathrm{G}(D,d) \setminus \mathrm{B}_{\mathrm{G}}(\mathrm{L}_1, \gamma_1)$ results in the inequality:

$$E_{\mu_1}(\mathrm{dist}(\mathbf{x}, \mathrm{L})) > \frac{(1 - \mu_1(\{\mathbf{0}\}))\cdot 2^{p-1}\cdot\gamma_1^p}{\left(\pi\sqrt{d}\right)^p \cdot \left(\psi_{\mu_1}^{-1}\left((1 + \mu_1(\{\mathbf{0}\}))/2\right)\right)^p}. \tag{3.52}$$

Now, combining (3.51) and (3.52) we have that

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

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

$$e_{l_p}(\mathcal{X},\mathrm{L})-e_{l_p}(\mathcal{X},\mathrm{L}_1)>\gamma_9 N/2 \text{ w.p. } \geq 1-\exp(-N\gamma_9^2/8)\,. \tag{3.53}$$

By Lemma 3.2.2 we have that for any $\mathrm{L}'\in\mathrm{G}(D,d)$ satisfying $\text{dist}_{\mathrm{G}}(\mathrm{L},\mathrm{L}')<(\gamma_9/4)^{1/p}$ and any $\mathbf{x}\in\mathrm{B}(\mathbf{0},1)$:

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

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

$$e_{l_p}(\mathcal{X},\mathrm{L}')-e_{l_p}(\mathcal{X},\mathrm{L}_1)>0 \text{ w.p. } \geq 1-\exp(-N\gamma_9^2/8)\,. \tag{3.54}$$

Following [51, Remark 8.4] we can cover $\mathrm{G}(D,d)\setminus\mathrm{B}_{\mathrm{G}}(\mathrm{L}_1,\gamma_1)$ 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 (3.53) is valid for its center w.p. $1-\exp(-N\gamma_9^2/8)$ and consequently (3.54) is valid for subspaces in that ball with the same probability. We thus conclude that (3.54) 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 (3.50), 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 $\min_{2\leq i\leq K}(\text{dist}_{\mathrm{G}}(\mathrm{L}_1,\mathrm{L}_i))$.

3.4 Proof of Theorem 2.4.2: Stability Analysis

Reduction of Theorem 2.4.2

We first explain how to reduce the proof of Theorem 2.4.2 when $0 < p \leq 1$ to the verification of a simpler statement. We then adapt this idea for proving the same theorem when both $p > 1$ and $K = 1$.

In order to prove Theorem 2.4.2 when $0 < p \leq 1$, i.e., prove that the global minimum of $e_{l_p}(\mathcal{X}, \mathrm{L})$ is in $\mathrm{B_G}(\mathrm{L}_1, f)$ w.o.p., we only need to show that there exists a constant $\gamma_1 > 0$ such that for any $\mathrm{L} \notin \mathrm{B_G}(\mathrm{L}_1, f)$:

$$E_{\mu_\epsilon}(e_{l_p}(\mathbf{x}, \mathrm{L})) > E_{\mu_\epsilon}(e_{l_p}(\mathbf{x}, \mathrm{L}_1)) + \gamma_1. \tag{3.55}$$

Indeed, we cover the compact space $\mathrm{G}(D, d) \setminus \mathrm{B_G}(\mathrm{L}_1, f)$ by small balls with radius $\gamma_1/2$. Then by using (3.55) 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 $\mathrm{L} \in \mathrm{G}(D, d) \setminus \mathrm{B_G}(\mathrm{L}_1, f)$ w.o.p. Equivalently, $\mathrm{G}(D, d) \setminus \mathrm{B_G}(\mathrm{L}_1, f)$ 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 (3.55) and thus Theorem 2.4.2 for $p \leq 1$, the following equation is sufficient:

$$E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x}, \mathrm{L})) > E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x}, \mathrm{L}_1)) + \gamma_1 + 2\epsilon^p, \text{ for any } \mathrm{L} \in \mathrm{G}(D, d) \setminus \mathrm{B_G}(\mathrm{L}_1, f). \tag{3.56}$$

We can similarly reduce Theorem 2.4.2 when $K = 1$ and $p > 1$ to the following condition:

$$E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x}, \mathrm{L})) > E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x}, \mathrm{L}_1)) + \gamma_1 + 2p\epsilon, \text{ for any } \mathrm{L} \in \mathrm{G}(D, d) \setminus \mathrm{B_G}(\mathrm{L}_1, f). \tag{3.57}$$

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 $\mathrm{dist_G}(\tilde{\mathrm{L}}_1, \tilde{\mathrm{L}}_2) < \eta$:

$$\mathrm{dist}(\mathbf{x}_1, \tilde{\mathrm{L}}_1)^p - \mathrm{dist}(\mathbf{x}_2, \tilde{\mathrm{L}}_1)^p < 1 - (1 - \eta)^p < p\eta, \tag{3.58}$$

and

$$\operatorname{dist}(\mathbf{x}_1, \tilde{\mathrm{L}}_1)^p - \operatorname{dist}(\mathbf{x}_1, \tilde{\mathrm{L}}_2)^p < 1 - (1-\eta)^p < p\eta. \tag{3.59}$$

When $p = 1$, (3.58) follows from the triangle inequality and (3.59) follows from Lemma 3.2.2, 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 $y_1^p - y_2^p < 1 - (1-\eta)^p$.

Proof of (3.56) and (3.57) and Conclusion of Theorem 2.4.2

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 3.2.3 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 (3.56) with $\gamma_1 := 2\epsilon^p$ as follows:

$$\begin{aligned} E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x}, \mathrm{L})) - E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x}, \mathrm{L}_1)) &\geq \left(\alpha_1 - \sum_{i=2}^K \alpha_i \right) E_{\tilde{\mu}_{1,\epsilon}}(e_{l_p}(\mathbf{x}, \mathrm{L})) \\ \geq \frac{\left(\alpha_1 - \sum_{i=2}^K \alpha_i \right) (1 - \mu_1(\{\mathbf{0}\}))\, 2^{p-1} f^p}{(\pi\sqrt{d})^p \left(\psi_{\mu_1}^{-1} \left(\frac{1+\mu_1(\{\mathbf{0}\})}{2} \right) \right)^p} &= 4\epsilon^p, \end{aligned} \tag{3.60}$$

where the second inequality applies Lemma 3.2.1.

Equation (3.57) (with $p > 1$) follows from the same argument of (3.60), where ϵ^p is now replaced by $p\epsilon$.

Remark on The Size of ϵ

If $0 < p \leq 1$ and

$$\epsilon > \frac{\left(\alpha_1 - \sum_{i=2}^K \alpha_i \right)^{\frac{1}{p}} (1 - \mu_1(\{\mathbf{0}\}))^{\frac{1}{p}}}{2^{\frac{3}{p}}\, \psi_{\mu_1}^{-1}\left(\frac{1+\mu_1(\{\mathbf{0}\})}{2} \right)} \tag{3.61}$$

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

$$\epsilon > \frac{\alpha_1 (1 - \mu_1(\{\mathbf{0}\}))\, 2^{p-3}}{p\, \pi^p d^{\frac{p}{2}}\, \psi_{\mu_1}^{-1}\left(\frac{1+\mu_1(\{\mathbf{0}\})}{2} \right)^p}, \tag{3.62}$$

then $f > \frac{\pi\sqrt{d}}{2}$, which implies that $\mathrm{B_G}(\mathrm{L}_1, f) = \mathrm{G}(D,d)$ (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 (3.61) or (3.62).

3.5 Proof of Theorem 2.4.3: Symmetry Arguments

First Reduction of Theorem 2.4.3

We use the same notation of Section 3.4, in particular, $\tilde{\mu}_\epsilon$. Theorem 2.4.3 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:

$$\gamma_{D,d}^K \left(\{\mathrm{L}_i\}_{i=1}^K \subset \mathrm{G}(D,d) : \mathrm{L}_1 = \mathrm{argmin}_{\mathrm{L}} E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x}, \mathrm{L})) \right) = 0. \tag{3.63}$$

Indeed, if (3.63) 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 (3.58) and (3.59) 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 2.4.3 and therefore (3.63) implies this theorem.

Second Reduction of Theorem 2.4.3

We define the operator

$$\mathbf{D}_{\mathrm{L},\mathbf{x},p} = P_{\mathrm{L}}(\mathbf{x})P_{\mathrm{L}}^{\perp}(\mathbf{x})^{T}\mathrm{dist}(\mathbf{x},\mathrm{L})^{(p-2)} \tag{3.64}$$

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 3.1.2, (3.63) follows from the condition:

$$\gamma_{D,d}^{K}\left(\{\mathrm{L}_i\}_{i=1}^{K} \subset \mathrm{G}(D,d) : E_{\tilde{\mu}_\epsilon}\left(\mathbf{D}_{\mathrm{L}_1,\mathbf{x},p}\right) = 0\right) = 0, \tag{3.65}$$

which we rewrite as follows:

$$\begin{aligned}
&\gamma_{D,d}^{K}\left(\{\mathrm{L}_i\}_{i=1}^{K} \subset \mathrm{G}(D,d) : E_{\tilde{\mu}_\epsilon}\left(\mathbf{D}_{\mathrm{L}_1,\mathbf{x},p}\right) = 0\right)\\
=&\gamma_{D,d}^{K}\left(\{\mathrm{L}_i\}_{i=1}^{K} \subset \mathrm{G}(D,d) : E_{\sum_{i=2}^{K}\alpha_i\tilde{\mu}_{i,\epsilon}}\left(\mathbf{D}_{\mathrm{L}_1,\mathbf{x},p}\right) = 0\right)\\
=&\gamma_{D,d}^{K}\left(\{\mathrm{L}_i\}_{i=1}^{K} \subset \mathrm{G}(D,d) : \sum_{i=2}^{K}\alpha_i\, h(\mathrm{L}_1,\mathrm{L}_i) = 0\right) = 0.
\end{aligned} \tag{3.66}$$

Since $\{\mathrm{L}_i\}_{i=1}^{K}$ are identically and independently distributed according to $\gamma_{D,d}$, Fubini's Theorem implies that (3.66) follows from the equation:

$$\gamma_{D,d}\left(\mathrm{L}_2 \in \mathrm{G}(D,d) : h(\mathrm{L}_1,\mathrm{L}_2) = \mathbf{C}(\mathrm{L}_1,\mathrm{L}_3,\cdots,\mathrm{L}_K)\right) = 0, \tag{3.67}$$

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

Third Reduction of Theorem 2.4.3

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 $1 \le i_1, i_2 \le d$:

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

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

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

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

Therefore, when we form $\mathbf{V}$ and $\mathbf{U}$ as in (3.12), the $d \times d$ matrix $\mathbf{V}h(\mathrm{L}_1, \mathrm{L}_2)\mathbf{U}^T$ is diagonal with the elements

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

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 L_1, L_3, L_4, $\cdots$, 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 (3.67) is implied by the following equation:

$$\gamma_{D,d}\left(\mathrm{L}_2 \in \mathrm{G}(D, d) : \lambda_1(h(\mathrm{L}_1, \mathrm{L}_2)) \in \{\sigma_i\}_{i=1}^{D}\right) = 0, \tag{3.68}$$

which we express as:

$$\gamma_{D,d}\left(\int_{\sum_{i=1}^{d} {a_1}^2 \leq 1} \cos\theta_1 \sin\theta_1 a_1^2 \left(\sum_{i=1}^{d} a_i^2 \sin^2\theta_i\right)^{\frac{p-2}{2}} \mathrm{d}V \in \{\sigma_i\}_{i=1}^{D}\right) \tag{3.69}$$

$$= 0.$$

Proof of (3.69) and Conclusion of Theorem 2.4.3

We first conclude (3.69) when $p = 2$. In this case

$$\int_{\sum_{i=1}^{d} a_1{}^2 \leq 1} \cos\theta_1 \sin\theta_1 a_1^2 \left(\sum_{i=1}^{d} a_i^2 \sin^2\theta_i \right)^{\frac{p-2}{2}} \mathrm{d}V \equiv \int_{\sum_{i=1}^{d} a_1{}^2 \leq 1} \cos\theta_1 \sin\theta_1 a_1^2 \, \mathrm{d}V \quad (3.70)$$

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, (3.69) (and consequently (3.63)) is verified in this case.

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

$$\int_{\sum_{i=1}^{d} a_1{}^2 \leq 1} \cos\theta_1 \sin\theta_1 a_1^2 \left(\sum_{i=1}^{d} a_i^2 \sin^2\theta_i \right)^{\frac{p-2}{2}} \mathrm{d}V \quad (3.71)$$

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

$$\gamma_{D,d} \left(h(\mathrm{L}_1, \mathrm{L}_2) \in \{\sigma_i\}_{i=1}^{D} | \{\theta_i\}_{i=1}^{d-1} \right) = 0. \quad (3.72)$$

Combining (3.72) and Fubini's Theorem, we conclude (3.69).

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 (3.91) 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 3.4).

3.6 Proof of Theorem 2.5.1: Recovery of Subspaces by Calculus on the Grassmannian

Preliminaries

We view the energy $e_{l_p}(\mathcal{X}, \mathrm{L}_1, \cdots, \mathrm{L}_K)$ as a function defined on $\mathrm{G}(D,d)^K$ while being conditioned on the fixed data set $\mathcal{X}$. Therefore, the minimizer of $e_{l_p}(\mathcal{X}, \mathrm{L}_1, \cdots, \mathrm{L}_K)$ is an element $(\mathrm{L}'_1, \cdots, \mathrm{L}'_K)$ in $\mathrm{G}(D,d)^K$. Since any permutation of its K coordinates in $\mathrm{G}(D,d)$ results in another minimizer, we sometimes say that the set $\{\mathrm{L}'_1, \cdots, \mathrm{L}'_K\}$ is a minimizer.

We denote $e_{l_p}(\mathbf{x}, \mathrm{L}_1, \cdots, \mathrm{L}_K) := e_{l_p}(\{\mathbf{x}\}, \mathrm{L}_1, \cdots, \mathrm{L}_K)$ and view it as a function on $\mathbb{R}^D \times \mathrm{G}(D,d)^K$.

We distinguish elements in $\mathrm{G}(D,d)^K$ by the l_∞ norm on the product space, i.e.,

$$\mathrm{dist}_{\mathrm{G}^\mathrm{K}}((\mathrm{L}_1, \cdots, \mathrm{L}_K), (\hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K)) = \max_{i=1,\cdots,K} (\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_i, \hat{\mathrm{L}}_i)). \tag{3.73}$$

We partition $\mathcal{X}$ into the subsets $\{\mathcal{X}_i\}_{i=0}^K$ with $\{N_i\}_{i=0}^K$ points sampled according to the measures $\{\mu_i\}_{i=0}^K$.

Auxiliary Lemmata

We formulate the following lemma, which will be used throughout this proof. It uses the constant τ_0 of (2.2) and the following notation w.r.t. the fixed d-subspaces $\mathrm{L}_1, \mathrm{L}_2, \cdots, \mathrm{L}_K, \hat{\mathrm{L}}_1, \hat{\mathrm{L}}_2, \cdots, \hat{\mathrm{L}}_K \in \mathrm{G}(D,d)$:

$$I(i) = \arg\min_{1 \le j \le K} \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_i, \hat{\mathrm{L}}_j) \ \ \forall\, 1 \le i \le K \tag{3.74}$$

and

$$d_0 = \min_{i_1, i_2, \cdots, i_K \in \mathcal{P}_K} \mathrm{dist}_{\mathrm{G}^\mathrm{K}}((\mathrm{L}_{i_1}, \cdots, \mathrm{L}_{i_K}), (\hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K)). \tag{3.75}$$

Lemma 3.6.1. *Suppose that* $\mathrm{L}_1, \cdots, \mathrm{L}_K, \hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K \in \mathrm{G}(D,d)$ *and* $0 < p \le 1$. *If* $(I(1), \cdots, I(K))$ *is a permutation of* $(1, \cdots, K)$, *then*

$$E_\mu e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K) - E_\mu e_{l_p}(\mathbf{x}, \mathrm{L}_1, \cdots, \mathrm{L}_K) \ge \left(\tau_0 \min_{1 \le j \le K} \alpha_j - \alpha_0\right) d_0^p.$$

On the other hand, if $(I(1), \cdots, I(K))$ *is not a permutation of* $(1, \cdots, K)$, *then*

$$\begin{aligned} &E_\mu e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K) - E_\mu e_{l_p}(\mathbf{x}, \mathrm{L}_1, \cdots, \mathrm{L}_K) \\ \ge &\tau_0 \left(\min_{1 \le j \le K} \alpha_j\right) \left(\min_{1 \le i, j \le K} \mathrm{dist}_{\mathrm{G}}{}^p(\mathrm{L}_i, \mathrm{L}_j)/2\right) - \alpha_0. \end{aligned}$$

Reduction of Theorem 2.5.1

We note that it is enough to prove that the set $\{\mathrm{L}_1^*, \cdots, \mathrm{L}_K^*\}$ minimizes w.o.p. the energy $e_{l_p}(\mathcal{X}_0 \cup \mathcal{X}_1, \mathrm{L}_1, \cdots, \mathrm{L}_K)$. Indeed, it follows from the immediate observation:

$$e_{l_p}\left(\sum_{i=2}^{K} \mathcal{X}_i, \mathrm{L}_1^*, \cdots, \mathrm{L}_K^*\right) = 0.$$

We will first prove that $\{\mathrm{L}_1^*, \cdots, \mathrm{L}_K^*\}$ is a minimizer in a local ball and later show that it is a minimizer in $\mathrm{G}(D,d)^K$.

Proof in a Local Ball by Direct Minimization

We will show that $(\mathrm{L}_1^*, \cdots, \mathrm{L}_K^*)$ is a minimizer w.o.p. of $e_{l_p}(\mathcal{X}_0 \cup \mathcal{X}_1, \mathrm{L}_1, \cdots, \mathrm{L}_K)$ in $\mathrm{B}_{\mathrm{G}}((\mathrm{L}_1^*, \cdots, \mathrm{L}_K^*), \gamma_1)$, for sufficiently small number γ_1. Consequently, $(\mathrm{L}_1^*, \cdots, \mathrm{L}_K^*)$ is also the minimizer w.o.p. of e_{l_p} in $\cup_{i_1,i_2,\cdots,i_K \in \mathcal{P}_K} \mathrm{B}_{\mathrm{G}}((\mathrm{L}_{i_1}^*, \cdots, \mathrm{L}_{i_K}^*), \gamma_1)$, where $\mathcal{P}_K$ is the set of all permutations of $(1, 2, \cdots, K)$. This restriction to a sufficiently small ball allows direct calculations on the Grassmannian.

In order to simplify notation in this part of the proof, we will adopt WLOG the convention that the RHS of (3.73) occurs at $i = 1$, i.e.,

$$\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1^*, \hat{\mathrm{L}}_1) = \max_{i=1,\cdots,K} (\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_i^*, \hat{\mathrm{L}}_i)). \tag{3.76}$$

Let $t_0 := \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1^*, \hat{\mathrm{L}}_1)$. For each $1 \leq i \leq K$, we parametrize according to arc length the geodesic lines from L_i^* to $\hat{\mathrm{L}}_i$ by functions $\mathrm{L}_i(t)$, $1 \leq i \leq K$, on the interval $[0, t_0]$ such that

$$\mathrm{L}_i(0) = \mathrm{L}_i^* \quad \text{and} \quad \mathrm{L}_i(t_0) = \hat{\mathrm{L}}_i. \tag{3.77}$$

We will prove that for some $\gamma_1 > 0$

$$\frac{\mathrm{d}}{\mathrm{d}t^p} \left(e_{l_p}(\mathcal{X}_0 \cup \mathcal{X}_1, \mathrm{L}_1(t), \cdots, \mathrm{L}_K(t))\right) > 0 \quad \text{for all } 0 \leq t \leq \gamma_1 \text{ w.o.p.} \tag{3.78}$$

This will clearly imply our desired result that $(\mathrm{L}_1^*, \cdots, \mathrm{L}_K^*)$ is a minimizer w.o.p. of $e_{l_p}(\mathcal{X}_0 \cup \mathcal{X}_1, \mathrm{L}_1, \cdots, \mathrm{L}_K)$ in $\mathrm{B}_{\mathrm{G}}((\mathrm{L}_1^*, \cdots, \mathrm{L}_K^*), \gamma_1)$.

Our proof of (3.78) is based on the following estimate:

$$\frac{\mathrm{d}}{\mathrm{d}t^p} \left(e_{l_p}(\mathbf{x}, \mathrm{L}_1(t), \cdots, \mathrm{L}_K(t))\right)\Big|_{t=0} \geq -\|\mathbf{x}\|. \tag{3.79}$$

In order to establish (3.79), we denote $j = \arg\min_{1\le i\le K} \operatorname{dist}(\mathbf{x}, \mathrm{L}_i^*)$ and apply Lemma 3.2.2 to obtain that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t^p}\left(e_{l_p}(\mathbf{x}, \mathrm{L}_1(t), \cdots, \mathrm{L}_K(t))\right)\Big|_{t=0} &= \lim_{t\to 0} \frac{\operatorname{dist}(\mathbf{x}, \mathrm{L}_j(t))^p - \operatorname{dist}(\mathbf{x}, \mathrm{L}_j(0))^p}{t^p} \\ \ge -\|\mathbf{x}\| \lim_{t\to 0} \frac{\operatorname{dist}_{\mathrm{G}}(\mathrm{L}_j(t), \mathrm{L}_j(0))^p}{t^p}.& \end{aligned} \tag{3.80}$$

We also note that for all $0 \le t \le t_0$:

$$\frac{\operatorname{dist}_{\mathrm{G}}(\mathrm{L}_j(t), \mathrm{L}_j(0))^p}{t^p} \ge \frac{\operatorname{dist}_{\mathrm{G}}(\mathrm{L}_1(t), \mathrm{L}_1(0))^p}{t^p} = 1. \tag{3.81}$$

Indeed, if $t = t_0$ the inequality in (3.81) follows from (3.76) and the equality follows from (3.77). Moreover, both of them extend to $0 \le t < t_0$ by the underlying property of arc length parametrization. Equation (3.79) thus follows from (3.80) and (3.81).

Combining (3.79) with Hoeffding's inequality, we obtain that

$$\frac{\mathrm{d}}{\mathrm{d}t^p}\left(e_{l_p}(\mathcal{X}_0, \mathrm{L}_1(t), \cdots, \mathrm{L}_K(t))\right)\Big|_{t=0} \ge -\sum_{\mathbf{x}\in\mathcal{X}_0} \|\mathbf{x}\| \ge -\alpha_0 N \text{ w.o.p.} \tag{3.82}$$

We similarly derive an equation analogous to (3.82) when replacing $\mathcal{X}_0$ with $\mathcal{X}_1$ by applying Lemma 3.2.1 and Hoeffding's inequality as follows:

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t^p}\left(e_{l_p}(\mathcal{X}_1, \mathrm{L}_1(t), \cdots, \mathrm{L}_K(t))\right)\Big|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t}\left(e_{l_1}(\mathcal{X}_1, \mathrm{L}_1(t))\right)\Big|_{t=0} \\ \ge &\frac{\tau_0\alpha_1 N}{2} \text{ w.o.p.}\end{aligned} \tag{3.83}$$

At last, combining (3.82), (3.83) and (2.5) we obtain that there exists $\gamma_1' \equiv \gamma_1'(D, d, K, p, \alpha_0, \alpha_1)$ such that w.o.p.

$$\frac{\mathrm{d}}{\mathrm{d}t^p}\left(e_{l_p}(\mathcal{X}_0 \cup \mathcal{X}_1, \mathrm{L}_1(t), \cdots, \mathrm{L}_K(t))\right)\Big|_{t=0} \ge \left(\frac{\tau_0\,\alpha_1}{2} - \alpha_0\right) N > \gamma_1'\, N.$$

Using the arguments of the proof of (3.31) we conclude that there exists a constant $\gamma_1 \equiv \gamma_1(D, d, K, p, \alpha_0, \alpha_1, \min_{2\le i\le K} \operatorname{dist}(\mathrm{L}_1^*, \mathrm{L}_i^*), \mu_0, \mu_1) > 0$ such that (3.78) holds.

Conclusion of Theorem 2.5.1 by a Global Estimate

In order to conclude the theorem it is enough to prove that $\{\mathrm{L}_1^*, \cdots, \mathrm{L}_K^*\}$ is a global minimizer w.o.p. of $e_{l_p}(\mathcal{X}_0 \cup \mathcal{X}_1, \mathrm{L}_1, \cdots, \mathrm{L}_K)$ in the set

$$\mathrm{GP}(D, d, \gamma_1) := \mathrm{G}(D, d)^K \setminus \cup_{i_1, i_2, \cdots, i_K \in \mathcal{P}_K} \mathrm{B}_{\mathrm{G}}((\mathrm{L}_{i_1}^*, \cdots, \mathrm{L}_{i_K}^*), \gamma_1). \tag{3.84}$$

Combining Lemma 3.6.1, the fact that $d_0 > \gamma_1$ (which follows from the definition of d_0 in (3.75)), Hoeffding's inequality and (2.5), we obtain that there exists $\gamma_2 \equiv \gamma_2\left(D, d, K, p, \alpha_0, \min_{1\le i\le K}\alpha_i, \min_{1\le$ 0 such that for any fixed $(\hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K) \in \mathrm{GP}(D, d, \gamma_1)$:

$$e_{l_p}(\mathcal{X}, \hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K) - e_{l_p}(\mathcal{X}, \mathrm{L}_1^*, \cdots, \mathrm{L}_K^*) > \gamma_2 N \ \text{ w.o.p.} \tag{3.85}$$

Following the proof of Theorem 2.4.1 (i.e., covering $\mathrm{GP}(D, d, \gamma_1)$ by balls where the estimate of Section 3.6 can be applied) we easily extend (3.85) w.o.p. for all K subspaces in the set $\mathrm{GP}(D, d, \gamma_1)$ (instead of fixed ones) and thus conclude the theorem.

3.7 Proof of Theorem 2.5.2: Stability to Noise and Some Counterexamples

Proof of Theorem 2.5.2

Following the argument of Section 3.4, we reduce the verification of Theorem 2.5.2 to proving that if for all permutations $i_1, \cdots, i_K \in \mathcal{P}_K$, $\hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K \in \mathrm{G}(D, d)$ satisfy that $\mathrm{dist}_{\mathrm{G^K}}((\mathrm{L}_{i_1}^*, \cdots, \mathrm{L}_{i_K}^*),$ $(\hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K)) > f$, then

$$E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K)) > E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x}, \mathrm{L}_1^*, \cdots, \mathrm{L}_K^*)) + \gamma_3 + 2\epsilon^p. \tag{3.86}$$

In view of Lemma 3.6.1, in order to conclude (3.86) it is sufficient to prove that

$$\left(\tau_0 \min_{1\le j\le K} \alpha_j - \alpha_0\right) f^p > \gamma_3 + 2\epsilon^p \tag{3.87}$$

and

$$\tau_0 \min_{1\le j\le K} \alpha_j \min_{1\le i,j\le K} \mathrm{dist}_{\mathrm{G}}{}^p(\mathrm{L}_i^*, \mathrm{L}_j^*)/2^p - \alpha_0 > \gamma_3 + 2\epsilon^p. \tag{3.88}$$

Setting $\gamma_3 = \epsilon^p$, (3.87) follows from (2.7) and (3.88) follows from (2.6).

Remark on the Size of ϵ

If

$$\epsilon < \pi\sqrt{d}\, 3^{-\frac{1}{p}} \left(\tau_0 \min_{1\le j\le K} \alpha_j - \alpha_0\right)^{\frac{1}{p}} /2, \tag{3.89}$$

then $f > \pi\sqrt{d}/2$, so that there is no restriction on the minimizer of (1.2) in $\mathrm{G}(D, d)^K$. It thus makes sense to further restrict ϵ to be at least lower than the right hand side of (3.89).

A Counterexample to Exact Asymptotic Recovery in the Setting of Theorem 2.5.2

Theorem 2.5.2 is not surprising and its proof is based on straightforward stability analysis. One may ask if it is possible to exactly recover the underlying subspaces as the number of sampled points, N, approaches infinity. This is true, for example, when $K = 1$ [44, Section 11.6][45] or $d = 0$ [52]. However, it is often not true when $d > 1$ and $K > 1$ as in the typical counterexample demonstrated in Figure 3.1(a). In this example, $D = 2$, $K = 2$, $d = 1$, $\alpha_0 = 0$ and the two underlying measures μ_1 and μ_2 (corresponding to the two underlying lines L_1^* and L_2^*) are uniformly distributed in the two gray regions demonstrated in this figure.

Our argument uses the following basic observation, which is also applied in Section 3.8 and proved in the appendix.

Lemma 3.7.1. *If* $\mathrm{L}_1^*, \cdots, \mathrm{L}_K^* \in \mathrm{G}(D, d)$ *and*

$$(\mathrm{L}_1^*, \cdots, \mathrm{L}_K^*) = \underset{(\hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K)}{\arg\min} E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K)),$$

then

$$\mathrm{L}_1^* = \underset{\mathrm{L} \in \mathrm{G}(D,d)}{\arg\min} E_{\tilde{\mu}_\epsilon}\left(e_{l_p}(\mathbf{x}, \mathrm{L}) I(\mathbf{x} \in \mathrm{Y}_1)\right). \tag{3.90}$$

We claim that for any fixed $p > 0$, the distance between the global minimizer of the energy (1.2) and $\{\mathrm{L}_1^*, \mathrm{L}_2^*\}$ is bounded from below w.o.p. by a positive constant. Indeed, assume on the contrary that this distance obtains the value zero or approaches zero as N approaches infinity. Then according to (3.90), the distance between L_1^* and the minimizer of $E_\mu(e_{l_p}(\mathbf{x}, \mathrm{L})\, I(\mathbf{x} \in \mathrm{Y}_1))$ is zero. However, this leads to a contradiction. Indeed, the region Y_1 is demonstrated in Figure 3.1(b). We also demonstrate there the best l_2 line for Y_1, denoted by $\tilde{\mathrm{L}}_1$, which is different than L_1^* and thus $E_{\mathbf{x} \in \mathrm{Y}_1}(e_{l_p}(\mathbf{x}, \tilde{\mathrm{L}}_1)) < E_{\mathbf{x} \in \mathrm{Y}_1}(e_{l_p}(\mathbf{x}, \mathrm{L}_1^*))$.

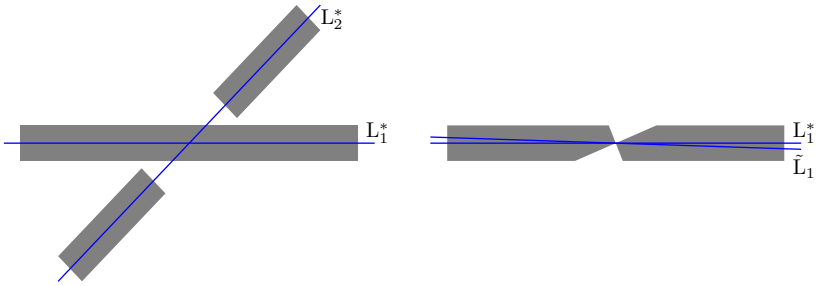
3.8 Proof of Theorem 2.5.3

3.8.1 Preliminaries

Notation

We designate the projection from $\mathbb{R}^D$ onto its subspace L by P_{L} and the corresponding orthogonal projection by $P_{\mathrm{L}}^{\perp}$. We define

$$\mathbf{D}_{\mathrm{L},\mathbf{x},p} = P_{\mathrm{L}}(\mathbf{x}) P_{\mathrm{L}}^{\perp}(\mathbf{x})^T \mathrm{dist}(\mathbf{x}, \mathrm{L})^{(p-2)}. \tag{3.91}$$



(a)Underlying lines and gray regions of uniform distributions around them.

(b)The gray region is the intersection of Y_1 with uniform distribution region around L_1. The best l_2 line in Y_1 is $\tilde{L}_1$

Figure 3.1: A counterexample showing that exact recovery with noise is impossible.

We form the regions $\{Y_i\}_{i=1}^K$, which are obtained by a Voronoi diagram (restricted to the unit ball) of the d-subspaces $\{L_i^*\}_{i=1}^K \subseteq G(D,d)$ as follows:

$$\begin{aligned} Y_i &\equiv Y_i(L_1^*, \cdots, L_K^*) \\ &= \{\mathbf{x} \in B(\mathbf{0},1) : \text{dist}(\mathbf{x}, L_i^*) < \text{dist}(\mathbf{x}, L_j^*) \ \ \forall j : 1 \le j \ne i \le K\}. \end{aligned} \tag{3.92}$$

When introducing additional subspaces denoted by $\hat{L}_2^* \in G(D,d)$ and $\tilde{L}_2^* \in G(D,d)$, we will also use the following short notation for $1 \le i \le K$:

$$\hat{Y}_i = Y_i(L_1^*, \hat{L}_2^*, L_3^*, \cdots, L_K^*), \quad \tilde{Y}_i = Y_i(L_1^*, \tilde{L}_2^*, L_3^*, \cdots, L_K^*) \tag{3.93}$$

$$\text{and} \quad Y_i = Y_i(L_1^*, L_2^*, L_3^*, \cdots, L_K^*). \tag{3.94}$$

We denote by $\bar{Y}_i$ the closure of Y_i, that is,

$$\bar{Y}_i = \{\mathbf{x} \in B(\mathbf{0},1) : \text{dist}(\mathbf{x}, L_i^*) \le \text{dist}(\mathbf{x}, L_j^*) \ \ \forall j : 1 \le j \ne i \le K\}. \tag{3.95}$$

Similarly, the closure of $\hat{Y}_i$ is $\bar{\hat{Y}}_i$.

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 $E \subseteq B(\mathbf{0},1) \cap L_i^*$: $\tilde{\mu}_{i,\epsilon}(E) = \mu_{i,\epsilon}(P_{L_i^*}^{-1}(E))$). Also, let $\tilde{\mu}_\epsilon := \alpha_0 \mu_0 + \sum_{i=1}^K \alpha_i \tilde{\mu}_{i,\epsilon}(E)$.

Let $\mathcal{L}_k$ denote the k-th dimensional Lebesgue measure. We denote $d^* = d \wedge (D-d)$ and let $\theta_{d^*}(L_i^*, L_j^*)$ be the d^*-th largest principal angle between the d-subspaces L_i^* and L_j^*.

For $L, L^* \in G(D,d)$, we define the "orthogonal subtraction" $\ominus$ as follows:

$$L^* \ominus L = L^* \cap (L \cap L^*)^\perp.$$

Auxiliary Lemmata

Using the notation above, we formulate two lemmata, which will be used throughout this proof. Lemma 3.7.1 is proved in the appendix, whereas the proof of Lemma 3.8.1 is identical to that of [41, Proposition 2.2] (while replacing sums by expectations).

Lemma 3.8.1. *For any* $\mathrm{L}^* \in \mathrm{G}(D,d)$ *and distribution* μ*, a necessary condition for* L^* *to be a local minimum of* $E_\mu\left(l_p(\mathbf{x},\mathrm{L})\right)$ *is:*

$$E_\mu(\mathbf{D}_{\mathrm{L}^*,\mathbf{x},p})) = \mathbf{0}. \tag{3.96}$$

The last lemma quantifies under some conditions the sensitivity of the region Y_j, for some $1 \leq j \leq K$, to perturbations in the subspace L_i, where $1 \leq i \leq K$ and $i \neq j$. WLOG we formulate it with $j = 1$ and $i = 2$ (note that it uses the short notation of (3.93)).

Lemma 3.8.2. *If* $\hat{\mathrm{L}}_2, \mathrm{L}_1^*, \mathrm{L}_2^*, \cdots, \mathrm{L}_K^*$ *are subspaces in* $\mathrm{G}(D,d)$ *such that* $\hat{\mathrm{L}}_2 \neq \mathrm{L}_2^*$*,*

$$\min_{j\neq 2}(\theta_{d^*}(\hat{\mathrm{L}}_2, \mathrm{L}_j^*)) > 0, \quad \min_{1\leq i\neq j\leq K}(\theta_{d^*}(\mathrm{L}_i^*, \mathrm{L}_j^*)) > 0, \tag{3.97}$$

$$\textit{and } \theta_{d^*}(\hat{\mathrm{L}}_2, \mathrm{L}_1^*) \vee \theta_{d^*}(\mathrm{L}_2^*, \mathrm{L}_1^*) \leq \min_{3\leq i\leq K} \theta_{d^*}(\mathrm{L}_i^*, \mathrm{L}_1^*), \tag{3.98}$$

then

$$\mathcal{L}_D\left((\hat{\mathrm{Y}}_1 \setminus \mathrm{Y}_1) \cup (\mathrm{Y}_1 \setminus \hat{\mathrm{Y}}_1)\right) > 0. \tag{3.99}$$

3.8.2 A Special Case

The proof of Theorem 2.5.2 is rather involved. In order to develop a simple intuition we provide an elementary proof of the very special case where $d = 1$, $p = 2$ and $K = 2$. For simplicity we also assume that $D = 2$, though our argument easily extends to $D > 2$. Figure 3.2 shows the two underlying lines L_1^* and L_2^* and their corresponding regions Y_1 and Y_2. We note that the best l_2 lines (i.e., lines minimizing the orthogonal least squares error) for μ_0 restricted to Y_1 and Y_2 are the central axes of those regions. Since $\alpha_0 > 0$, the best l_2 lines for μ restricted to Y_1 and Y_2 (denoted by $\tilde{\mathrm{L}}_1$ and $\tilde{\mathrm{L}}_2$ respectively) must reside between the latter best l_2 lines for μ_0 and L_1^* and L_2^*. In particular, they are different from L_1^* and L_2^* as demonstrated in the figure. Therefore, $E_{\tilde{\mu}}(e_{l_p}(\mathbf{x}, \mathrm{L}_1^*, \mathrm{L}_2^*)) > E_{\tilde{\mu}}(e_{l_p}(\mathbf{x}, \tilde{\mathrm{L}}_1, \tilde{\mathrm{L}}_2))$. This implies that w.o.p. $e_{l_p}(\mathcal{X}, \mathrm{L}_1^*, \mathrm{L}_2^*) > e_{l_p}(\mathcal{X}, \tilde{\mathrm{L}}_1, \tilde{\mathrm{L}}_2)$.

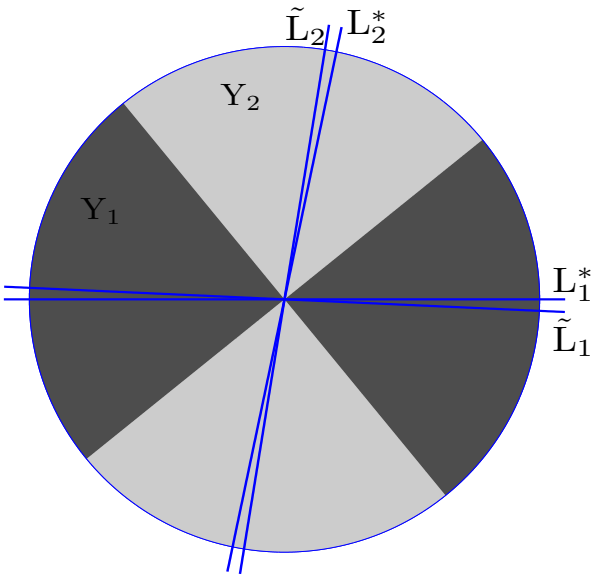


Figure 3.2: Illustrative proof of Theorem 2.5.3 in the special case.

3.8.3 Reduction to Simpler Statements

We follow by gradually reducing the general statement of Theorem 2.5.2 to simpler formulations.

Reduction I: Using the Voronoi-type Regions $\{\mathrm{Y}_i\}_{i=1}^K$

We will show here that the following equation implies Theorem 2.5.3:

$$\gamma_{D,d}^K(\{\mathrm{L}_i^*\}_{i=1}^K \subset \mathrm{G}(D,d) : E_{\mu_0}(I(\mathbf{x} \in \mathrm{Y}_j)\, \mathbf{D}_{\mathrm{L}_j^*,\mathbf{x},p}) = \mathbf{0}\ \forall\, 1 \leq j \leq K) = 0. \tag{3.100}$$

First, we apply the argument of [41, Section 3.6.1] (and thus the assumptions on $\{\nu_i\}_{i=1}^K$ specified in Definition 2.3.1) to obtain that Theorem 2.5.3 follows by the equation:

$$\begin{aligned}&\gamma_{D,d}^K\Big(\{\mathrm{L}_i^*\}_{i=1}^K \subset \mathrm{G}(D,d) : (\mathrm{L}_1^*, \cdots, \mathrm{L}_K^*) \\ &= \underset{(\mathrm{L}_1,\cdots,\mathrm{L}_K)}{\arg\min}\ E_{\tilde{\mu}_\epsilon}\big(e_{l_p}(\mathbf{x}, \mathrm{L}_1, \cdots, \mathrm{L}_K)\big)\Big) = 0.\end{aligned} \tag{3.101}$$

Next, applying Lemma 3.7.1 we conclude that (3.101) is a direct consequence of the equation:

$$\gamma_{D,d}^{K}\left(\{\mathrm{L}_i^*\}_{i=1}^{K}\subset \mathrm{G}(D,d):\right.$$
$$\left.\mathrm{L}_j^*=\underset{\mathrm{L}\in \mathrm{G}(D,d)}{\arg\min}\, E_{\tilde{\mu}_\epsilon}\left(e_{l_p}(\mathbf{x},\mathrm{L})I(\mathbf{x}\in \mathrm{Y}_j)\right)\ \ \forall\, 1\leq j\leq K\right)=0. \tag{3.102}$$

Furthermore, applying Lemma 3.8.1 with $\mu=\tilde{\mu}_\epsilon|_{\mathrm{Y}_j}$, we obtain that (3.102) follows by the equation

$$\gamma_{D,d}^{K}\left(\{\mathrm{L}_i^*\}_{i=1}^{K}\subset \mathrm{G}(D,d): E_{\tilde{\mu}_\epsilon}(I(\mathbf{x}\in \mathrm{Y}_j)\,\mathbf{D}_{\mathrm{L}_j^*,\mathbf{x},p})=0\ \forall\, 1\leq j\leq K\right)=0. \tag{3.103}$$

At last we conclude the desired reduction by noting that (3.103) and (3.100) are equivalent (indeed, the only relevant components of the measure $\tilde{\mu}_\epsilon$ in (3.103) are μ_0 and μ_j and the corresponding expectation according to μ_j is zero).

Reduction II: From K Subspaces to a Single Subspace

We reduce (3.100) so that its underlying condition involves a single subspace as follows:

$$\gamma_{D,d}(\mathrm{L}_2^*\in \mathrm{G}(D,d):\min_{1\leq i\neq j\leq K}\theta_{d^*}(\mathrm{L}_i^*,\mathrm{L}_j^*)>0,$$
$$\underset{2\leq i\leq K}{\arg\min}\,\theta_{d^*}(\mathrm{L}_1^*,\mathrm{L}_i^*)=2,\ E_{\mu_0}(I(\mathbf{x}\in \mathrm{Y}_1)\,\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p})=\mathbf{0})=0. \tag{3.104}$$

We remark that some of the underlying technical conditions of (3.104) appear in (3.97) and (3.98) and will be better understood later when applying Lemma 3.8.2.

We verify this reduction as follows. WLOG (3.104) can be equivalently formulated by replacing L_2^* with L_k^*, for some $3\leq k\leq K$, while letting $\arg\min_{2\leq i\leq K}\theta_{d^*}(\mathrm{L}_1^*,\mathrm{L}_i^*)=k$. Combining this observation with elementary properties of measures we have that

$$\gamma_{D,d}^{K}\Big(\{\mathrm{L}_i^*\}_{i=1}^{K}\subset \mathrm{G}(D,d): E_{\mu_0}\big(I(\mathbf{x}\in \mathrm{Y}_j)\,\mathbf{D}_{\mathrm{L}_j^*,\mathbf{x},p}\big)=\mathbf{0}\ \ \forall 1\leq j\leq K\Big)$$
$$\leq\sum_{k=2}^{K}\int_{\mathrm{G}(D,d)^{K-1}}\gamma_{D,d}\Big(\mathrm{L}_k^*:\min_{1\leq i\neq j\leq K}\theta_{d^*}(\mathrm{L}_i^*,\mathrm{L}_j^*)>0,\underset{2\leq i\leq K}{\arg\min}\,\theta_{d^*}(\mathrm{L}_1^*,\mathrm{L}_i^*)=k,$$
$$E_{\mu_0}(I(\mathbf{x}\in \mathrm{Y}_1)\,\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p})=\mathbf{0}\,\Big|\,\{\mathrm{L}_i^*\}_{1\leq i\neq k\leq K}\Big)\cdot \mathrm{d}\left(\gamma_{D,d}^{K-1}(\{\mathrm{L}_i^*\}_{1\leq i\neq k\leq K})\right)$$
$$+\gamma_{D,d}^{K}\left(\{\mathrm{L}_i^*\}_{i=1}^{K}\subset \mathrm{G}(D,d):\min_{1\leq i,j\leq K}\theta_{d^*}(\mathrm{L}_i^*,\mathrm{L}_j^*)=0\right)=0.$$

3.8.4 Concluding the Cases $d = 1$ and $d = D - 1$

We assume first that $d = 1$. We conclude the theorem in this case by proving (3.104) and then extend the analysis to the case $d = D - 1$.

Reduction of (3.104) with Additional Condition on the Grassmannian

We fix $\mathbf{v}_1$ to be one of the two unit vectors spanning L_1^* and denote by $\mathbf{u}_1$ the unit vector spanning $(\mathrm{L}_1^* + \mathrm{L}_2^*) \cap \mathrm{L}_1^{*\perp}$ having orientation such that for any point $\mathbf{x} \in \mathrm{L}_2^*$: $(\mathbf{x}^T \mathbf{u}_1)\,(\mathbf{x}^T \mathbf{v}_1) \geq 0$. We will prove that (3.104) follows from the following equation, which introduces a restriction on the Grassmannian:

$$\gamma_{D,d}\Big(\mathrm{L}_2^* \in \mathrm{G}(D,d) : \min_{1 \leq i \neq j \leq K} \theta_{d^*}(\mathrm{L}_i^*, \mathrm{L}_j^*) > 0,\ \arg\min_{2 \leq i \leq K} \theta_{d^*}(\mathrm{L}_1^*, \mathrm{L}_i^*) = 2, \\ E_{\mu_0}(I(\mathbf{x} \in \mathrm{Y}_1)\, \mathbf{D}_{\mathrm{L}_1^*, \mathbf{x}, p}) = \mathbf{0} \;\Big|\; (\mathrm{L}_1^* + \mathrm{L}_2^*) \cap \mathrm{L}_1^{*\perp} = \mathrm{Sp}(\mathbf{u}_1)\Big) = 0. \tag{3.105}$$

We define the following subset of the sphere S^{D-1}: $\Omega_0 = \{\mathbf{x} \in S^{D-1} : \mathbf{x} \perp \mathbf{v}\}$, and a measure ω on Ω_0 such that for any $\mathrm{A} \subseteq \Omega_0$: $\omega(\mathrm{A}) = \gamma_{D,d}(\mathrm{L}_2^* \in \mathrm{G}(D,d) : (\mathrm{L}_1^* + \mathrm{L}_2^*) \cap \mathrm{L}_1^{*\perp} \in \mathrm{Sp}(\mathrm{A}))$. Using this notation, (3.105) implies (3.104) as follows:

$$\begin{aligned}
&\gamma_{D,d}\Big(\mathrm{L}_2^* \in \mathrm{G}(D,d) : \min_{1 \leq i \neq j \leq K} \theta_{d^*}(\mathrm{L}_i^*, \mathrm{L}_j^*) > 0,\ \arg\min_{2 \leq i \leq K} \theta_{d^*}(\mathrm{L}_1^*, \mathrm{L}_i^*) = 2, \\
&E_{\mu_0}(I(\mathbf{x} \in \mathrm{Y}_1)\, \mathbf{D}_{\mathrm{L}_1^*, \mathbf{x}, p}) = \mathbf{0}\Big) \\
&= \int_{\Omega_0} \gamma_{D,d}\Big(\mathrm{L}_2^* : \min_{1 \leq i \neq j \leq K} \theta_{d^*}(\mathrm{L}_i^*, \mathrm{L}_j^*) > 0,\ \arg\min_{2 \leq i \leq K} \theta_{d^*}(\mathrm{L}_1^*, \mathrm{L}_i^*) = 2, \\
&E_{\mu_0}\big(I(\mathbf{x} \in \mathrm{Y}_1)\, \mathbf{D}_{\mathrm{L}_1^*, \mathbf{x}, p}\big) = \mathbf{0} \Big| (\mathrm{L}_1^* + \mathrm{L}_2^*) \cap \mathrm{L}_1^{*\perp} = \mathrm{Sp}(\mathbf{u}_1)\Big)\, \mathrm{d}\big(\omega(\mathbf{u}_1)\big) = 0.
\end{aligned}$$

Proof of (3.105)

We will show that at most one element satisfies the underlying condition of (3.105) (i.e., it is a member of the set for which $\gamma_{D,d}$ is evaluated). Assume on the contrary that there are two subspaces $\hat{\mathrm{L}}_2^*$ and $\tilde{\mathrm{L}}_2^*$ satisfying this underlying condition with corresponding angles $\hat{\theta} = \theta_{d^*}(\mathrm{L}_1^*, \hat{\mathrm{L}}_2^*)$ and $\tilde{\theta} = \theta_{d^*}(\mathrm{L}_1^*, \tilde{\mathrm{L}}_2^*)$ in $[0, \pi/2]$, where WLOG $\hat{\theta} > \tilde{\theta}$. Using the notation of (3.93), we have that

$$E_{\mu_0}\left(I(\mathbf{x} \in \tilde{\mathrm{Y}}_1 \setminus \hat{\mathrm{Y}}_1)\, \mathbf{D}_{\mathrm{L}_1^*, \mathbf{x}, p}\right) - E_{\mu_0}\left(I(\mathbf{x} \in \hat{\mathrm{Y}}_1 \setminus \tilde{\mathrm{Y}}_1)\, \mathbf{D}_{\mathrm{L}_1^*, \mathbf{x}, p}\right)$$

$$=2\cdot\left(E_{\mu_0}(I(\mathbf{x}\in\tilde{\mathrm{Y}}_1)\,\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p})-E_{\mu_0}(I(\mathbf{x}\in\hat{\mathrm{Y}}_1)\,\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p})\right)=\mathbf{0}-\mathbf{0}=\mathbf{0}. \tag{3.106}$$

Consequently,

$$E_{\mu_0}\left(I(\mathbf{x}\in\tilde{\mathrm{Y}}_1\setminus\hat{\mathrm{Y}}_1)\,\mathbf{v}_1^T\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p}\mathbf{u}_1\right)-E_{\mu_0}\left(I(\mathbf{x}\in\hat{\mathrm{Y}}_1\setminus\tilde{\mathrm{Y}}_1)\,\mathbf{v}_1^T\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p}\mathbf{u}_1\right)=0. \tag{3.107}$$

Defining

$$\theta_{\mathbf{u}_1,\mathbf{v}_1}(\mathbf{x})=\arctan\frac{\mathbf{u}_1\cdot\mathbf{x}}{\mathbf{v}_1\cdot\mathbf{x}}$$

and

$$\mathrm{Y}_{1,\hat{2}}=\{\mathbf{x}\in\mathrm{B}(\mathbf{0},1):\mathrm{dist}(\mathbf{x},\mathrm{L}_1^*)<\min_{3\le i\le K}\mathrm{dist}(\mathbf{x},\mathrm{L}_i^*)\},$$

we express the regions $\hat{\mathrm{Y}}_1$ and $\tilde{\mathrm{Y}}_1$ as follows:

$$\hat{\mathrm{Y}}_1=\mathrm{Y}_{1,\hat{2}}\cap\{\mathbf{x}\in\mathrm{B}(\mathbf{0},1):\hat{\theta}/2-\pi/2<\theta_{\mathbf{u}_1,\mathbf{v}_1}(\mathbf{x})<\hat{\theta}/2\}, \tag{3.108}$$

$$\tilde{\mathrm{Y}}_1=\mathrm{Y}_{1,\hat{2}}\cap\{\mathbf{x}\in\mathrm{B}(\mathbf{0},1):\tilde{\theta}/2-\pi/2<\theta_{\mathbf{u}_1,\mathbf{v}_1}(\mathbf{x})<\tilde{\theta}/2\}. \tag{3.109}$$

Figure 3.3 clarifies (3.108) and (3.109) in the special case where $d=1$ and $K=2$.

Combining (3.108) and (3.109) with the definition of $\mathbf{D}_{\mathrm{L},\mathbf{x},p}$ in (3.91), we obtain that

$$\hat{\mathrm{Y}}_1\setminus\tilde{\mathrm{Y}}_1\subset\{\mathbf{x}\in\mathrm{B}(\mathbf{0},1):\ \mathbf{v}_1^T\mathbf{x}\mathbf{x}^T\mathbf{u}_1\equiv\mathrm{dist}(\mathbf{x},\mathrm{L}_1^*)^{(2-p)}\mathbf{v}_1^T\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p}\mathbf{u}_1>0\} \tag{3.110}$$

and

$$\tilde{\mathrm{Y}}_1\setminus\hat{\mathrm{Y}}_1\subset\{\mathbf{x}\in\mathrm{B}(\mathbf{0},1):\ \mathbf{v}_1^T\mathbf{x}\mathbf{x}^T\mathbf{u}_1\equiv\mathrm{dist}(\mathbf{x},\mathrm{L}_1^*)^{(2-p)}\mathbf{v}_1^T\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p}\mathbf{u}_1<0\}. \tag{3.111}$$

It follows from Lemma 3.8.2 that $\mathcal{L}_D\left((\tilde{\mathrm{Y}}_1\setminus\hat{\mathrm{Y}}_1)\cup(\hat{\mathrm{Y}}_1\setminus\tilde{\mathrm{Y}}_1)\right)>0$. Therefore, $\mathcal{L}_D\left(\mathrm{B}(\mathbf{0},r)\cap((\tilde{\mathrm{Y}}_1\setminus\hat{\mathrm{Y}}_1)\cup($ 0. Since the restriction of $\mathcal{L}_D$ to the $\mathrm{B}(\mathbf{0},r)$ is absolutely continuous with respect to μ_0, we also have that $\mu_0\left(\mathrm{B}(\mathbf{0},r)\cap((\tilde{\mathrm{Y}}_1\setminus\hat{\mathrm{Y}}_1)\cup(\hat{\mathrm{Y}}_1\setminus\tilde{\mathrm{Y}}_1))\right)>0$. However, this contradicts (3.107), (3.110) and (3.111), i.e., it proves (3.105) and therefore the theorem in current special case.

The Case $d=D-1$

We note that the proof of the above case ($d=1$) can be adapted to the current case ($d=D-1$). This is done by letting $\mathbf{v}_1$ be one of the two unit vectors spanning $\mathrm{L}_1^*\cap(\mathrm{L}_1^*\cap\mathrm{L}_2^*)^\perp$ (note that $\dim(\mathrm{L}_1^*)=D-1$ and $\dim(\mathrm{L}_1^*\cap\mathrm{L}_2^*)=d-2$ so that $\dim(\mathrm{L}_1^*\cap(\mathrm{L}_1^*\cap\mathrm{L}_2^*)^\perp)=1$) and $\mathbf{u}_1$ be the unit vector of $(\mathrm{L}_1^*+\mathrm{L}_2^*)\cap\mathrm{L}_1^\perp$ with a similar orientation as in the case where $d=1$.

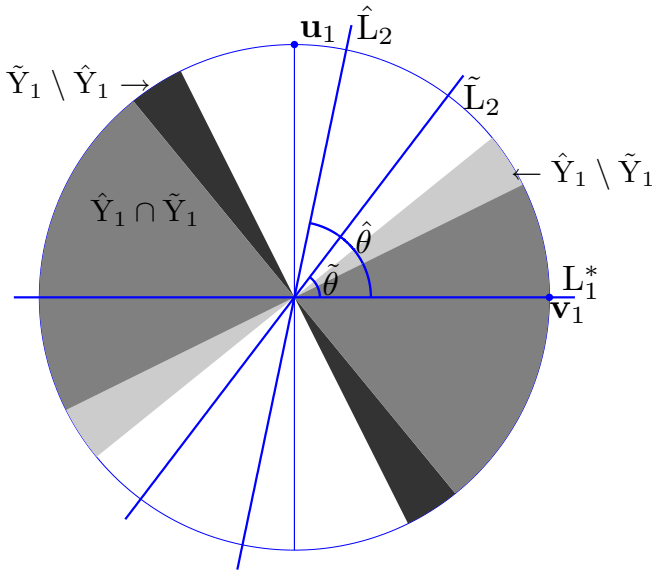


Figure 3.3: The regions $\hat{\mathrm{Y}}_1$ and $\tilde{\mathrm{Y}}_1$ and the relation to $\hat{\theta}$ and $\tilde{\theta}$ when $d = 1$ and $K = 2$.

3.8.5 Concluding the Cases where $d \neq 1$ and $d \neq D - 1$

Reduction of (3.104) with Additional Condition on the Grassmannian

In analogy to Section 3.8.4, we reduce (3.104) by introducing an additional condition on the Grassmannian.

Denoting by $B(\mathbb{R}^D, \mathbb{R}^D)$ the space of linear operators from $\mathbb{R}^D$ to itself, we define

$$\begin{aligned}\Omega_1 =&\{(P_1, P_2) \in B(\mathbb{R}^D, \mathbb{R}^D) : \exists\, \mathrm{L} \in \mathrm{G}(D, d) \text{ not orthogonal to } \mathrm{L}_1^*,\\ &\text{s.t. } \dim(\mathrm{L}_1^* \ominus \mathrm{L}) > 1,\ P_{\mathrm{L}_1^*}^T P_{\mathrm{L}} P_{\mathrm{L}_1^*} = P_1, P_{\mathrm{L}_1^*}^{\perp T} P_{\mathrm{L}} P_{\mathrm{L}_1^*}^{\perp} = P_2\}\end{aligned}$$

and the measure ω_1 on Ω_1 as follows: For any set $\mathrm{A} \subseteq \Omega_1$

$$\omega_1(\mathrm{A}) = \gamma_{D,d}\left(\mathrm{L} \in \mathrm{G}(D, d) :\ (P_{\mathrm{L}_1^*}^T P_{\mathrm{L}} P_{\mathrm{L}_1^*}, P_{\mathrm{L}_1^*}^{\perp T} P_{\mathrm{L}} P_{\mathrm{L}_1^*}^{\perp}) \in \mathrm{A}\right).$$

Using this notation we reduce (3.104) as follows:

$$\gamma_{D,d}\Big(\mathrm{L}_2^* \in \mathrm{G}(D, d) : \mathrm{L}_1^* \not\perp \mathrm{L}_2^*,\ \dim(\mathrm{L}_1^* \cap \mathrm{L}_2^{*\perp}) > 1, \min_{1 \leq i \neq j \leq K} \theta_{d^*}(\mathrm{L}_i^*, \mathrm{L}_j^*) > 0, \tag{3.112}$$

$$\arg\min_{2 \leq i \leq K} \theta_{d^*}(\mathrm{L}_1^*, \mathrm{L}_i^*) = 2, E_{\mu_0}\big(I(\mathbf{x} \in \mathrm{Y}_1)\, \mathbf{D}_{\mathrm{L}_1^*, \mathbf{x}, p}\big) = \mathbf{0}$$

$$\Big|(P_{\mathrm{L}_1^*}^T P_{\mathrm{L}_2^*} P_{\mathrm{L}_1^*}, P_{\mathrm{L}_1^*}^{\perp T} P_{\mathrm{L}_2^*} P_{\mathrm{L}_1^*}^{\perp}) = (P_1, P_2) \in \Omega_1\Big) = 0.$$

Indeed,

$$\begin{aligned}
&\gamma_{D,d}\Big(\mathrm{L}_2^* \in \mathrm{G}(D,d) : \min_{1\le i\neq j\le K} \theta_{d^*}(\mathrm{L}_i^*, \mathrm{L}_j^*) > 0,\ \arg\min_{2\le i\le K} \theta_{d^*}(\mathrm{L}_1^*, \mathrm{L}_i^*) = 2,\\
&E_{\mu_0}\big(I(\mathbf{x} \in \mathrm{Y}_1)\, \mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p}\big) = \mathbf{0}\Big)\\
\le &\int_{\Omega_1} \gamma_{D,d}\Big(\mathrm{L}_2^* : \mathrm{L}_1^* \text{ is not orthogonal to } \mathrm{L}_2^*,\ \dim(\mathrm{L}_1^* \ominus \mathrm{L}_2^*) > 1,\\
&\min_{1\le i\neq j\le K} \theta_{d^*}(\mathrm{L}_i^*, \mathrm{L}_j^*) > 0,\ \arg\min_{2\le i\le K} \theta_{d^*}(\mathrm{L}_1^*, \mathrm{L}_i^*) = 2, E_{\mu_0}\big(I(\mathbf{x} \in \mathrm{Y}_1)\, \mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p}\big) = \mathbf{0}\Big|\\
&(P_{\mathrm{L}_1^*}^T P_{\mathrm{L}_2^*} P_{\mathrm{L}_1^*}, P_{\mathrm{L}_1^*}^{\perp T} P_{\mathrm{L}_2^*} P_{\mathrm{L}_1^*}^{\perp}) = (P_1, P_2) \in \Omega_1\Big)\, \mathrm{d}\big(\tilde{\mu}(P_1, P_2)\big)\\
&+ \gamma_{D,d}\big(\mathrm{L}_2^* \in \mathrm{G}(D,d) : \dim(\mathrm{L}_1^* \ominus \mathrm{L}_2^*) \le 1, \text{ or } \mathrm{L}_2^* \perp \mathrm{L}_1^*\big) = 0 + 0 = 0.
\end{aligned}$$

Bulk of the Proof

We prove (3.112) by using the following two lemmata, which are proved below (Sections 3.8.5 and 3.8.5).

Lemma 3.8.3. *If* $\dim(\mathrm{L}_1^* \ominus \mathrm{L}_2^*) \ge 2$ *and* L_1^* *is not orthogonal to* L_2^**, then the set*

$$\mathrm{Z} = \{\mathrm{L} \in \mathrm{G}(D,d) : P_{\mathrm{L}_1^*}(P_{\mathrm{L}_2^*} - P_{\mathrm{L}})P_{\mathrm{L}_1^*} = 0,\ P_{\mathrm{L}_1^*}^{\perp}(P_{\mathrm{L}_2} - P_{\mathrm{L}})P_{\mathrm{L}_1^*}^{\perp} = 0\}$$

is infinite.

Lemma 3.8.4. *If* $\tilde{\mathrm{L}}_2^*, \hat{\mathrm{L}}_2^* \in \mathrm{G}(D,d)$ *satisfy* $\tilde{\mathrm{L}}_2^* \neq \hat{\mathrm{L}}_2^*$, $\theta_{d^*}(\hat{\mathrm{L}}_2^*, \mathrm{L}_1^*) \vee \theta_{d^*}(\mathrm{L}_2^*, \mathrm{L}_1^*)$ $\le \min_{3\le i\le K} \theta_{d^*}(\mathrm{L}_i^*, \mathrm{L}_1^*)$, $P_{\mathrm{L}_1^*}(P_{\hat{\mathrm{L}}_2^*} - P_{\tilde{\mathrm{L}}_2^*})P_{\mathrm{L}_1^*} = 0$ *and* $P_{\mathrm{L}_1^*}^{\perp}(P_{\hat{\mathrm{L}}_2^*} - P_{\tilde{\mathrm{L}}_2^*})P_{\mathrm{L}_1^*}^{\perp} = 0$*, then either* $\hat{\mathrm{L}}_2^*$ *or* $\tilde{\mathrm{L}}_2^*$ *will not satisfy the condition in* (3.112)*.*

Lemma 3.8.3 implies that there are infinite subspaces $\mathrm{L} \in \mathrm{G}(D,d)$ satisfying $P_{\mathrm{L}_1^*}^T P_{\mathrm{L}} P_{\mathrm{L}_1^*} = P_1$ and $P_{\mathrm{L}_1^*}^{\perp T} P_{\mathrm{L}} P_{\mathrm{L}_1^*}^{\perp} = P_2$. On the other hand, Lemma 3.8.4 implies that only one subspace out of these infinite subspaces satisfies the underlying condition of (3.112) and thus the latter equation is proved. We remark that the idea of this proof is somewhat similar to that of the previous case where $d = 1$ or $d = D - 1$. In this case, Lemma 3.8.3 is analogous to the fact that there is a degree of freedom in choosing L_2^* in (3.105) (since we can choose any $\theta_{d^*}(\mathrm{L}_1^*, \mathrm{L}_2^*) < \min_{3\le i\le K} \theta_{d^*}(\mathrm{L}_1^*, \mathrm{L}_i^*)$). Moreover, Lemma 3.8.4 is analogous to the fact that there were not two subspaces $\hat{\mathrm{L}}_2^*$ and $\tilde{\mathrm{L}}_2^*$ satisfying the underlying condition of (3.105).

Proof of Lemma 3.8.3

We denote $\tilde{\mathrm{L}}_1 = \mathrm{L}_1^* \ominus (\mathrm{L}_1^* \cap \mathrm{L}_2^*)$ and $\tilde{\mathrm{L}}_2 = \mathrm{L}_2^* \ominus (\mathrm{L}_1^* \cap \mathrm{L}_2^*)$. The idea of the proof is to construct a one-to-one function $g : S^{D-1} \cap \tilde{\mathrm{L}}_2 \to \mathrm{Z}$. Then, using this function and the fact that $\dim(\tilde{\mathrm{L}}_2) = \dim(\mathrm{L}_1^*) - \dim(\mathrm{L}_2^* \cap \mathrm{L}_1^*) \geq 2$, we conclude that Z, which contains $g(S^{D-1} \cap \tilde{\mathrm{L}}_2)$, is infinite.

For any $\mathbf{u}_0 \in S^{D-1} \cap \tilde{\mathrm{L}}_2$, we arbitrarily fix $\mathbf{v}_0 = \mathbf{v}_0(\mathbf{u}_0)$ as one of the two unit vectors spanning $\tilde{\mathrm{L}}_1 \cap \left(\tilde{\mathrm{L}}_2 \ominus \mathrm{Sp}(\mathbf{u}_0)\right)^{\perp}$. The vector $\mathbf{v}_0$ exists since

$$\begin{aligned}\dim\left(\tilde{\mathrm{L}}_1 \cap \left(\tilde{\mathrm{L}}_2 \ominus \mathrm{Sp}(\mathbf{u}_0)\right)^{\perp}\right) \geq & \dim(\tilde{\mathrm{L}}_1) + \dim\left(\left(\tilde{\mathrm{L}}_2 \ominus \mathrm{Sp}(\mathbf{u}_0)\right)^{\perp}\right) - D \\ = & d + (D - d + 1) - D = 1.\end{aligned}$$

We define the function g as follows:

$$g(\mathbf{u}_0) = \mathrm{Sp}(\mathbf{u}_0 - 2(\mathbf{v}_0^T\mathbf{u}_0)\mathbf{v}_0, \mathrm{L}_2^* \ominus \mathrm{Sp}(\mathbf{u}_0)).$$

We first claim that the image of g is contained in Z. Indeed, we note that

$$\begin{aligned}P_{g(\mathbf{u}_0)} - P_{\mathrm{L}_2^*} = & (\mathbf{u}_0 - 2(\mathbf{v}_0^T\mathbf{u}_0)\mathbf{v}_0)^T(\mathbf{u}_0 - 2(\mathbf{v}_0^T\mathbf{u}_0)\mathbf{v}_0) - \mathbf{u}_0^T\mathbf{u}_0 \\ = & -2(\mathbf{v}_0^T\mathbf{u}_0)\left(\mathbf{v}_0^T(\mathbf{u}_0 - (\mathbf{v}_0^T\mathbf{u}_0)\mathbf{v}_0) + (\mathbf{u}_0 - (\mathbf{v}_0^T\mathbf{u}_0)\mathbf{v}_0)^T\mathbf{v}_0\right).\end{aligned} \tag{3.113}$$

Combining (3.113) with the following two facts: $\mathbf{v}_0 \in \mathrm{L}_1^*$ and $\mathbf{u}_0 - (\mathbf{v}_0^T\mathbf{u}_0)\mathbf{v}_0 \in \mathrm{L}_1^{*\perp}$ we obtain that $g(\mathbf{u}_0) \in \mathrm{Z}$.

At last, we prove that g is one-to-one and thus conclude the proof. If on the contrary there exist $\mathbf{u}_1, \mathbf{u}_2 \in S^{D-1} \cap \tilde{\mathrm{L}}_2$ such that $\mathbf{u}_1 \neq \mathbf{u}_2$ and $g(\mathbf{u}_1) = g(\mathbf{u}_2)$, then $g(\mathbf{u}_1) = \mathrm{Sp}(g(\mathbf{u}_1), g(\mathbf{u}_2)) \supseteq (\mathrm{L}_2^* \ominus \mathrm{Sp}(\mathbf{u}_1)) + (\mathrm{L}_2^* \ominus \mathrm{Sp}(\mathbf{u}_2)) \supseteq \mathrm{L}_2^*$. Since $\dim(g(\mathbf{u}_1)) = \dim(\mathrm{L}_2^*)$ we conclude that $g(\mathbf{u}_1) = \mathrm{L}_2^*$. On the other hand we claim that for any $\mathbf{u}_0 \in S^{D-1} \cap \tilde{\mathrm{L}}_2$: $g(\mathbf{u}_0) \neq \mathrm{L}_2^*$ and thus obtain a contradiction. Indeed, since $\mathbf{u}_0 \in \tilde{\mathrm{L}}_2$, $\mathbf{v}_0 \in \tilde{\mathrm{L}}_1$ and L_1^* is not orthogonal to L_2^* we have that $\mathbf{v}_0^T\mathbf{u}_0 \neq 0$ and consequently $\mathbf{u}_0 - (\mathbf{v}_0^T\mathbf{u}_0)\mathbf{v}_0 \neq \mathbf{u}_0$. Applying the latter observation in (3.113), we obtain that $P_{g(\mathbf{u}_0)} \neq P_{\mathrm{L}_2^*}$ and consequently $g(\mathbf{u}_0) \neq \mathrm{L}_2^*$.

Proof of Lemma 3.8.4

We assume by contrary that both $\hat{\mathrm{L}}_2^*$ and $\tilde{\mathrm{L}}_2^*$ satisfy the underlying condition of (3.104) and conclude a contradiction.

We arbitrarily fix here $\mathbf{x} \in \hat{\mathrm{Y}}_1 \setminus \tilde{\mathrm{Y}}_1$ (using the notation of (3.93)). We note that $\mathrm{dist}(\mathbf{x}, \mathrm{L}_1^*) < \mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}}_2^*)$ and $\mathrm{dist}(\mathbf{x}, \mathrm{L}_1^*) < \arg\min_{3 \le i \le K} \mathrm{dist}(\mathbf{x}, \mathrm{L}_i^*)$. Since $\mathbf{x} \notin \tilde{\mathrm{Y}}_1$: $\mathrm{dist}(\mathbf{x}, \mathrm{L}_1^*) > \mathrm{dist}(\mathbf{x}, \tilde{\mathrm{L}}_2^*)$ and thus

$$\mathrm{dist}(\mathbf{x}, \tilde{\mathrm{L}}_2^*) < \mathrm{dist}(\mathbf{x}, \mathrm{L}_1^*) < \mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}}_2^*). \tag{3.114}$$

Consequently,

$$\mathbf{x}^T (P_{\hat{\mathrm{L}}_2^*} - P_{\tilde{\mathrm{L}}_2^*}) \mathbf{x} = \mathrm{dist}(\mathbf{x}, \tilde{\mathrm{L}}_2^*)^2 - \mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}}_2^*)^2 < 0. \tag{3.115}$$

We partition $P_{\hat{\mathrm{L}}_2^*} - P_{\tilde{\mathrm{L}}_2^*}$ into four parts: $P_{\mathrm{L}_1^*}(P_{\hat{\mathrm{L}}_2^*} - P_{\tilde{\mathrm{L}}_2^*}) P_{\mathrm{L}_1^*}$, $P_{\mathrm{L}_1^*}^{\perp}(P_{\hat{\mathrm{L}}_2^*} - P_{\tilde{\mathrm{L}}_2^*}) P_{\mathrm{L}_1^*}^{\perp}$, $P_{\mathrm{L}_1^*}(P_{\hat{\mathrm{L}}_2^*} - P_{\tilde{\mathrm{L}}_2^*}) P_{\mathrm{L}_1^*}^{\perp}$, and $P_{\mathrm{L}_1^*}^{\perp}(P_{\hat{\mathrm{L}}_2^*} - P_{\tilde{\mathrm{L}}_2^*}) P_{\mathrm{L}_1^*}$. The first two are zero, and the last two are adjoint to each other; we thus only consider $P_{\mathrm{L}_1^*}(P_{\hat{\mathrm{L}}_2^*} - P_{\tilde{\mathrm{L}}_2^*}) P_{\mathrm{L}_1^*}^{\perp}$. Let its SVD be

$$P_{\mathrm{L}_1^*}(P_{\hat{\mathrm{L}}_2^*} - P_{\tilde{\mathrm{L}}_2^*}) P_{\mathrm{L}_1^*}^{\perp} = \mathbf{U}\mathbf{\Sigma}\mathbf{V} = \sum_{i=1}^{d} \sigma_i \mathbf{u}_i \mathbf{v}_i^T. \tag{3.116}$$

We can express the SVD of $P_{\hat{\mathrm{L}}_2^*} - P_{\tilde{\mathrm{L}}_2^*}$ using (3.116) and the partition mentioned above as follows:

$$P_{\hat{\mathrm{L}}_2^*} - P_{\tilde{\mathrm{L}}_2^*} = \sum_{i=1}^{d} \sigma_i (\mathbf{u}_i \mathbf{v}_i^T + \mathbf{v}_i \mathbf{u}_i^T). \tag{3.117}$$

Combining (3.115) and (3.117), we obtain that

$$\sum_{i=1}^{n} \sigma_i \mathbf{u}_i^T \mathbf{x}\mathbf{x}^T \mathbf{v}_i = \mathbf{x}^T (\sum_{i=1}^{n} \sigma_i (\mathbf{u}_i \mathbf{v}_i^T + \mathbf{v}_i \mathbf{u}_i^T)) \mathbf{x}/2 < 0. \tag{3.118}$$

We define a function $f : \mathbb{R}^{D \times D} \to \mathbb{R}$ such that for any $\mathbf{A} \in \mathbb{R}^{D \times D}$: $f(\mathbf{A}) = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i^T \mathbf{A} \mathbf{v}_i$. Using (3.118) and the fact that $\{\mathbf{u}_i\}_{i=1}^d \in \mathrm{L}_1^*$ and $\{\mathbf{v}_i\}_{i=1}^d \in \mathrm{L}_1^{*\perp}$, we deduce that

$$\begin{aligned} f(\mathbf{D}_{\mathrm{L}_1^*, \mathbf{x}, p}) &= \mathrm{dist}(\mathbf{x}, \mathrm{L}_1^*)^{(p-2)} \, f(P_{\mathrm{L}_1^*}(\mathbf{x}) P_{\mathrm{L}_1^*}^{\perp}(\mathbf{x})^T) \\ &= \mathrm{dist}(\mathbf{x}, \mathrm{L}_1^*)^{(p-2)} \sum_{i=1}^{n} \sigma_i \mathbf{u}_i^T P_{\mathrm{L}_1^*}(\mathbf{x}) P_{\mathrm{L}_1^*}^{\perp}(\mathbf{x})^T \mathbf{v}_i \\ &= \mathrm{dist}(\mathbf{x}, \mathrm{L}_1^*)^{(p-2)} \sum_{i=1}^{n} \sigma_i \mathbf{u}_i^T \mathbf{x}\mathbf{x}^T \mathbf{v}_i < 0. \end{aligned} \tag{3.119}$$

Similarly, for any point $\mathbf{x} \in \tilde{\mathrm{Y}}_1 \setminus \hat{\mathrm{Y}}_1$:

$$f(\mathbf{D}_{\mathrm{L}_1^*, \mathbf{x}, p}) > 0. \tag{3.120}$$

Combining (3.106), (3.119), (3.120), Lemma 3.8.2 and the linearity of f we conclude the following contradiction establishing the current lemma:

$$\begin{aligned} 0 &= f\left(E_{\mu_0}(I(\mathbf{x}\in\tilde{\mathrm{Y}}_1\setminus\hat{\mathrm{Y}}_1)\,\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p}) - E_{\mu_0}(I(\mathbf{x}\in\hat{\mathrm{Y}}_1\setminus\tilde{\mathrm{Y}}_1)\,\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p})\right) \\ &= f\left(E_{\mu_0}(I(\mathbf{x}\in\tilde{\mathrm{Y}}_1\setminus\hat{\mathrm{Y}}_1)\,\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p})\right) - f\left(E_{\mu_0}(I(\mathbf{x}\in\hat{\mathrm{Y}}_1\setminus\tilde{\mathrm{Y}}_1)\,\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p})\right) > 0. \quad (3.121) \end{aligned}$$

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, one can estimate from below both κ_0 and δ_0 by the following number:

$$\max_{i=1.2,\cdots,K}\left(E_{\mu_\epsilon}\left(e_{l_p}(\mathbf{x},\mathrm{L}_i^*)I(\mathbf{x}\in\mathrm{Y}_i)\right) - \min_{\mathrm{L}\in\mathrm{G}(D,d)} E_{\mu_\epsilon}\left(e_{l_p}(\mathbf{x},\mathrm{L})I(\mathbf{x}\in\mathrm{Y}_i)\right)\right)\Big/4,$$

where $\mathrm{Y}_i = \mathrm{Y}_i(\mathrm{L}_1^*,\mathrm{L}_2^*,\cdots,\mathrm{L}_K^*)$. In the case of $p\geq 2$ one can also estimate by:

$$\frac{\|\max_{1\leq i\leq K} E_{\mu_\epsilon}(\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p}I(\mathbf{x}\in\mathrm{Y}_i))\|_2^2}{dD2^{p+5}}.$$

Chapter 4

New HLM Algorithms

In this chapter, we introduce two HLM algorithms based on the minimization of energy (1.2) with $p = 1$, MKF and LBF algorithms. We also introduce SLBF algorithm, which is a combination of spectral algorithm [53] with the initialization strategy used in LBF algorithm. we test the performance of these algorithms on both artificial data and the data from real-world applications.

4.1 The MKF Algorithm

We introduce here the MKF algorithm and estimate its storage and running time. We then discuss some technical details of our implementation.

4.1.1 Description of Algorithm

The MKF algorithm partitions a data set $\mathrm{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \subseteq \mathbb{R}^D$ into K clusters X_1, X_2, ..., X_K, with each cluster approximated by a d-dimensional linear subspace.

We start with a notational convention for linear subspaces. For each $1 \leq i \leq K$, let $\mathbf{P}_i$ be the $d \times D$ matrix whose rows are the orthogonal basis of the linear subspace approximating X_i, and note that $\mathbf{P}_i \mathbf{P}_i^T = \mathbf{I}_{d \times d}$. We identify the approximating subspaces of clusters $\mathrm{X}_1, \ldots, \mathrm{X}_K$ with the matrices $\mathbf{P}_1, \ldots, \mathbf{P}_K$.

We rewrite the energy in (1.2) with $p = 1$ for the partition $\{\mathrm{X}_i\}_{i=1}^K$ and the corresponding

subspaces $\{\mathbf{P}_i\}_{i=1}^K$:

$$\mathcal{E}(\{\mathrm{X}_i\}_{i=1}^K, \{\mathbf{P}_i\}_{i=1}^K) = \sum_{i=1}^K \sum_{\mathbf{x}\in\mathrm{X}_i} ||\mathbf{x} - \mathbf{P}_i^T\mathbf{P}_i\mathbf{x}||. \tag{4.1}$$

The MKF algorithm tries to partition the data into clusters $\{\mathrm{X}_i\}_{i=1}^K$ minimizing the above energy. Since the underlying flats are linear subspaces, we can normalize the elements of X to lie on the unit sphere, so that $||\mathbf{x}_j|| = 1$ for each $1 \leq j \leq N$, and express the energy function $\mathcal{E}$ as follows:

$$\begin{aligned}\mathcal{E}(\{\mathrm{X}_i\}_{i=1}^K, \{\mathbf{P}_i\}_{i=1}^K) &= \sum_{i=1}^K \sum_{\mathbf{x}\in\mathrm{X}_i} \sqrt{||\mathbf{x} - \mathbf{P}_i^T\mathbf{P}_i\mathbf{x}||^2} \\ &= \sum_{i=1}^K \sum_{\mathbf{x}\in\mathrm{X}_i} \sqrt{1 - ||\mathbf{P}_i\mathbf{x}||^2}.\end{aligned} \tag{4.2}$$

To minimize this energy, the MKF algorithm uses the method of stochastic gradient descent [54]. The derivative of the energy with respect to a given matrix $\mathbf{P}_i$ is

$$\frac{\partial\mathcal{E}}{\partial\mathbf{P}_i} = -\sum_{\mathbf{x}\in\mathrm{X}_i} \frac{\mathbf{P}_i\mathbf{x}\mathbf{x}^T}{\sqrt{1 - ||\mathbf{P}_i\mathbf{x}||^2}}. \tag{4.3}$$

The algorithm needs to adjust $\mathbf{P}_i$ according to the component of the derivative orthogonal to $\mathbf{P}_i$. The part of the derivative that is parallel to the subspace $\mathbf{P}_i$ is

$$\frac{\partial\mathcal{E}}{\partial\mathbf{P}_i}\mathbf{P}_i^T\mathbf{P}_i = -\sum_{\mathbf{x}\in\mathrm{X}_i} \frac{\mathbf{P}_i\mathbf{x}\mathbf{x}^T\mathbf{P}_i^T\mathbf{P}_i}{\sqrt{1 - ||\mathbf{P}_i\mathbf{x}||^2}}. \tag{4.4}$$

Hence the orthogonal component is

$$\mathrm{d}\mathbf{P}_i = \sum_{\mathbf{x}\in\mathrm{X}_i} \mathrm{d}_\mathbf{x}\mathbf{P}_i, \tag{4.5}$$

where

$$\mathrm{d}_\mathbf{x}\mathbf{P}_i = -\frac{(\mathbf{P}_i\mathbf{x}\mathbf{x}^T - \mathbf{P}_i\mathbf{x}\mathbf{x}^T\mathbf{P}_i^T\mathbf{P}_i)}{\sqrt{1 - ||\mathbf{P}_i\mathbf{x}||^2}}. \tag{4.6}$$

In view of the above calculations, the algorithm proceeds by picking a point $\mathbf{x}^*$ at random from the set, and then deciding which $\mathbf{P}_{i^*}$ that point currently belongs to. Then it applies the update $\mathbf{P}_{i^*} \mapsto \mathbf{P}_{i^*} - \mathrm{d}t\,\mathrm{d}_{\mathbf{x}^*}\mathbf{P}_{i^*}$, where $\mathrm{d}t$ (the "time step") is a parameter chosen by the user. It repeats this process until some convergence criterion is met, and assigns the data points to their nearest subspaces $\{\mathbf{P}_i\}_{i=1}^K$ to obtain the K clusters. This is summarized in Algorithm 1.

Algorithm 1 Median K-flats (MKF)

Require: $\mathrm{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \subseteq \mathbb{R}^D$: data, normalized onto the unit sphere, d: dimension of subspaces, K: number of subspaces, $\{\mathbf{P}_i\}_{i=1}^K$: the initialized subspaces. $\mathrm{d}t$: step parameter.

Ensure: A partition of X into K disjoint clusters $\{\mathrm{X}\}_{i=1}^K$.

Steps:

1. Pick a random point $\mathbf{x}^*$ in X
2. Find its closest subspace $\mathbf{P}_{i^*}$, where
$$i^* = \mathrm{argmax}_{1 \le i \le K} ||\mathbf{P}_i \mathbf{x}||$$
3. Compute $\mathrm{d}_{\mathbf{x}^*}\mathbf{P}_{i^*}$ by Eq. (4.6)
4. Update $\mathbf{P}_{i^*}$: $\mathbf{P}_{i^*} \mapsto \mathbf{P}_{i^*} - \mathrm{d}t\, \mathrm{d}_{\mathbf{x}^*}\mathbf{P}_{i^*}$
5. Orthogonalize $\mathbf{P}_{i^*}$
6. Repeat steps 1-5 until convergence[1]
7. Assign each $\mathbf{x}_i$ to the nearest subspace

4.1.2 Complexity and Storage of the Algorithm

Note that the data set does not need to be kept in memory, so the storage requirement of the algorithm is $O(K \cdot d \cdot D)$, due to the K $d \times D$ matrices $\{\mathbf{P}_i\}_{i=1}^K$.

Finding the nearest subspace to a given point costs $O(K \cdot d \cdot D)$ operations. Computing the update costs $O(d \cdot D)$, and orthogonalizing $\mathbf{P}_{i^*}$ costs $O(d^2 \cdot D)$. Consequently, each iteration is $O(K \cdot d \cdot D + d^2 \cdot D)$. If n_s denotes the number of sampling iterations performed, then the total running time of the MKF algorithm is $O(n_s \cdot K \cdot d \cdot D + n_s \cdot d^2 \cdot D)$.

In our experiments we use $\mathrm{d}t = 0.01$. With this choice, the number of sampling iterations n_s is typically about 10^4. Usually n_s increases as the data becomes more complex (i.e., more flats, more outliers, etc), but in our experiments it never exceeded $3 \cdot 10^4$.

4.1.3 Initialization

Although the algorithm often works well with a random initialization of $\{\mathbf{P}_i\}_{i=1}^K$, it can many times be improved with a more careful initialization. We propose a farthest insertion method in Algorithm 2 below.

If the data has little noise and few outliers, then empirically, this initialization greatly increases the likelihood of obtaining the correct subspaces. On the other hand, in the case of sufficiently large noise or outliers, the initialization of Algorithm 2 does not work significantly better than random initializations, since the local structure of the data is obscured.

Notice that the initialization of Algorithm 2 also works for affine subspaces, so we can use it to initialize other iterative methods, such as K-flats.

4.1.4 Some Implementation Odds and Ends

Because the algorithm is randomized and the objective function may have many local minima, it is useful to restart the algorithm several times as often practiced in the K-flats algorithm. We can choose the best set of flats over all the restarts either measured in the l_1 sense or in the l_2 sense, depending on the application.

[1] In our experiments we checked the energy functional of Eq. (4.1) every 1000 iterations. We stopped if the ratio between current energy and the previous one was in the range (0.999,1.001). However, the computation of the energy functional depends on the size of the data. For large data sets we can obtain an online algorithm by replacing the ratio of the energy functionals, with e.g., the sum of squares of sines of principal angles between the corresponding subspaces.

Algorithm 2 Initialization for $\{\mathbf{P}_i\}_{i=1}^K$

Require: $\mathrm{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \in \mathbb{R}^{D\times N}$: data, d: dimension, K: number of d-flats

Ensure: $\{\mathbf{P}_i\}_{i=1}^K$: K subspaces.

For $i = 1$ to K, **do**

- If $i = 1$, Pick a random point $\hat{\mathbf{x}}$ in X; otherwise pick the point $\hat{\mathbf{x}}$ with the largest distance from the available planes $\{P_1, P_2, \cdots, P_{i-1}\}$
- Find the smallest integer j such that

$$\dim(\mathrm{Sp}(j\,\mathrm{NN}(\hat{\mathbf{x}}) - \hat{\mathbf{x}})) = d,$$

  where $j\,\mathrm{NN}(\hat{\mathbf{x}})$ denotes the set of j-nearest neighbors of $\hat{\mathbf{x}}$
- Let $\mathbf{P}_i$ be the affine space spanned by $\hat{\mathbf{x}}$ and $j\,\mathrm{NN}(\hat{\mathbf{x}})$

end

The MKF algorithm we have presented is designed for data sampled from linear subspaces of the same dimension. For affine subspaces, similar as in [10] we can add a homogeneous coordinate so that subspaces become linear. Empirically, it works well for clean cases with little noise or few outliers. However, we are still working on the true affine model, to make the algorithm more accurate and robust.

Also, for mixed dimensions of subspaces, i.e., when the dimensions $d_1, d_2, \cdots, d_K$ are not identical, we can set d to be $\max(d_1, d_2, \cdots, d_K)$ to implement the MKF algorithm (similarly as in [15]). Experiments show that this method works well if there exists a comparably small difference among $\{d_i\}_{i=1}^K$.

4.2 The Local Best-fit Flats Heuristic and the LBF and SLBF Algorithms

Our goal is to solve the HLM problem, that is, to partition a data set $\mathrm{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \subseteq \mathbb{R}^D$ into K clusters X_1, X_2, ..., X_K, with each cluster approximated by a d-dimensional affine subspace. In this section, we assume that all flats have the same known dimension d and that

the number of flats K is known. The cases of unknown K and mixed dimensions are addressed to some extent in Section 4.3. In this section we describe two methods that have at their heart an estimation of the locally affine scale of the data.

Both methods, LBF and SLBF, break into two parts. Their first part finds a set of candidate flats, i.e., affine subspaces. It starts by randomly selecting points from the data, and then chooses a "best neighborhood" around each point (see Section 4.2.1). The best fit flats (in L^2 sense) for these neighborhoods, which we refer to as local best-fit flats, are collected as *candidates*. We remark that the number of candidates is an input parameter in LBF.

The two algorithms process the candidates in different ways: LBF uses energy minimization and SLBF uses a spectral approach. Their sketches appear in Algorithms 4 and 6 and a more detailed explanation is in Sections 4.2.2 and 4.2.3.

4.2.1 Choosing the Optimal Neighborhood

Choosing the correct neighborhood is crucial for the success of both proposed methods, and is in some sense the central problem of this paper. If the neighborhood is too small, even if the point is in a good affine cluster, then a small amount of noise in the data will result in a flat which does not match most of the points in the affine cluster. If the neighborhood is too large, it will contain points from more than one affine cluster, and the resulting best fit flat will again not match any of the actual data points. While it is possible to take a guess at the correct scale as a parameter, we have found that it is possible to choose the correct scale reasonably well automatically.

What we will do is start at the smallest scale (say $d+1$) and look at larger and larger neighborhoods of a given point $\mathbf{x}_0$. At the smallest scale, any noise causes the local neighborhood to have higher dimension than d. As we add points to the neighborhood, it becomes better and better approximated in an average sense by its best fit flat, until points belonging to other flats enter the neighborhood. We thus take the neighborhood which is the first local minimum of the 'average error' to the neighborhoods best fit flat; for the average error β_2 of a neighborhood $\mathcal{N}$ of $\mathbf{x}_0$ to their best fit flat we use the formula:

$$\beta_2(\mathcal{N}) = \min_{d\text{-flats } L} \sqrt{\frac{\sum_{\mathbf{y}\in\mathcal{N}} ||\mathbf{y} - P_L\mathbf{y}||^2}{|\mathcal{N}|(\max_{\mathbf{x}\in\mathcal{N}} ||\mathbf{x}-\mathbf{x}_0||)^2}}, \tag{4.7}$$

where P_L denotes the projection onto the flat L. This notion of scaled error introduced and

utilized in [55, 38, 56], and considered recently in [57] for dimension estimation. The procedure we have just described is summarized in Algorithm 3.

Algorithm 3 Neighborhood size selection for HLM by randomized local best fit flats

Require: $\mathrm{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \subseteq \mathbb{R}^D$: data, $\mathbf{x}$: a point in X, S: start size, T: step size, l, m (optional): mean shifts parameters.

Ensure: $\mathcal{N}(\mathbf{x})$: a neighborhood of $\mathbf{x}$.

Steps:

• (Optional) Update the point $\mathbf{x}$ as the center of its l-nearest neighborhood in X, while repeating m times

• $k = -1$

repeat

• k:=k+1

• Let $\mathcal{N}_k$ be the set of the $S + kT$ nearest points in X to $\mathbf{x}$

• Set $\tilde{L}_k$ to be the best fit flat to $\mathcal{N}_k$

• Compute $\beta_2(k) := \beta_2(\mathcal{N}_k)$ according to (4.7)

until $k > 1$ and $\beta_2(k-1) < \min\{\beta_2(k-2), \beta_2(k)\}$

• Output $\mathcal{N}(\mathbf{x}) := \mathcal{N}_{k-1}$

The following theorem tries to justify our strategy of fitting the correct scale around each point. We work with a "geometric" set of assumptions in the continuous setting, where our data set will be presumed to be a collection of tubes around flats. This corresponds roughly to a probabilistic setting of sampling according to mixtures of uniform distributions around subsets of d-flats. For convenience we assume infinite tubes but restrict to local scales.

The analog of the discrete β_2 of (4.7) when having an underlying continuous set Ω in a ball of center $\mathbf{x}$ and radius r is defined as follows:

$$\beta_2(\mathbf{x}, r) = \min_L \sqrt{\int_{\Omega \cap B(\mathbf{x}, r)} \left(\frac{\mathrm{dist}(\mathbf{x}, L)}{r}\right)^2 \frac{\mathrm{d}\mathbf{x}}{\mathrm{Vol}(\Omega \cap B(\mathbf{x}, r))}},$$

where the minimum is over all d-flats L (see also [56]).

Theorem 4.2.1. *Let $K \geq 2$, $d < D$, L_i, $i = 1, \ldots, K$, be K d-flats in $\mathbb{R}^D$, and $\Omega_i := T(L_i, w_i)$ be K tubes in $\mathbb{R}^D$ around these flats of comparable widths $\{w_i\}_{i=1}^K$.*

For fixed $1 \leq i^* \leq K$ *and fixed* $\mathbf{x} \in L_{i^*}$*, let*

$$\mathbf{y} = \mathbf{y}(\mathbf{x}) = \arg\min_{\mathbf{y} \in \Omega \setminus \Omega_{i^*}} \text{dist}(\mathbf{y}, \mathbf{x}) \tag{4.8}$$

and

$$r_0 = r_0(\mathbf{x}) = \text{dist}(\mathbf{y}, \mathbf{x}). \tag{4.9}$$

Assume that $r_0 > w_{i^*}$*. Then the function* $\beta_2(\mathbf{x}, r)$ *is constant for* r *in* $[0, w_{i^*}]$*, comparable to a function which is decreasing for a sufficiently large subinterval of* $[w_{i^*}, r_0]$*, and satisfies the inequality*

$$\beta_2(\mathbf{x}, (1+\epsilon) \cdot r_0) \gtrapprox \beta_2(\mathbf{x}, r_0) \tag{4.10}$$

for sufficiently small ϵ*, i.e., it has an "approximate" local minimum in the interval* $[r_0, (1+\epsilon) \cdot r_0]$*. If* $d \leq 4$*, then* $\epsilon \approx w_{i^*}/r_0$*, and if* $d > 4$ *then* $\epsilon \approx (w_{i^*}/r_0)^{4/d}$*. As* w_{i^*}/r_0 *approaches zero, all comparability constants mentioned above approach one.*

4.2.2 The LBF Algorithm

The LBF algorithm searches for a good set of flats from the candidates (described above) in a greedy fashion. A measure of goodness of a K tuple of flats G is chosen; here, it will be the average l_1 distance of each point to its nearest flat, i.e., the energy G defined in (1.2). After randomly initializing K flats from the list of candidates, p passes are made through the data points. One of the current choices of flats is removed, and all the other candidates are tried in its place. If G decreases, we replace the current flat with the one which gives the lowest value for G. We then move to the next pass, picking a random flat, etc.

The simplest choice of G is the sum of the squared distances of each point in X to its nearest flat, i.e., having the power 2 in (1.2). However, in some special scenarios the l_1 energy of (1.2) is more robust to outliers than the mean squared error (see [58] and Theorem 2.5.3 for theoretical support). One can also imagine using spectral distances that measure the smoothness of the clusters with respect to some kernel, or many other global energy functionals of a partition. The nice thing about this method is that it allows for energy functionals which may be hard to minimize; since we are only testing the energy of our candidate configurations, as long as we can compute the energy of a partition quickly, we can run the greedy descent.

Algorithm 4 LBF: energy minimization over randomized local best-fit flats

Require: $\mathrm{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \subseteq \mathbb{R}^D$: data, d: dimension of subspaces, C: number of candidate planes, K: number of output flats/clusters, p: number of passes, S and T: parameters for local scale calculation.

Ensure: A partition of X into K disjoint clusters $\{\mathrm{X}_i\}_{i=1}^K$, each approximated by flats $\{F_i\}_{i=1}^K$.

Steps:

1. Pick C random points in X
2. For each of the C points find appropriate local scale using Algorithm 3
3. Generate C candidate flats $L_1, ..., L_C$ from the best fit flats to the neighborhoods from the previous step
4. Choose K flats from the candidates using Algorithm 5; collect these in $\mathcal{L}$
5. Partition X by sending points to nearest flats in $\mathcal{L}$

Algorithm 5 Greedy l_1 candidate selection for HLM by randomized local best fit flats

Require: $\mathrm{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \subseteq \mathbb{R}^D$: data, K: number of flats, $L_1, ..., L_C$: candidate flats, and p: number of passes.

Ensure: A set of K "active" flats $\mathcal{L} \subset \{L_1, ..., L_C\}$.

Steps:

Initialize $\mathcal{L}$ by randomly choosing K "active" flats $L_{A_1}, ..., L_{A_K}$

for pass $= 1$ to p **do**

Pick a random flat $L_{A_l} \subset \mathcal{L}$ $(1 \leq l \leq K)$

for $j = 1$ to $C - K$ **do**

- Pick one of the "inactive" flats L_j and form the collection of flats $\tilde{\mathcal{L}} = L_j \bigcup \mathcal{L} \setminus L_{A_l}$
- Set $s_j = \sum_{i=1}^N \min_{L \in \tilde{\mathcal{L}}} ||x_i - P_L x_i||$

end for

If $\min s_j < \sum_{i=1}^N \min_{L \in \{L_{A_1}, ..., L_{A_K}\}} ||x_i - P_L x_i||$, set $L_{A_l} := L_{\arg\min s_j}$

end for

4.2.3 The SLBF Algorithm

We can also process the candidate subspaces via a spectral clustering method. We first find the neighborhoods $\{\mathcal{N}_i\}_{i=1}^N$ for all points $\{\mathbf{x}_i\}_{i=1}^N$ via algorithm 3 and fit d-flats $\{L_i\}_{i=1}^N$ in these neighborhoods, then form the $N \times N$ matrices $\mathbf{S}$ and $\hat{\mathbf{S}}$ as follows:

$$\mathbf{S}_{i,j} = \sqrt{\text{dist}(\mathbf{x}_i, L_j)\,\text{dist}(\mathbf{x}_j, L_i)}, \tag{4.11}$$

and

$$\hat{\mathbf{S}}_{i,j} = \exp(-\mathbf{S}_{i,j}/2\sigma_j^2) + \exp(-\mathbf{S}_{i,j}/2\sigma_i^2), \tag{4.12}$$

where

$$\sigma_j = \sqrt{\min_{d\text{-flats L}} \sum_{\mathbf{x}\in\mathcal{N}_j} \text{dist}(\mathbf{x}, L)^2}. \tag{4.13}$$

Finally, we apply spectral clustering using the matrix $\hat{\mathbf{S}}$. More precisely, we follow the main algorithm of [53, Section 2], while replacing the matrix A there by $\hat{\mathbf{S}}$, multiplying the unit eigenvectors of Step 3 (of [53, Section 2]) by the corresponding square roots of eigenvalues and skipping step 4. We remark that the two last changes to [53, Section 2] are common to spectral-based manifold learning algorithms (unlike spectral clustering), so that the similarity matrix $\hat{\mathbf{S}}$ is considered as a gram matrix, see e.g., Euclidean MDS [59] and ISOMAP [60]. We describe this algorithm in Algorithm 6, and refer to it as the SLBF (spectral LBF) algorithm.

4.2.4 Adaptation of the Proposed Algorithms to Motion Segmentation Data

Note that the first minimum in the Theorem 4.2.1 excludes the left endpoint, and thus $k = 0$ is excluded in Algorithm 3). In our experiments, we noticed that on data without too much noise, it is useful to allow the first scale to count as a local minimum and allow $k = 0$ in Algorithm 3). We refer to the implementation of LBF and SLBF with those two techniques tailored for motion segmentation data as LBF-MS and SLBF-MS.

4.2.5 Complexity and Storage of LBF and SLBF

We first discuss the complexity of Algorithm 3, which is used in both the LBF and the SLBF algorithms. In order to obtain $\beta_2(\mathcal{N}_k)$, we need to obtain the top d singular values of the $|\mathcal{N}_k| \times D$ data matrix representing the $|\mathcal{N}_k|$ points, which requires a complexity of $O(d \cdot D \cdot |\mathcal{N}_k|)$.

Algorithm 6 SLBF: spectral clustering based on best-fit flats

Require: $\mathrm{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \subseteq \mathbb{R}^D$: data, λ: a parameter (or several parameters if we use step 7, with default values $[2, 2e, 2e^2, \cdots, 2e^6]$), other parameters used by Algorithm 3.

Ensure: A partition of X into K disjoint clusters $\{\mathrm{X}_i\}_{i=1}^K$, each approximated by a single flat.

Steps:

For each point $\mathbf{x}_i$, fit a subspace L_i by Algorithm 3

Construct the $N \times N$ matrix $\mathbf{S}$ and $\hat{\mathbf{S}}$ by (4.11), (4.12) and (4.13)

Let $\mathbf{D}$ be the $N \times N$ diagonal matrix, such that $\mathbf{D}_{i,i} = \sum_{j=1}^N \hat{\mathbf{S}}_{i,j}$

Normalize $\hat{\mathbf{S}}$ by: $\hat{\mathbf{S}} = \mathbf{D}^{-\frac{1}{2}} \hat{\mathbf{S}} \mathbf{D}^{-\frac{1}{2}}$

Let U be the $N \times K$ matrix whose columns are the top K eigenvectors of $\hat{\mathbf{S}}$, and Σ be the $K \times K$ matrix representing the top K eigenvalues of $\hat{\mathbf{S}}$

Apply K-means to the rows of $N \times K$ matrix $U\,\Sigma^{1/2}$ and partition X accordingly

(optional) repeat steps 2-6 with various values of λ in (4.13) to obtain several segmentation, and choose the segmentation with the smallest following error:

$$\sum_{i=1}^{K} \min_{d\text{-flat } L} \left(\sum_{\mathbf{x} \in \mathrm{X}_i} \mathrm{dist}^2(\mathbf{x}, L_i) \right) \tag{4.14}$$

To find $\mathcal{N}(x)$, we need to generate $\beta_2(\mathcal{N}_k)$ for any $|\mathcal{N}_k| = S + kT$, where $k = 1, 2, \cdots, (N - S)/T$, hence the complexity for obtaining $\mathcal{N}(x)$ is of order:

$$O(d \cdot D \cdot \sum_{k=1}^{(N-S)/T} (S + kT)) \leq O(d \cdot D \cdot N^2/2T).$$

We thus note that if T is in the order of N, e.g., $T = \max(N/300, 2)$, the total complexity of Algorithm 3 is $O(d \cdot D \cdot N)$.

Next, we clarify the complexity of Algorithm 4. For step 2 of this algorithm we need to run Algorithm 3 C times and thus its complexity of order $O(C \cdot d \cdot D \cdot N)$. Step 3 of Algorithm 4, requires C SVD decompositions for C matrices of size at most $N \times D$, in order to obtain the first d vectors in $\mathbb{R}^D$. It thus also has a complexity at most $O(C \cdot d \cdot D \cdot N)$.

Step 4 of Algorithm 4 is an application of Algorithm 5. The later algorithm requires the evaluation of the $N \times C$ matrix representing the distances $||x_i - P_{L_j} x_i||$ between $\mathrm{X} = \{x_1, x_2, \cdots, x_N\}$ and $L_1, L_1, \cdots, L_C$. This costs $O(C \cdot d \cdot D \cdot N)$ operations, since each distance from a point to a subspace costs $O(d \cdot D)$. Moreover, the p passes in Algorithm 5 have complexity of order $O(p \cdot (C - K) \cdot N)$. Therefore, step 4 of Algorithm 4 has a complexity of order $O(C \cdot N \cdot (d \cdot D + p))$. At last, Step 5 of Algorithm 4 has a complexity of order $O(K \cdot d \cdot D \cdot N)$, which comes from the construction of the $N \times K$ matrix of distances from N points to K subspaces. Combining these complexities together, we have an overall complexity of $O(C \cdot N \cdot (d \cdot D + p))$ for the LBF Algorithm.

As for the storage of LBF, we note that the data set is saved in an $N \times D$ matrix, the candidate subspaces are organized in C projection matrices of size $D \times d$ and that the algorithm also requires an $N \times C$ matrix of distances between the data points and the C candidate subspaces. Therefore, the storage of LBF is in the order of $O(D \cdot N + C \cdot D \cdot d + N \cdot C)$.

We conclude with the complexity and storage of the SLBF algorithm. Step 1 of Algorithm 6 has a complexity of $O(d \cdot D \cdot N^2)$, since Algorithm 3 has a complexity of $O(d \cdot D \cdot N)$ and it is applied to every point in the set X. The most expensive calculation of steps 1-1 in Algorithm 6 is the construction of $\mathbf{S}$, which requires a complexity of $O(d \cdot D \cdot N^2)$. The SVD decomposition in step 1 has a complexity of $O(K \cdot N^2)$ and the K-means algorithm in step 1 has a complexity of $O(n_s \cdot K \cdot N \cdot D)$, where n_s is the iterations in K-means.

Combining these complexities together, we have an overall complexity of $O((K + d \cdot D) \cdot N^2 + n_s \cdot K \cdot D \cdot N)$ for the SLBF Algorithm. Moreover, it stores the data set in a $D \times N$

matrix, the candidate subspaces in N $D \times d$ matrices (recall that in SLBF every data point is assigned a subspace and thus $C = N$) and it also uses the $N \times N$ matrix $\mathbf{S}$. Therefore , the storage of the SLBF Algorithm is in the order of $O(N \cdot D \cdot d + N^2)$.

4.3 Experimental Results

In this section, we conduct experiments on artificial and real data sets to verify the effectiveness of the proposed algorithm in comparison to other hybrid linear modeling (HLM) algorithms.

We measure the accuracy of those algorithms by the rate of misclassified points with outliers excluded, that is

$$\text{error}\% = \frac{\text{\# of misclassified inliers}}{\text{\# of total inliers}} \times 100\% \,. \tag{4.15}$$

In all the experiments below, the number C in Algorithm 5 is $70 \cdot K$, where K is the number of subspaces, the number p in Algorithm 5 is $5 \cdot K$, and the numbers S and T in Algorithm 3 are $2 \cdot d$ and 2 respectively, where d is the dimension of the subspace. According to our experience, LBF and SLBF are very robust to changes in parameters, but unsurprisingly, there is a general trade off between accuracy (higher C, higher p, smaller T), and run time (lower C, lower p, larger T). We have chosen these parameters for a balance between run time and accuracy. Nevertheless, we have insisted to use the same parameters for all data sets and experiments, even though particular parameters could obtain even better or near perfect results for some of the data sets. These experiments run on a computer with Intel Core 2 CPU at 2.66GHz and 2 GB memory.

4.3.1 Clustering Results on Simulated Data

We compare our algorithm with the following algorithms: Mixtures of PPCA (MoPPCA) [4], K-flats (KF) [9], Local Subspace Analysis (LSA) [11], Spectral Curvature Clustering (SCC) [15], Random Sample Consensus (RANSAC) for HLM [12], Agglomerative Lossy Compression (ALC) [14] and GPCA with voting (GPCA) [13]. Throughout the rest of the paper, we use the Matlab codes of the GPCA, MoPPCA, RANSAC and KF algorithm from http://perception.csl.uiuc.edu/gpca, the SCC algorithm from http://www .math.umn.edu/∼lerman/scc, the LSA algorithm from http://www.vision.jhu.edu/db, the ALC algorithm from http://perception.csl.uiuc.edu/coding/m and the SSC algorithm from http://www.cis.jhu.edu/∼ehsan/ssc.htm.

Table 4.1: Mean percentage of misclassified points in simulation.

Linear	$2^2 \in \mathbb{R}^4$		$4^2 \in \mathbb{R}^6$		$2^4 \in \mathbb{R}^4$		$10^2 \in \mathbb{R}^{15}$		$(4,5,6) \in \mathbb{R}^{10}$	
Outl. %	5	30	5	30	5	30	5	30	5	30
LSCC	3.0	6.9	2.3	2.6	7.7	22.4	0.5	3.8	1.8	28.2
LSCC-MS	3.8	10.0	2.4	4.1	8.5	36.7	0.7	31.9	1.4	19.8
LSA	18.7	19.6	10.9	12.7	44.3	21.0	7.6	9.9	6.1	6.6
KF	3.0	15.8	2.5	18.4	9.4	34.3	0.8	33.8	0.8	30.6
MoPPCA	3.1	14.2	2.5	17.7	8.4	34.2	0.9	38.8	1.4	34.7
GPCA	19.7	30.9	11.7	35.9	29.2	43.9	10.2	42.6	10.1	45.4
LBF	2.6	3.7	2.5	2.3	6.4	11.5	1.3	1.9	1.5	1.5
LBF-MS	2.7	3.0	2.6	2.6	6.6	11.7	1.7	2.2	1.4	1.5
SLBF	5.2	6.3	5.6	7.0	18.5	23.9	5.4	6.2	2.7	2.4
SLBF-MS	8.7	11.7	5.9	6.6	33.5	46.6	3.9	4.8	2.4	2.6
RANSAC (oracle)	3.3	2.6	2.3	2.2	8.6	9.8	0.9	6.7	1.8	1.4
RANSAC (ϵ from LBF)	2.4	2.8	2.2	2.5	5.8	7.5	30.4	42.8	0.7	13.5
ALC (oracle)	4.0	3.4	13.1	16.3	27.5	30.1	50.0	50.0	5.3	36.1
ALC (ϵ from LBF)	4.1	5.7	8.7	10.0	9.9	14.0	50.0	50.0	2.3	1.8
SSC	24.8	34.3	32.2	43.5	49.2	52.8	18.9	44.9	32.4	54.0
MKF	4.6	5.8	2.0	2.1	9.6	18.8	0.1	0.1	0.9	0.7

Affine	$2^2 \in \mathbb{R}^4$		$4^2 \in \mathbb{R}^6$		$2^4 \in \mathbb{R}^4$		$10^2 \in \mathbb{R}^{15}$		$(4,5,6) \in \mathbb{R}^{10}$	
Outl. %	5	30	5	30	5	30	5	30	5	30
SCC	0.0	0.6	0.0	0.0	0.2	0.5	0.0	0.7	0.0	5.8
SCC-MS	0.0	2.2	0.0	0.5	1.4	5.8	0.0	0.0	0.0	3.1
LSA	11.8	11.0	5.3	4.7	45.0	41.7	0.0	0.0	1.0	1.1
KF	7.3	15.1	9.9	26.0	19.7	37.1	11.1	24.9	7.3	23.5
MoPPCA	25.6	23.7	27.8	38.3	45.5	39.8	37.1	45.2	42.9	46.8
GPCA	13.8	14.4	22.6	22.1	33.6	32.4	36.0	29.6	26.7	29.1
LBF	0.2	2.0	0.0	0.7	0.3	4.5	0.0	0.3	0.0	0.0
LBF-MS	0.2	2.7	0.1	1.5	0.8	5.2	0.0	0.5	0.0	0.0
SLBF	0.0	1.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
SLBF-MS	0.0	0.1	0.0	0.0	0.2	0.1	0.0	0.0	0.0	0.0
RANSAC (oracle)	13.2	12.2	11.5	11.2	31.5	28.4	2.6	9.2	1.1	2.2
RANSAC (ϵ from LBF)	6.3	41.2	10.1	47.9	10.5	45.8	40.7	N/A	N/A	61.8
ALC (oracle)	0.0	0.0	0.0	0.0	35.9	25.1	0.0	40.0	0.0	65.0
ALC (ϵ from LBF)	0.7	0.4	0.0	0.0	0.7	0.3	0.0	0.0	0.0	0.0
SSC	1.1	1.9	0.1	0.1	6.6	6.4	0.0	0.0	0.0	0.0

Table 4.2: Mean running time for linear-subspaces cases and affine-subspaces cases.

Linear	$2^2 \in \mathbb{R}^4$		$4^2 \in \mathbb{R}^6$		$2^4 \in \mathbb{R}^4$		$10^2 \in \mathbb{R}^{15}$		$(4,5,6) \in \mathbb{R}^{10}$	
Outl. %	5	30	5	30	5	30	5	30	5	30
LSCC	0.7	0.8	16.0	1.8	2.1	2.0	13.3	5.7	5.1	8.4
LSCC-MS	0.5	0.5	1.2	1.4	1.7	1.5	5.1	5.6	4.0	4.6
LSA	8.8	16.0	11.1	20.8	28.3	54.4	31.3	31.5	38.2	54.4
KF	0.5	0.6	0.5	0.8	1.4	1.8	1.9	1.0	1.1	2.8
MoPPCA	0.2	0.5	0.3	0.7	1.2	2.0	1.7	1.1	1.0	3.3
GPCA	3.5	7.6	9.8	19.0	20.9	29.7	30.3	31.6	39.1	57.8
LBF	0.5	0.5	0.5	0.5	2.2	2.7	0.7	0.8	1.2	1.4
LBF-MS	0.4	0.5	0.4	0.5	2.0	2.6	0.5	0.6	1.0	1.3
SLBF	10.5	20.7	11.8	21.7	90.1	174.9	12.0	23.3	31.3	64.2
SLBF-MS	13.2	24.0	13.1	24.4	152.0	202.0	13.2	23.5	39.5	72.4
RANSAC (oracle)	8.0	9.5	15.7	16.1	14.8	18.3	44.4	46.4	104.0	103.7
RANSAC (ϵ from LBF)	9.2	11.4	17.9	19.0	19.3	23.8	48.7	44.9	114.3	141.6
ALC (oracle)	12.0	23.2	16.2	33.6	68.7	136.3	37.7	172.6	53.1	180.1
ALC (ϵ from LBF)	18.9	28.0	19.6	37.9	59.2	121.9	73.1	152.4	79.7	151.6
SSC	162.8	236.2	170.8	247.9	382.7	591.3	184.1	276.6	298.3	437.9
MKF	3.2	4.5	5.9	5.4	7.2	6.5	8.3	8.3	7.0	10.7

Affine	$2^2 \in \mathbb{R}^4$		$4^2 \in \mathbb{R}^6$		$2^4 \in \mathbb{R}^4$		$10^2 \in \mathbb{R}^{15}$		$(4,5,6) \in \mathbb{R}^{10}$	
Outl. %	5	30	5	30	5	30	5	30	5	30
SCC	0.9	1.0	1.7	2.0	5.1	2.5	6.1	13.7	5.6	6.0
SCC-MS	0.7	0.7	1.4	1.6	2.2	2.2	5.4	6.0	4.6	4.8
LSA	8.7	16.1	11.1	20.8	28.6	54.0	21.1	32.2	38.3	54.0
KF	0.5	0.6	0.6	0.7	2.4	1.4	0.6	1.7	1	1.4
MoPPCA	0.5	0.5	0.7	0.6	2.9	1.4	1.3	1.9	1.9	2.0
GPCA	2.4	6.9	5.1	9.8	11.2	26.1	20.2	31.9	38.4	49.9
LBF	0.5	0.6	0.5	0.6	2.2	2.8	0.7	0.8	1.2	1.5
LBF-MS	0.4	0.5	0.4	0.5	2.0	2.7	0.5	0.6	1.0	1.3
SLBF	9.4	19.1	8.8	19.0	71.8	143.1	9.2	19.4	35.1	61.4
SLBF-MS	10.5	21.7	10.1	21.9	79.9	175.5	10.4	21.1	40.1	66.7
RANSAC (oracle)	12.0	14.4	19.6	20.6	23.9	29.5	48.2	51.5	79.7	84.9
RANSAC (ϵ from LBF)	13.4	12.6	22.7	23.5	28.4	32.1	49.3	N/A	N/A	60.6
ALC (oracle)	13.3	25.2	15.2	39.1	61.0	119.3	18.5	43.0	39.7	92.7
ALC (ϵ from LBF)	13.4	26.8	15.6	29.8	55.2	113.6	29.8	55.5	47.9	85.2
SSC	160.2	226.8	176.0	255.3	386.6	592.4	202.9	311.9	338.6	504.1

For the SCC algorithm, we also tried a slightly modified version tailored for motion segmentation as in step 1 of Algorithm 6, which we refer to as SCC-MS (SCC for motion segmentation): Following the notation of [Algorithm 2][15] we let the matrix $\mathbf{U}$ be the $N \times K$ matrix whose columns are the top K left singular vectors of $\mathbf{A}_C^*$ and also denote by Σ the diagonal $K \times K$ matrix whose elements are the top K left singular values of $\mathbf{A}_C^*$. Then the K-means step of SCC-MS is applied directly to the rows of the $N \times K$ matrix $U\,\Sigma^{1/2}$ (as opposed to applying it to U (or its row-wise normalization by 1) in SCC).

The MoPPCA algorithm is always initialized with a random guess of the membership of the data points. The LSCC algorithm is initialized by randomly picking $100 \times K$ $(d+1)$-tuples (following [15]), and KF are initialized with random guess. Since algorithms like KF tend to converge to local minimum, we use 10 restarts for MoPPCA, 30 restarts for KF, and recorded the misclassification rate of the one with the smallest l_2 error for both of these algorithms. The number of restarts was restricted by the running time and accuracy. In SSC algorithm, we set the value λ to be 0.01, as suggested in the code.

The RANSAC and ALC algorithms for HLM [12, 14] depends on a user supplied inlier threshold. RANSAC (oracle) and ALC (oracle) uses the oracle inlier bound given by the true noise variance of the model and thus clearly has an advantage over the other algorithms listed. RANSAC (ϵ from LBF) and ALC (ϵ from LBF) estimates the inlier threshold by the local best fit flats heuristic of this paper. That is, it fits N neighborhoods for all N points using this heuristic and estimates the least error of approximation by d-flat in these N neighborhoods. The inlier bound ϵ is then the average of these errors. For some cases this algorithm run into error, then the result is reported as N/A. The reason for this is that the RANSAC algorithm for HLM [12] is very sensitive to the estimate of ϵ and an overestimate can result in removal of points belonging to more than one subspace, so that the algorithm may exhaust all points before detecting K subspaces. ALC (ϵ from LBF) may result in a number of clusters different than K, though we still use the same identification error as in (4.15), while comparing the true label with all permutations of the computed label and use the one with smallest error. We remark that due to outliers we could not effectively use the voting procedure for ALC described later.

The simulated data represents various instances of K linear subspaces in $\mathbb{R}^D$. If their dimensions are fixed and equal d, we follow [15] and refer to the setting as $d^K \in \mathbb{R}^D$. If they are mixed, then we follow [13] and refer to the setting as $(d_1, \ldots, d_K) \in \mathbb{R}^D$. Fixing K and d (or $d_1, \ldots, d_K$), we randomly generate 100 different instances of corresponding hybrid linear

Table 4.3: The mean and median percentage of misclassified points.

2-motion	Checker		Traffic		Articulated		All	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
GPCA	6.09	1.03	1.41	0.00	2.88	0.00	4.59	0.38
LLMC 5	4.37	0.00	0.84	0.00	6.16	1.37	3.62	0.00
LSA $4K$	2.57	0.27	5.43	1.48	4.10	1.22	3.45	0.59
LBF($4K$,3)	3.65	0.00	3.89	0.00	4.43	0.15	3.78	0.00
LBF-MS($4K$,3)	2.90	0.00	1.64	0.00	2.51	0.06	2.54	0.00
SLBF($2F$,3)	1.59	0.00	0.20	0.00	0.80	0.00	1.16	0.00
SLBF-MS($2F$,3)	1.28	0.00	0.21	0.00	0.94	0.00	0.98	0.00
SCC($4K$,3)	2.42	0.00	4.44	0.00	9.51	2.04	3.60	0.00
SCC-MS($4K$,3)	2.00	0.00	0.35	0.00	4.11	1.12	1.77	0.00
SSC-N($4K$,3)	1.29	0.00	0.29	0.00	0.97	0.00	1.00	0.0
MSL	4.46	0.00	2.23	0.00	7.23	0.00	4.14	0.00
RANSAC	6.52	1.75	2.55	0.21	7.25	2.64	5.56	1.18
MKF	3.70	0.00	0.90	0.00	6.80	0.00	3.26	0.00
3-motion	Checker		Traffic		Articulated		All	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
GPCA	31.95	32.93	19.83	19.55	16.85	28.66	28.66	28.26
LLMC $4K$	12.01	9.22	7.79	5.47	9.38	9.38	11.02	6.81
LLMC 5	10.70	9.21	2.91	0.00	5.60	5.60	8.85	3.19
LSA $4K$	5.80	1.77	25.07	23.79	7.25	7.25	9.73	2.33
LSA 5	30.37	31.98	27.02	34.01	23.11	23.11	29.28	31.63
LBF($4K$,3)	8.50	1.26	16.31	13.52	20.75	20.75	10.77	1.70
LBF-MS($4K$,3)	6.97	1.15	7.06	0.62	21.38	21.38	7.81	0.98
SLBF($2F$,3)	4.57	0.94	0.38	0.00	2.66	2.66	3.63	0.64
SLBF-MS($2F$,3)	3.33	0.39	0.24	0.00	2.13	2.13	2.64	0.22
SCC($4K$,3)	7.80	1.04	8.05	2.37	7.07	7.07	7.81	0.67
SCC-MS($4K$,3)	6.28	0.80	4.09	0.58	7.22	7.22	5.89	0.68
SSC-N($4K$,3)	3.22	0.29	0.53	0.00	2.13	2.13	2.62	0.22
MSL	10.38	4.61	1.80	0.00	2.71	2.71	8.23	1.76
RANSAC	25.78	26.01	12.83	11.45	21.38	21.38	22.94	22.03
MKF	14.50	12.00	3.06	0.01	15.90	15.90	12.29	6.23

models according to the code in http://perception.csl.uiuc.edu/gpca. More precisely, for each of the 100 experiments, K linear subspaces of the corresponding dimensions in $\mathbb{R}^D$ are randomly generated. The random variables sampled within each subspace are sums of two other variables. One of them is sampled from a uniform distribution in a d-dimensional ball of radius 1 of that subspace (centered at the origin for the case of linear subspaces). The other is sampled from a D-dimensional multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $0.05^2 \cdot \mathbf{I}_{D\times D}$. Then, for each subspace 250 samples are generated according to the distribution just described. Next, the data is further corrupted with 5% or 30% uniformly distributed outliers in a cube of sidelength determined by the maximal distance of the former 250 samples to the origin (using the same code). Since most algorithms (SCC, LSA, MoPPCA, LBF, SLBF, RANSAC, SSC) do not support mixed dimensions natively, we assume each subspace has the maximum dimension in the experiment. GPCA and ALC support mixed dimensions natively, and GPCA is the only algorithm for which we specify the dimensions for each subspace in mixed-dimension case (note that the knowledge of dimensions are unnecessary in ALC algorithm).

The mean (over 100 instances) misclassification rates of the various algorithms is recorded in Table 4.1. The mean running time is shown in Table 4.2. From Table 4.1 we can see that our algorithms, i.e., LBF, LBF-MS, SLBF, SLBF-MS, perform well in various artificial instances of hybrid linear modeling (with both linear subspaces and affine subspaces), and their advantage is especially obvious with many outliers and affine subspaces. The robustness to outliers is a result of our use of both l_1 loss function (see e.g., [58]) and random sampling. The SLBF and SLBF-MS are better at the affine cases because of their use of spectral clustering. Also unlike many other methods, the proposed methods natively supports affine subspace models (the particular data has non-intersecting subspaces, which makes advantageous to some other algorithms, e.g., SSC). The results of RANSAC (ϵ from LBF) and ALC (ϵ from LBF) show that the local best-fit heuristic can be effectively used to estimate the main parameter of RANSAC and ALC, i.e., to estimate the local noise. Table 4.2 shows that the running time of LBF/LBF-MS is less than the running time of most other algorithms, especially GPCA, LSA, RANSAC, ALC and SSC. The difference is large enough that we can also use the proposed algorithm as an initialization for the others. LBF and LBF-MS algorithms are slower than a single run of K-flats, but it usually takes many restarts of K-flats to get a decent result. Notice that the choice of C and p in our algorithm function in a similar manner to the number of restarts in KF. SLBF and SLBF-MS

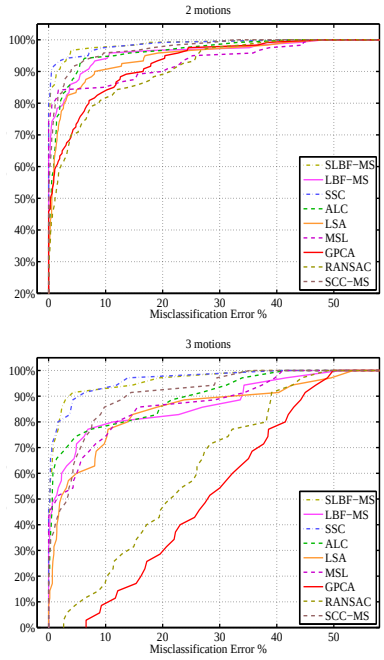


Figure 4.1: The misclassification rate of some algorithms for the Hopkins 155 database.

cost more time when N is large, because of the construction of the $N \times N$ matrix in spectral clustering, but it is still faster than SSC, which also depends on a spectral matrix.

4.3.2 Clustering results on motion segmentation data

We test the proposed algorithms on the Hopkins 155 database of motion segmentation, which is available at http://www.vision.jhu.edu/data/hopkins155. This data contains 155 video sequences along with the coordinates of certain features extracted and tracked for each sequence in all its frames. The main task is to cluster the feature vectors (across all frames) according to the different moving objects and background in each video.

More formally, for a given video sequence, we denote the number of frames by F. In each sequence, we have either one or two independently moving objects, and the background can also move due to the motion of the camera. We let K be the number of moving objects plus the background, so that K is 2 or 3 (and distinguish accordingly between two-motions and three-motions). For each sequence, there are also N feature points $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N \in \mathbb{R}^3$

that are detected on the objects and the background. Let $\mathbf{z}_{ij} \in \mathbb{R}^2$ be the coordinates of the feature point $\mathbf{y}_j$ in the i^{th} image frame for every $1 \leq i \leq F$ and $1 \leq j \leq N$. Then $\mathbf{z}_j = [\mathbf{z}_{1j}, \mathbf{z}_{2j}, \cdots, \mathbf{z}_{Fj}] \in \mathbb{R}^{2F}$ is the trajectory of the j^{th} feature point across the F frames. The actual task of motion segmentation is to separate these trajectory vectors $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N$ into K clusters representing the K underlying motions.

It has been shown [2] that under the affine camera model, the trajectory vectors corresponding to different moving objects and the background across the F image frames live in distinct affine subspaces of dimension at most three in $\mathbb{R}^{2F}$. Following this theory, we implement our algorithm with $d = 3$.

Table 4.4: Average total computation times for all 155 sequences.

RANSAC	LBF-MS	LBF	SCC-MS	SLBF-MS	SLBF	MKF	SSC
60s	73s	91s	196s	28min	31min	236min	427min

We compare our algorithm with the following ones: improved GPCA for motion segmentation (GPCA) [61], K-flats (KF) [9] (implemented for linear subspaces), Local Linear Manifold Clustering (LLMC) [62], Local Subspace Analysis (LSA) [11], Multi Stage Learning (MSL) [63], Spectral Curvature Clustering (SCC) [15] and SCC-MS (see description earlier), Sparse Subspace Clustering (SSC) [64], and RANSAC for HLM [12].

For GPCA (improved for motion segmentation), LLMC, LSA, MSL and RANSAC (for HLM), we copy the results from http://www.vision.jhu.edu/data/hopkins155 (they are based on experiments reported in [65] and [62]). We perform our own experiments for SCC, SCC-MS, SSC-N (SSC-B is not reported since it did not perform as well as SSC-N), LBF, LBF-MS, SLBF, SLBF-MS, we perform the experiments on our own and record the mean misclassification rate and the median misclassification rate for each algorithm for any fixed K (two or three-motions) and for the different type of motions ("checker", "traffic" and "articulated"). Each experiment (testing the latter set of algorithms) was repeated 500 times. The average misclassification rates and running time are recorded in Table 4.3 and demonstrated in Figure 4.1. In Figure 4.1, the y-axis represent the percentage of data sets that have misclassification rates (under corresponding algorithms) lower than that of x-axis.

Our misclassification errors for SCC are different than [66] and [67] and our misclassification errors for SSC are different than [64] (the difference between our and their results differ more than twice the standard deviations of errors obtained here). This can be explained by

possible evolutions of the codes since then (at least for SSC). We remark though that the misclassification errors of SCC-MS here are even slightly better than the misclassification errors of SCC in [66].

From Table 4.3 and Figure 4.1 we can see that our algorithms work well for the Hopkins database. Of all the methods tested, SLBF-MS and SSC-N are the most accurate algorithms. Besides SLBF/SLBF-MS and SSC-N, only SSC-MS is better than LBF-MS. However, From Table 4.4, LBF-MS ran more than 100 times faster than SSC-N and SLBF-MS is also more than 10 times faster than SSC. In many of the cases, the l_1 energy (as well as the l_2 energy) was lower for the labels obtained by LBF than the true labels. We thus suspect that the reason SLBF/SLBF-MS works better than LBF/LBF-MS is that good clustering of the Hopkins data requires additional type of information (e.g., spectral information) to be combined with subspace clustering (i.e., hybrid linear modeling).

Chapter 5

Discussion

In the minimization of (1.1), 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 [39] as well as sequential HLM as in [12] (while using l_p minimization with $0 < p \leq 1$ in the spirit of [21, 22]).

In the minimization of (1.2), we studied the effectiveness of l_p minimization for recovering all underlying K subspaces when identically and independently sampling from a spherically symmetric HLM measure. In particular, we demonstrated a phase transition phenomenon around $p = 1$. We also showed how to generalize this study in order to nearly recover the subspaces in the case of additive noise.

We discuss here possible extensions, counterexamples indicating impossibility of extensions and some open directions.

5.1 Implementation and Relation to Other Algorithms

To minimize (1.1), one can approximate the geometric l_1 minimizer by gradient descent or stochastic gradient descent (see e.g., [17]). 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 [33] and the outlier pursuit [40]. It is possible that by carefully choosing the tuning parameter of [40], the rows of the low rank matrix obtained by [40] span the d-subspace that minimizes the l_1 energy in (1.1).

5.2 Obstacles for Convex Recovery of Multiple Subspaces

When we fix X, the minimization of (1.2) is not a geodesically convex problem, since $\mathrm{G}(D,d)^K$ is not a geodesically convex set. For example, when $K=1$, $D=2$ and $d=1$, the geodesic arc between subspaces $x_1=0$ and $x_2=0$ is not unique: it can either pass through $x_1=x_2$ or $x_1=-x_2$.

There are various methods for recovering or nearly recovering a single subspaces by convex optimization. In particular, we commented in Proposition 3.1.2 on the possibility to relate our work on l_p-recovery of a single subspace to such efforts. Given a function f on $\mathrm{G}(D,d)$, these efforts look for a convex set $\mathbb{H}$, function $g:\mathbb{H}\to\mathrm{G}(D,d)$ and a convex function $h:\mathbb{H}\to\mathbb{R}$ with minimizer $\hat{\mathbf{x}}$, such that $g(\hat{\mathbf{x}})$ is the minimizer of f. However, these strategies can not extend to multiple subspaces. Indeed, a successful convex optimization strategy needs to be based on a convex and permutation-invariant function $h_{\mathrm{X}}(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_K)$ whose set of minimizers is all permutations of $\hat{\mathbf{x}}_1,\hat{\mathbf{x}}_2,\cdots,\hat{\mathbf{x}}_K$, where $\hat{\mathbf{x}}_i\in g^{-1}(\mathrm{L}_i^*)$ for all $1\le i\le K$. However, the convexity of such a function implies that

$$\left(\frac{\sum_{i=1}^K\hat{\mathbf{x}}_i}{K},\frac{\sum_{i=1}^K\hat{\mathbf{x}}_i}{K},\cdots,\frac{\sum_{i=1}^K\hat{\mathbf{x}}_i}{K}\right)$$

is also a minimizer of h_{X}. This contradicts our assumption and also implies a rather large set of minimizers for such a convex energy h_{X}.

5.3 Extending Our Theory by More General Distributions

The strict spherical symmetry of the distributions $\{\mu_i\}_{i=0}^K$ in Theorems 2.4.1 and 2.4.2 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.,

$$\alpha_0 < C_\star(\mu_0, \mu_1), \tag{5.1}$$

and the condition

$$C_1(\mu_1)\alpha_1 > \sum_{i=2}^{K} C_i(\mu_i)\alpha_i + C_0(\mu_0)\alpha_0 \tag{5.2}$$

instead of (2.3). We also need to replace the corresponding part of the denominator of (2.4) 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 3.1.2 by assuming that both μ_0 and μ_1 are approximately spherically symmetric (with bounded support) as well conditions (5.1) and (5.2). 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 3.1.2 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 (3.91).

In Theorems 2.5.1 and 2.5.2, the strict symmetry of the distribution generating $\{\mu_i\}_{i=1}^K$ is not necessary. Indeed, it is enough to assume that these distributions are supported within bounded sets of $\mathrm{L}_i^*\}_{i=1}^K$ respectively and that each of them is approximately spherically symmetric. That is, its derivative with respect to a spherically symmetric distribution with the same support is bounded away from 0 and ∞. However, this will require the following stronger conditions instead of (2.5) and (2.6):

$$\alpha_0 < \tau_0 \cdot \min_{1\le i\le K} C_i \cdot \alpha_i \cdot \left(1 \wedge \min_{1\le i,j\le K} \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_i^*, \mathrm{L}_j^*)^p / 2^p\right) \tag{5.3}$$

and

$$\epsilon < 3^{-\frac{1}{p}} \left(\tau_0 \cdot \min_{1\le i\le K} C_i \cdot \alpha_i \cdot \left(1 \wedge \min_{1\le i,j\le K} \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_i^*, \mathrm{L}_j^*)^p / 2^p\right) - \alpha_0\right)^{\frac{1}{p}}, \tag{5.4}$$

where $C_i \equiv C_i(\mu_i, p)$ for $1 \le i \le K$.

We remark that the underlying distributions of Theorem 2.5.3 are already pretty general.

5.4 Distributions Resulting in Counterexamples for our Theory

There are several typical cases with settings different than above, where the underlying subspaces cannot be recovered by minimizing the energy (1.2) for all $p > 0$. The first typical example is when there is an outlier with sufficiently large magnitude so that the minimizer of (1.2) contains a subspace passing through this outlier, which is different than any of the underlying subspaces.

The second example is when the distribution of outliers lies on another subspace, $\mathrm{L}_0^* \in \mathrm{G}(D,d)$, and $\alpha_0 > \min_{1\leq i\leq K} \alpha_i$, then L_0^* is contained in the minimizer of (1.2). In both cases the outliers are not approximately spherically symmetric.

For the last example we assume for simplicity that $D = 2$, $d = 1$, $K = 2$ and underlying uniform distributions (of outliers and along the two underlying lines) restricted to the unit disk. We further assume that the two lines have angles ϵ and $-\epsilon$ w.r.t. the x-axis. By choosing ϵ sufficiently small the x-axis and y-axis provide a smaller value for the energy (1.2) than the underlying lines. We note that in this case (2.5) (or its variant above) does not hold (due to the small size of $\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_i^*, \mathrm{L}_j^*)$).

5.5 Preference for $p = 1$ over $p < 1$

Our theory explains why l_p minimization needs to be performed with $p \leq 1$ in order to recover multiple subspaces. We would like to mention here a different phase transition, which follows from Proposition 3.1.1 and shows why l_1 minimization is preferable to l_p minimization with $0 < p < 1$. We note that Proposition 3.1.1 implies that if the data set $\mathcal{X} \subset \mathbb{R}^D$ and the subspaces $\{\mathrm{L}_i^*\}_{i=1}^K \subset \mathrm{G}(D,d)$ satisfy the condition: $\mathrm{Sp}(\mathcal{X} \cap \mathrm{L}_i^*) = \mathrm{L}_i^*$ for each $1 \leq i \leq K$, then any L_i for $1 \leq i \leq K$ is a local minimum of the energy (1.1), and the set $\{\mathrm{L}_i^*\}_{i=1}^K$ is a local minimum of the energy (1.2). Furthermore, many subspaces satisfy this condition (in particular, w.o.p. d-subspaces spanned by randomly sampled d vectors). Therefore the minimization with $p < 1$ in (1.1) and (1.2) will often lead to plenty of local minima and this is often not the case with $p = 1$.

5.6 The Case of Affine Subspaces

Our analysis was restricted to linear subspaces, though it can be formally extended to affine subspaces intersecting a fixed ball. Indeed, we can consider the affine Grassmannian [42, 68], which distinguishes between subspaces according to both their offsets with respect to the origin (i.e., distances to closest linear subspaces of the same dimension) and their orientations (based on principal angles of the shifted linear subspaces). The assumption above on the affine subspaces (i.e., their offsets are less than the radius of the intersecting ball) restricts them to be in a compact subspace of the affine Grassmannian as necessary to our analysis. We can also generalize (A.3) (with a different function ψ_{μ_1}) and the estimates on δ_0 and κ_0 in Section 3.8.5 to the case of affine subspaces. We remark though that it is not obvious whether the metric on the affine Grassmannian is relevant for our applications, since it mixes two different quantities of different units (i.e., offset values and orientations) so that one can arbitrarily weigh their contributions. We remark that 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.

Moreover, the minimization of the energy (1.2) over affine subspaces seems to result in more local minima than in the linear case, which can partially explain why numerical heuristics for minimizing (1.2) do not perform well with affine subspaces as they do with linear ones. We are interested in further explanation of this phenomenon.

5.7 The Case of Mixed Dimensions

It will be interesting to try to extend our analysis to linear subspaces of mixed dimensions $d_1, \ldots, d_K$, known in advance. We believe that it is possible in this case to prove that if the percentage of outlier is sufficiently small, then the set of underlying subspaces coincides with the set minimizing the energy (1.2) whenever $0 < p \leq 1$. The idea of the proof seems to be closer to the current proof while using the same distance for subspaces of the same dimension and defining the distance $\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1, \mathrm{L}_2)$ between linear subspaces L_1 and L_2 of different dimensions (with some abuse of notation) as follows: If $\dim(\mathrm{L}_1) < \dim(\mathrm{L}_2)$, then $\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1, \mathrm{L}_2) = \min_{\substack{\mathrm{L} \in \mathrm{L}_2 \\ \dim(\mathrm{L}) = \dim(\mathrm{L}_1)}} \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1, \mathrm{L})$. We remark that the expected bound on the percentage of outliers is $\min_{1 \leq i,j \leq K} \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_i^*, \mathrm{L}_j^*)$.

5.8 Further Performance Guarantees for l_p-based HLM Algorithms

Our theory studies the recovery of the underlying subspaces by the l_p energy of (1.2). It will be interesting (though even more difficult) to extend such results to heuristics (like the K-subspaces) which try to minimize this energy in practice.

5.9 Asymptotic Rates of Convergence and Sample Complexity

In 3.7 we demonstrated simple instances when one cannot asymptotically recover by l_p minimization the underlying subspaces, when noise around those subspaces is present. One may still inquire about the existence of asymptotic limit different than the underlying subspaces and quantify the rate of convergence (depending on the mixture model parameters) to that limit. That is, assume that $\{\hat{\mathrm{L}}_1, \hat{\mathrm{L}}_2\}$ is the minimizer of $E_\mu\big(l_p(\mathbf{x}, \mathrm{L}_1, \mathrm{L}_2)\big)$ and $\{\hat{\mathrm{L}}_1^N, \hat{\mathrm{L}}_2^N\}$ is the minimizer of $E_{\mu_N}\big(l_p(\mathbf{x}, \mathrm{L}_1, \mathrm{L}_2)\big)$, where μ_N is an empirical distribution of sampling N points from distribution μ. We first ask whether $\mathrm{dist}\left(\{\hat{\mathrm{L}}_1, \hat{\mathrm{L}}_2\}, \{\hat{\mathrm{L}}_1^N, \hat{\mathrm{L}}_2^N\}\right) \to 0$ as $N \to \infty$. If true, then we ask about the asymptotic rates of convergence. This will then allow a definition of a sample complexity for multiple subspaces as the number $N = N(\epsilon, D, d, K)$ of samples that are required to achieve a prediction error within ϵ of the optimal one for K d-subspaces in $\mathbb{R}^D$.

References

[1] Nanda Kambhatla and Todd K. Leen. Fast non-linear dimension reduction. In *Advances in Neural Information Processing Systems 6*, pages 152–159. Morgan Kaufmann, 1994.

[2] J. Costeira and T. Kanade. A multibody factorization method for independently moving objects. *International Journal of Computer Vision*, 29(3):159–179, 1998.

[3] P. H. S. Torr. Geometric motion segmentation and model selection. *Phil. Trans. Royal Society of London A*, 356:1321–1340, 1998.

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

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

[6] P. Tseng. Nearest q-flat to m points. *Journal of Optimization Theory and Applications*, 105(1):249–252, April 2000.

[7] K. Kanatani. Motion segmentation by subspace separation and model selection. In *Proc. of 8th ICCV*, volume 3, pages 586–591. Vancouver, Canada, 2001.

[8] K. Kanatani. Evaluation and selection of models for motion segmentation. In *7th ECCV*, volume 3, pages 335–349, May 2002.

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

[10] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12), 2005.

[11] J. Yan and M. Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and nondegenerate. In *ECCV*, volume 4, pages 94–106, 2006.

[12] 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.

[13] Y. Ma, A. Y. Yang, H. Derksen, and R. Fossum. Estimation of subspace arrangements with applications in modeling and segmenting mixed data. *SIAM Review*, 50(3):413–458, 2008.

[14] Y. Ma, H. Derksen, W. Hong, and J. Wright. Segmentation of multivariate mixed data via lossy coding and compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1546–1562, September 2007.

[15] G. Chen and G. Lerman. Spectral curvature clustering (SCC). *Int. J. Comput. Vision*, 81(3):317–330, 2009.

[16] Akram Aldroubi, Carlos Cabrelli, and Ursula Molter. Optimal non-linear models for sparsity and sampling. *Journal of Fourier Analysis and Applications*, 14(5-6):793–812, December 2008.

[17] 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.

[18] T. Zhang, A. Szlam, Y. Wang, and G. Lerman. Randomized hybrid linear modeling by local best-fit flats. pages 1927 –1934, jun. 2010.

[19] 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.

[20] 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.

[21] Philip H. S. Torr and Andrew 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.

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

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

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

[25] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions (Wiley Series in Probability and Statistics)*. Wiley-Interscience, New York, revised edition, April 2005.

[26] Peter J. Rousseeuw and Annick 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.

[27] Ricardo A. Maronna, R. Douglas Martin, and Victor J. Yohai. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 2006. Theory and methods.

[28] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61 (electronic), 1998.

[29] 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.

[30] David 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.

[31] David 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.

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

[33] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? Submitted, Dec. 2009, arXiv:0912.3599.

[34] 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.

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

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

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

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

[39] Chris Ding, Ding Zhou, Xiaofeng He, and Hongyuan 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.

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

[41] G. Lerman and T. Zhang. l_p-Recovery of the most significant subspace among multiple subspaces with outliers. Submitted December 2010. Available at http://arxiv.org/abs/1012.4116.

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

[43] Nuno Pinho da Silva and João Paulo Costeira. Subspace segmentation with outliers: A grassmannian approach to the maximum consensus subspace. In *CVPR*. IEEE Computer Society, 2008.

[44] T. W. Anderson and Theodore W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley-Interscience, 2 edition, September 1984.

[45] J. Shawe-taylor, C. Williams, N. Cristianini, and J. Kandola. On the eigenspectrum of the Gram matrix and the generalisation error of kernel PCA. *IEEE Transactions on Information Theory*, 51(1):2510–2522, 2005.

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

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

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

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

[50] Stanislaw 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.

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

[52] David Pollard. A central limit theorem for k-means clustering. *Annals of Probability*, 10:919–926, 1982.

[53] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pages 849–856, 2001.

[54] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006.

[55] P. W. Jones. Rectifiable sets and the traveling salesman problem. *Invent Math*, 102(1):1–15, 1990.

[56] G. Lerman. Quantifying curvelike structures of measures by using L_2 Jones quantities. *Comm. Pure Appl. Math.*, 56(9):1294–1365, 2003.

[57] Y.-M. Jung, A.V. Little, M. Maggioni, and L. Rosasco. Multiscale estimation of intrinsic dimensionality of noisy data sets. preprint.

[58] G. Lerman and T. Zhang. Probabilistic recovery of multiple subspaces in point clouds by geometric l_p minimization. Submitted April 2010. Available at http://arxiv.org/abs/1002.1994.

[59] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman and Hall, 2nd edition, 2001.

[60] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[61] R. Vidal, R. Tron, and R. Hartley. Multiframe motion segmentation with missing data using powerfactorization and gpca. *Int. J. Comput. Vision*, 79(1):85–105, 2008.

[62] A. Goh and R. Vidal. Segmenting motions of different types by unsupervised manifold clustering. In *CVPR 07*, 2007.

[63] Y. Sugaya and K. Kanatani. Multi-stage unsupervised learning for multi-body motion segmentation. *IEICE Transactions on Information and Systems*, E87-D(7):1935–1942, 2004.

[64] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 09)*, pages 2790 – 2797, 2009.

[65] R. Tron and R. Vidal. A benchmark for the comparison of 3-d motion segmentation algorithms. In *CVPR*, 2007.

[66] G. Chen and G. Lerman. Motion segmentation by SCC on the Hopkins 155 database. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on Computer Vision*, pages 759–764, Kyoto, Japan.

[67] F. Lauer and C. Schnorr. Spectral clustering of linear subspaces for motion segmentation. pages 678 –685, sep. 2009.

[68] D. A. Klain and G.-C. Rota. *Introduction to Geometric Probability*. Cambridge University Press, 1997.

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 $\mathbb{R}^D$:

$$\psi_\mu(t) < \frac{2d}{\pi}\, t. \tag{A.1}$$

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

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

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

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

A.2 Proof of Lemma 3.2.1

We will use the following inequality for any $1 \leq j \leq K$, which is proved in Section A.2:

$$\mu_1\left(\mathbf{x} \in \mathrm{B}(\mathbf{0},1) \cap \mathrm{L}_1^* : \mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}}_j) < \beta \, \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1^*, \hat{\mathrm{L}}_j)\right) \leq \psi_{\mu_1}\left(\frac{\pi\sqrt{d}}{2}\beta\right) \quad \forall \beta > 0. \tag{A.3}$$

We denote $\beta_1 = \frac{2}{\pi\sqrt{d}}\psi_{\mu_1}^{-1}\left(\frac{1+(2K-1)\mu_1(\{\mathbf{0}\})}{2K}\right)$ (the existence of $\psi_{\mu_1}^{-1}\left(\frac{1+(2K-1)\mu_1(\{\mathbf{0}\})}{2K}\right)$ will be proved later) and combine (A.3) with the fact that $\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1^*, \hat{\mathrm{L}}_j) \geq \epsilon$ for any $1 \leq j \leq K$ to obtain that

$$\begin{aligned}
&\mu_1\left(\mathbf{x} \in \mathrm{B}(\mathbf{0},1) \cap \mathrm{L}_1^* \setminus \{\mathbf{0}\} : \mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}}_1) < \beta_1 \, \epsilon\right) \\
=&\mu_1\left(\mathbf{x} \in \mathrm{B}(\mathbf{0},1) \cap \mathrm{L}_1^* \setminus \{\mathbf{0}\} : \mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}}_1) < \beta_1 \, \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1^*, \hat{\mathrm{L}}_1)\right) \\
\leq&\frac{1+(2K-1)\mu_1(\{\mathbf{0}\})}{2K} - \mu(\{\mathbf{0}\}) = \frac{1-\mu_1(\{\mathbf{0}\})}{2K}.
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\mu_1\left(\mathbf{x} \in \mathrm{B}(\mathbf{0},1) \cap \mathrm{L}_1^* : \mathrm{dist}(\mathbf{x}, \cup_{j=1}^K \hat{\mathrm{L}}_1) \geq \beta_1 \epsilon\right) \geq 1 - \mu(\{\mathbf{0}\}) \\
&- \sum_{i=1}^K \mu_1\left(\mathbf{x} \in \mathrm{B}(\mathbf{0},1) \cap \mathrm{L}_1^* \setminus \{\mathbf{0}\} : \mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}}_i) < \beta_1 \, \epsilon\right) \geq (1 - \mu_1(\{\mathbf{0}\}))/2,
\end{aligned}$$

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

$$E_{\mu_1}\left(e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1)\right) \geq \beta_1^p \epsilon^p/2 = \frac{(1-\mu_1(\{\mathbf{0}\}))2^{p-1}\epsilon^p}{(\pi\sqrt{d})^p \psi_{\mu_1}^{-1}\left(\frac{1+(2K-1)\mu_1(\{\mathbf{0}\})}{2K}\right)^p} = \tau_0 \, \epsilon^p.$$

The existence of $\psi_{\mu_1}^{-1}\left(\frac{1+(2K-1)\mu_1(\{\mathbf{0}\})}{2K}\right)$ will follow from the following observation:

$$\begin{aligned}
&\mu_1(\mathrm{L}) = 0 \ \text{ for any affine subspace } \mathrm{L} \subset \mathrm{L}_1, \\
&\mu_1(\mathrm{L}) = \mu_1(\{\mathbf{0}\}) \text{ for any linear subspace } \mathrm{L} \subsetneq \mathrm{L}_1,
\end{aligned} \tag{A.4}$$

We prove it as follows: Assume that d_0 is the smallest dimension for which there exists a subspace L_0 such that (A.4) 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 (A.4), 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 $\psi_{\mu_1}^{-1}\left(\frac{1+(2K-1)\mu_1(\{\mathbf{0}\})}{2K}\right)$ is concluded.

Proof of (A.3)

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), the volume of the d-dimensional unit ball by v_d and

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

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

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

We prove (A.3), 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 3.2.2

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 3.2.3

We assume WLOG that $i=1$ in (3.6). We thus need to prove that for all $\hat{\mathrm{L}} \in \mathrm{G}(D,d)$:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}_1\in\mu_1}(\mathrm{dist}(\mathbf{x}_1,\hat{\mathrm{L}})^p) + \mathbb{E}_{\mathbf{x}_2\in\mu_2}(\mathrm{dist}(\mathbf{x}_2,\hat{\mathrm{L}})^p) \\
\geq &\mathbb{E}_{\mathbf{x}_1\in\mu_1}(\mathrm{dist}(\mathbf{x}_1,\mathrm{L}_1)^p) + \mathbb{E}_{\mathbf{x}_2\in\mu_2}(\mathrm{dist}(\mathbf{x}_2,\mathrm{L}_1)^p).
\end{aligned} \tag{A.5}$$

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 (A.5). First of all, we only need to consider subspaces

$$\hat{\mathrm{L}} \in \mathrm{L}_1 + \mathrm{L}_2\,. \tag{A.6}$$

Indeed, the LHS of (A.5) 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

$$\mathrm{Sp}(\hat{\mathbf{v}}_i, \mathbf{v}_i) \nsubseteq \hat{\mathrm{L}} \;\text{ for all } 1 \leq i \leq k. \tag{A.7}$$

Indeed, WLOG let $i=1$ and suppose on the contrary to (A.7) that $\hat{\mathbf{v}}_1, \mathbf{v}_1 \in \hat{\mathrm{L}}$. Since $\hat{\mathrm{L}}$ is d-dimensional, there exists $2 \leq j \leq d$ (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 $\hat{\mathbf{x}} = \sum_{i=1}^d a_i\hat{\mathbf{v}}_i \in \mathrm{L}_2$:

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

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

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

and by direct integration we have that

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

We can thus replace the subspace $\hat{\mathrm{L}}$ with the subspace $\tilde{\mathrm{L}}$, which satisfies (A.7) (for $i = 1$, but can similarly be changed for all $1 < i \leq K$).

It follows from (A.6) and (A.7) 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 $\hat{\mathbf{x}} = \sum_{i=1}^d a_i \hat{\mathbf{v}}_i \in \mathrm{L}_2$:

$$\mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}}) = \sqrt{\sum_{i=1}^d \sin^2\theta_i^* a_i^2} \text{ and } \mathrm{dist}(\hat{\mathbf{x}}, \hat{\mathrm{L}}) = \sqrt{\sum_{i=1}^d \sin^2(\theta_i - \theta_i^*) a_i^2} \tag{A.8}$$

and

$$\mathrm{dist}(\mathbf{x}, \mathrm{L}_1) = 0 \text{ and } \mathrm{dist}(\hat{\mathbf{x}}, \mathrm{L}_1) = \sqrt{\sum_{i=1}^d \sin^2\theta_i a_i^2}. \tag{A.9}$$

Combining (A.8), (A.9), the triangle inequality (for "sine vectors" in $\mathbb{R}^d$) and the subadditivity of the sine function, we conclude that

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

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

$$\mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}})^p + \mathrm{dist}(\hat{\mathbf{x}}, \hat{\mathrm{L}})^p \geq \mathrm{dist}(\hat{\mathbf{x}}, \mathrm{L}_1)^p = \mathrm{dist}(\hat{\mathbf{x}}, \mathrm{L}_1)^p + \mathrm{dist}(\mathbf{x}, \mathrm{L}_1)^p. \tag{A.10}$$

Integrating (A.10) w.r.t. the uniform distribution we conclude (A.5) and thus prove the lemma.

A.5 Proof of (3.5)

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

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

We further note that for any $0 < r \leq 1$, the set $\{\mathbf{x} \in \mathrm{B}(\mathbf{0}, 1) \cap \mathrm{L}_1 : \mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}}) = r\}$ 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:

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

A.6 Proof of (3.27)

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

A.7 Proof of Lemma 3.6.1

We define

$$M = \mathrm{argmax}_{1 \leq i \leq K} \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_i^*, \hat{\mathrm{L}}_{I(i)}).$$

Assume first that $(I(1), \cdots, I(K))$ is a permutation of $(1, \cdots, K)$, then I has an inverse function, I^{-1}. Using the definition of I we have

$$\min_{1 \leq j \leq K} \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_M^*, \hat{\mathrm{L}}_j) = \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_M^*, \hat{\mathrm{L}}_{I(M)}) \tag{A.11}$$

$$= \mathrm{dist}_{\mathrm{G^K}}((\mathrm{L}_1^*, \cdots, \mathrm{L}_K^*), (\hat{\mathrm{L}}_{I(1)}, \cdots, \hat{\mathrm{L}}_{I(K)})) = d_0.$$

Combining (A.11) with Lemma 3.2.1 we obtain that

$$\begin{aligned} & E_{\mu_M} e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K) - E_{\mu_M} e_{l_p}(\mathbf{x}, \mathrm{L}_1^*, \cdots, \mathrm{L}_K^*) \\ & = E_{\mu_M} e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K) > \tau_0\, d_0^p. \end{aligned} \tag{A.12}$$

For any $\mathbf{x} \in \mathcal{X}_0$, let $m(\mathbf{x}) = \arg\min_{1 \le i \le K} \mathrm{dist}(\mathbf{x}, \mathrm{L}_i^*)$, $\hat{m}(\mathbf{x}) = \arg\min_{1 \le i \le K} \mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}}_i)$ and note that

$$\begin{aligned} & e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K) - e_{l_p}(\mathbf{x}, \mathrm{L}_1^*, \cdots, \mathrm{L}_K^*) = \mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}}_{\hat{m}(\mathbf{x})})^p \\ & - \mathrm{dist}(\mathbf{x}, \mathrm{L}_{m(\mathbf{x})}^*)^p \ge \mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}}_{\hat{m}(\mathbf{x})})^p - \mathrm{dist}(\mathbf{x}, \mathrm{L}_{I^{-1}(\hat{m}(\mathbf{x}))}^*)^p \\ & \ge -\|\mathbf{x}\|^p \mathrm{dist}_{\mathrm{G}}(\hat{\mathrm{L}}_{\hat{m}(\mathbf{x})}, \mathrm{L}_{I^{-1}(\hat{m}(\mathbf{x}))}^*)^p \ge -\|\mathbf{x}\|^p d_0^p \ge -d_0^p, \end{aligned} \tag{A.13}$$

where the second inequality in (A.13) uses Lemma 3.2.2. Therefore,

$$E_{\mu_0} e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K) - E_{\mu_0} e_{l_p}(\mathbf{x}, \mathrm{L}_1^*, \cdots, \mathrm{L}_K^*) > -d_0^p. \tag{A.14}$$

At last, we observe that

$$\begin{aligned} & E_{\mu} e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K) - E_{\mu} e_{l_p}(\mathbf{x}, \mathrm{L}_1^*, \cdots, \mathrm{L}_K^*) \\ & \ge \alpha_M \left(E_{\mu_M} e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K) - E_{\mu_M} e_{l_p}(\mathbf{x}, \mathrm{L}_1^*, \cdots, \mathrm{L}_K^*) \right) \\ & + \alpha_0 \left(E_{\mu_0} e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K) - E_{\mu_0} e_{l_p}(\mathbf{x}, \mathrm{L}_1^*, \cdots, \mathrm{L}_K^*) \right). \end{aligned} \tag{A.15}$$

Combining (A.12), (A.14) and (A.15), the lemma is proved in this case.

Next, we assume that $I(1), \cdots, I(K)$ is not a permutation of $1, 2, \cdots, K$, where we use some of the notation introduced above. In this case, there exist $1 \le n_1, n_2 \le K$ such that $I(n_1) = I(n_2)$ and consequently

$$\begin{aligned} & 2 \min_{1 \le j \le K} \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_M^*, \hat{\mathrm{L}}_j) = 2\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_M^*, \hat{\mathrm{L}}_{I(M)}) \ge \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_{n_1}^*, \hat{\mathrm{L}}_{I(n_1)}) \\ & + \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_{n_2}^*, \hat{\mathrm{L}}_{I(n_2)}) \ge \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_{n_1}^*, \mathrm{L}_{n_2}^*) \ge \min_{1 \le i,j \le K} \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_i^*, \mathrm{L}_j^*). \end{aligned} \tag{A.16}$$

Combining (A.16) and Lemma 3.2.1 (applied with $\epsilon = \min_{1 \le i,j \le K} \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_i^*, \mathrm{L}_j^*)/2$), we obtain that

$$E_{\mu_M} e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K) - E_{\mu_M} e_{l_p}(\mathbf{x}, \mathrm{L}_1^*, \cdots, \mathrm{L}_K^*) \tag{A.17}$$

$$> \tau_0 \, (\min_{1 \le i,j \le K} \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_i^*, \mathrm{L}_j^*)/2)^p.$$

Using the above notation for $m(\mathbf{x})$ and $\hat{m}(\mathbf{x})$ we get that for any $\mathbf{x} \in \mathcal{X}_0$:

$$\begin{aligned} & e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K) - e_{l_p}(\mathbf{x}, \mathrm{L}_1^*, \cdots, \mathrm{L}_K^*) \\ & = \mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}}_{\hat{m}(\mathbf{x})}) - \mathrm{dist}(\mathbf{x}, \mathrm{L}_{m(\mathbf{x})}^*) \ge -1 \end{aligned} \tag{A.18}$$

and consequently

$$E_{\mu_0} e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1, \cdots, \hat{\mathrm{L}}_K) - E_{\mu_0} e_{l_p}(\mathbf{x}, \mathrm{L}_1^*, \cdots, \mathrm{L}_K^*) > -1. \tag{A.19}$$

The lemma is concluded by combing (A.15), (A.17) and (A.19).

A.8 Proof of Lemma 3.7.1

The proof is an immediate consequence of the following inequality involving an arbitrary $\tilde{\mathrm{L}}_1 \in \mathrm{G}(D, d)$:

$$\begin{aligned} & 0 \le E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x}, \tilde{\mathrm{L}}_1, \mathrm{L}_2^*, \cdots, \mathrm{L}_K^*)) - E_{\tilde{\mu}_\epsilon}(e_{l_p}(\mathbf{x}, \mathrm{L}_1^*, \cdots, \mathrm{L}_K^*)) \\ & \le E_{\tilde{\mu}_\epsilon}(I(\mathbf{x} \in \mathrm{Y}_1) e_{l_p}(\mathbf{x}, \tilde{\mathrm{L}}_1)) + \sum_{2 \le i \le K} E_{\tilde{\mu}_\epsilon}(I(\mathbf{x} \in \mathrm{Y}_i) e_{l_p}(\mathbf{x}, \mathrm{L}_i^*)) \\ & - \sum_{1 \le i \le K} E_{\tilde{\mu}_\epsilon}(I(\mathbf{x} \in \mathrm{Y}_i) e_{l_p}(\mathbf{x}, \mathrm{L}_i^*)) \\ & = E_{\tilde{\mu}_\epsilon}(I(\mathbf{x} \in \mathrm{Y}_1) e_{l_p}(\mathbf{x}, \tilde{\mathrm{L}}_1)) - E_{\tilde{\mu}_\epsilon}(I(\mathbf{x} \in \mathrm{Y}_1) e_{l_p}(\mathbf{x}, \mathrm{L}_1^*)). \end{aligned}$$

A.9 Proof of Lemma 3.8.2: Geometric Sensitivity

We will first show that there exists $\mathbf{x}_0 \in \mathrm{B}(\mathbf{0}, 1)$ such that

$$\mathrm{dist}(\mathbf{x}_0, \mathrm{L}_1^*) = \mathrm{dist}(\mathbf{x}_0, \mathrm{L}_2^*) < \min_{3 \le i \le K} \mathrm{dist}(\mathbf{x}_0, \mathrm{L}_i^*). \tag{A.20}$$

We verify (A.20) in two cases: $d^* = d$ and $d^* = D - d$. We will then prove that (A.20) implies (3.99). Throughout the proof we denote the principal vectors of L_2^* and L_1^* by $\{\hat{\mathbf{v}}_i\}_{i=1}^{d^*}$ and $\{\mathbf{v}_i\}_{i=1}^{d^*}$ respectively.

Part I: Proof of (A.20) when $d^* = d$

We define

$$\mathbf{x}_0 = (\hat{\mathbf{v}}_{d^*} + \mathbf{v}_{d^*})/\|\hat{\mathbf{v}}_{d^*} + \mathbf{v}_{d^*}\|$$

and arbitrarily fix $i_0 > 3$ and $\mathbf{v}_0 \in \mathrm{L}^*_{i_0}$. We will show that

$$\mathrm{ang}(\mathbf{x}_0, \mathbf{v}_0) > \theta_{d^*}(\mathrm{L}^*_2, \mathrm{L}^*_1)/2 \tag{A.21}$$

and consequently conclude (A.20) as follows:

$$\mathrm{dist}(\mathbf{x}_0, \mathrm{L}^*_{i_0}) \geq \mathrm{ang}(\mathbf{x}_0, \mathbf{v}_0) > \theta_{d^*}(\mathrm{L}^*_2, \mathrm{L}^*_1)/2 = \mathrm{dist}(\mathbf{x}_0, \mathrm{L}^*_1) = \mathrm{dist}(\mathbf{x}_0, \mathrm{L}^*_2).$$

We can easily verify a weaker version of (A.21) where the inequality is not necessarily strict. Indeed, using elementary geometric estimates and the fact that the intersections of the d-subspaces $\{\mathrm{L}^*_i\}_{i=1}^K$ are empty (which follows from (3.97)), we obtain that

$$\begin{aligned}\mathrm{ang}(\mathbf{x}_0, \mathbf{v}_0) &\geq \mathrm{ang}(\mathbf{v}_{d^*}, \mathbf{v}_0) - \mathrm{ang}(\mathbf{v}_{d^*}, \mathbf{x}_0) \geq \theta_{d^*}(\mathrm{L}^*_{i_0}, \mathrm{L}^*_1) - \theta_{d^*}(\mathrm{L}^*_2, \mathrm{L}^*_1)/2 \\ &\geq \theta_{d^*}(\mathrm{L}^*_2, \mathrm{L}^*_1) - \theta_{d^*}(\mathrm{L}^*_2, \mathrm{L}^*_1)/2 = \theta_{d^*}(\mathrm{L}^*_2, \mathrm{L}^*_1)/2. \end{aligned} \tag{A.22}$$

At last, we show that (A.22) cannot be an equality. Indeed, if the first inequality in (A.22) is an equality, then $\mathbf{v}_0$, $\mathbf{v}_{d^*}$ and $\mathbf{x}_0$ are on a geodesic line within the sphere S^{D-1}. Combining this with the assumption that all other inequalities in (A.22) are equalities, we obtain that $\mathrm{ang}(\mathbf{x}_0, \mathbf{v}_0) = \theta_{d^*}(\mathrm{L}^*_2, \mathrm{L}^*_1)/2 = \mathrm{ang}(\mathbf{x}_0, \mathbf{v}_{d^*}) = \mathrm{ang}(\mathbf{x}_0, \hat{\mathbf{v}}_{d^*})$. This implies that either $\mathbf{v}_0 = \hat{\mathbf{v}}_{d^*}$ or $\mathbf{v}_0 = \mathbf{v}_{d^*}$, which contradicts (3.97).

Part II: Proof of (A.20) when $d^* = D - d$

It follows from basic dimension equalities of subspaces and (3.97) that for all $2 \leq i \leq K$: $\dim(\mathrm{L}^*_1 \cup \mathrm{L}^*_i) = D$ and $\dim(\mathrm{L}^*_1 \cap \mathrm{L}^*_i) = 2d - D$. We denote by K_0 the integer in $\{0, \ldots, K\}$ such that for any $3 \leq i \leq K_0$: $\mathrm{L}^*_1 \cap \mathrm{L}^*_i = \mathrm{L}^*_1 \cap \mathrm{L}^*_2$ and for any $i > K_0$: $\mathrm{L}^*_1 \cap \mathrm{L}^*_i \neq \mathrm{L}^*_1 \cap \mathrm{L}^*_2$ (the existence of K_0 may require reordering of the indices of the subspaces $\{\mathrm{L}^*_i\}_{i=3}^K$). In order to define $\mathbf{x}_0$ in the current case, we let $\mathbf{x}_1 = (\hat{\mathbf{v}}_{d^*} + \mathbf{v}_{d^*})/\|\hat{\mathbf{v}}_{d^*} + \mathbf{v}_{d^*}\|$, $\mathbf{x}_2$ be an arbitrarily fixed unit vector in $\mathrm{L}^*_1 \cap (\mathrm{L}^*_2 \setminus \cup_{K_0 < i \leq K} \mathrm{L}^*_i)$, $\epsilon_0 = \mathrm{dist}(\mathbf{x}_2, \cup_{K_0 < i \leq K} \mathrm{L}^*_i)$ and

$$\mathbf{x}_0 = \mathbf{x}_2/2 + \epsilon_0\, \mathbf{x}_1/5.$$

We first claim that

$$\operatorname{dist}(\mathbf{x}_0, \mathrm{L}_1^*) = \operatorname{dist}(\mathbf{x}_0, \mathrm{L}_2^*) < \min_{3 \leq j \leq K_0} \operatorname{dist}(\mathbf{x}_0, \mathrm{L}_j^*). \tag{A.23}$$

Indeed, we can remove $\mathrm{L}_1^* \cap \mathrm{L}_2^*$ from the subspaces $\{\mathrm{L}_i^*\}_{i=1}^{K_0}$ and obtain subspaces of dimension $D - d$ intersecting each other at the origin. We can then rewrite (A.23) by replacing $\{\mathrm{L}_i^*\}_{i=1}^{K_0}$ with their reduced version and $\mathbf{x}_0$ with $\mathbf{x}_1$. The argument of Section A.9 thus proves this equation.

We conclude (A.20) by combining (A.23) with the following observation:

$$\begin{aligned}\operatorname{dist}(\mathbf{x}_0, \mathrm{L}_1^*) = \epsilon_0 \operatorname{dist}(\mathbf{x}_1, \mathrm{L}_1^*)/5 \leq \epsilon_0/5 < \operatorname{dist}(\mathbf{x}_2/2, \cup_{K_0 < j \leq K} \mathrm{L}_j^*) - \epsilon_0/5 \\ \leq \operatorname{dist}(\mathbf{x}_2/2 + \epsilon_0\, \mathbf{x}_1/5, \cup_{K_0 < j \leq K} \mathrm{L}_j^*) = \min_{K_0 < i \leq K} \operatorname{dist}(\mathbf{x}_0, \mathrm{L}_i^*).\end{aligned} \tag{A.24}$$

Part III: Deriving (3.99) from (A.20) in a Simple Case

We note that (A.20) implies that

$$\mathbf{x}_0 \in (\mathrm{Y}_1 \cup \mathrm{Y}_2 \cup (\bar{\mathrm{Y}}_1 \cap \bar{\mathrm{Y}}_2)) \cap (\hat{\mathrm{Y}}_1 \cup \hat{\mathrm{Y}}_2 \cup (\bar{\hat{\mathrm{Y}}}_1 \cap \bar{\hat{\mathrm{Y}}}_2)) \tag{A.25}$$

and consequently

$$\mathrm{B}(\mathbf{x}_0, \epsilon) \subset (\mathrm{Y}_1 \cup \mathrm{Y}_2 \cup (\bar{\mathrm{Y}}_1 \cap \bar{\mathrm{Y}}_2)) \cap (\hat{\mathrm{Y}}_1 \cup \hat{\mathrm{Y}}_2 \cup (\bar{\hat{\mathrm{Y}}}_1 \cap \bar{\hat{\mathrm{Y}}}_2)). \tag{A.26}$$

We will deduce here (3.99) from (A.26) in the simpler case: $\bar{\hat{\mathrm{Y}}}_1 \cap \bar{\hat{\mathrm{Y}}}_2 \cap \mathrm{B}(\mathbf{x}_0, \epsilon) \neq \bar{\mathrm{Y}}_1 \cap \bar{\mathrm{Y}}_2 \cap \mathrm{B}(\mathbf{x}_0, \epsilon)$.

Using (A.26) and the fact that $\mathcal{L}_D(\bar{\mathrm{Y}}_1 \cap \bar{\mathrm{Y}}_2) = 0$, we may choose $\mathbf{y} \in (\bar{\hat{\mathrm{Y}}}_1 \cap \bar{\hat{\mathrm{Y}}}_2 \cap \mathrm{B}(\mathbf{x}_0, \epsilon)) \cap (\mathrm{Y}_1 \cup \mathrm{Y}_2)$; WLOG we assume instead of the later condition that $\mathbf{y} \in (\bar{\hat{\mathrm{Y}}}_1 \cap \bar{\hat{\mathrm{Y}}}_2 \cap \mathrm{B}(\mathbf{x}_0, \epsilon)) \cap \mathrm{Y}_1$. By slightly perturbing $\mathbf{y}$ we can choose another point $\mathbf{y}_0$ such that $\mathbf{y}_0 \in \hat{\mathrm{Y}}_2$ and $\mathbf{y}_0 \in \mathrm{Y}_1 \setminus \hat{\mathrm{Y}}_1$. It follows from the continuity of the distance function that there exists a small $\eta > 0$ such that $(\hat{\mathrm{Y}}_1 \setminus \mathrm{Y}_1) \cup (\mathrm{Y}_1 \setminus \hat{\mathrm{Y}}_1) \supseteq \mathrm{Y}_1 \setminus \hat{\mathrm{Y}}_1 \supset \mathrm{B}(\mathbf{y}_0, \eta)$, which proves (3.99).

Part IV: Deriving (3.99) from (A.20) in the Complementary Case

At last, we assume that $\bar{\hat{\mathrm{Y}}}_1 \cap \bar{\hat{\mathrm{Y}}}_2 \cap \mathrm{B}(\mathbf{x}_0, \epsilon) = \bar{\mathrm{Y}}_1 \cap \bar{\mathrm{Y}}_2 \cap \mathrm{B}(\mathbf{x}_0, \epsilon)$. We show here that it leads to the contradiction: $\hat{\mathrm{L}}_2 = \mathrm{L}_2^*$.

We note that the sets of solutions in $\mathrm{B}(\mathbf{x}_0, \epsilon)$ of the equations $\mathbf{x}^T(P_{\mathrm{L}_1^*} - P_{\mathrm{L}_2^*})\mathbf{x} = 0$ and $\mathbf{x}^T(P_{\mathrm{L}_1^*} - P_{\hat{\mathrm{L}}_2})\mathbf{x} = 0$ are $\bar{\hat{\mathrm{Y}}}_1 \cap \bar{\hat{\mathrm{Y}}}_2 \cap \mathrm{B}(\mathbf{x}_0, \epsilon)$ and $\bar{\mathrm{Y}}_1 \cap \bar{\mathrm{Y}}_2 \cap \mathrm{B}(\mathbf{x}_0, \epsilon)$ respectively. In view of (A.26), these solution sets coincide. They are $(D-1)$-manifolds and thus their $(D-1)$-dimensional tangent spaces at $\mathbf{x}_0$, i.e., $\mathbf{x}_0^T(P_{\mathrm{L}_1^*} - P_{\mathrm{L}_2^*}) = \mathbf{0}$ and $\mathbf{x}_0^T(P_{\mathrm{L}_1^*} - P_{\hat{\mathrm{L}}_2}) = \mathbf{0}$, also coincide. Consequently we have that $\mathbf{x}_0^T(P_{\mathrm{L}_1^*} - P_{\mathrm{L}_2^*}) = t_0\,\mathbf{x}_0^T(P_{\mathrm{L}_1^*} - P_{\hat{\mathrm{L}}_2})$ for some $t_0 \neq 0$. Similarly, for any $\mathbf{x}_1 \in \bar{\hat{\mathrm{Y}}}_1 \cap \bar{\hat{\mathrm{Y}}}_2 \cap \mathrm{B}(\mathbf{x}_0, \epsilon)$, we have $\mathbf{x}_1^T(P_{\mathrm{L}_1^*} - P_{\mathrm{L}_2^*}) = t_1\,\mathbf{x}_1^T(P_{\mathrm{L}_1^*} - P_{\hat{\mathrm{L}}_2})$ for some $t_1 \neq 0$. We note that $t_1 = t_0$ by the following argument: $t_1\,\mathbf{x}_1^T(P_{\mathrm{L}_1^*} - P_{\hat{\mathrm{L}}_2})\mathbf{x}_0 = \mathbf{x}_1^T(P_{\mathrm{L}_1^*} - P_{\mathrm{L}_2^*})\mathbf{x}_0 = t_0\,\mathbf{x}_1^T(P_{\mathrm{L}_1^*} - P_{\hat{\mathrm{L}}_2})\mathbf{x}_0$. Therefore, there exists $t \neq 0$ such that for any $\mathbf{x}_1 \in \bar{\hat{\mathrm{Y}}}_1 \cap \bar{\hat{\mathrm{Y}}}_2 \cap \mathrm{B}(\mathbf{x}_0, \epsilon)$:

$$\mathbf{x}_1^T(P_{\mathrm{L}_1^*} - P_{\mathrm{L}_2^*}) = t\,\mathbf{x}_1^T(P_{\mathrm{L}_1^*} - P_{\hat{\mathrm{L}}_2}). \tag{A.27}$$

Since the tangent space of $\bar{\hat{\mathrm{Y}}}_1 \cap \bar{\hat{\mathrm{Y}}}_2 \cap \mathrm{B}(\mathbf{x}_0, \epsilon)$ (or equivalently $\mathbf{x}^T(P_{\mathrm{L}_1^*} - P_{\hat{\mathrm{L}}_2})\mathbf{x} = 0$) at $\mathbf{x}_0$ has dimension $D-1$, the subspace $\mathrm{L}_0^* = \mathrm{Sp}(\bar{\hat{\mathrm{Y}}}_1 \cap \bar{\hat{\mathrm{Y}}}_2 \cap \mathrm{B}(\mathbf{x_0}, \epsilon))$ has dimension at least $D-1$. In view of (A.27), L_0^* satisfies

$$P_{\mathrm{L}_0^*}(P_{\mathrm{L}_1^*} - P_{\mathrm{L}_2^*}) = t\,P_{\mathrm{L}_0^*}(P_{\mathrm{L}_1^*} - P_{\hat{\mathrm{L}}_2}). \tag{A.28}$$

Due to the symmetry of $(P_{\mathrm{L}_1^*} - P_{\hat{\mathrm{L}}_2})$ and $(P_{\mathrm{L}_1^*} - P_{\mathrm{L}_2^*})$, we have the following equivalent formulation of (A.28):

$$(P_{\mathrm{L}_1^*} - P_{\mathrm{L}_2^*})P_{\mathrm{L}_0^*} = (P_{\mathrm{L}_1^*} - P_{\hat{\mathrm{L}}_2})P_{\mathrm{L}_0^*}. \tag{A.29}$$

Furthermore, using the fact that $(P_{\mathrm{L}_1^*} - P_{\hat{\mathrm{L}}_2})$ and $(P_{\mathrm{L}_1^*} - P_{\mathrm{L}_2^*})$ have trace 0, we obtain that

$$\begin{aligned}\mathrm{tr}(P_{\mathrm{L}_1^{*\perp}}(P_{\mathrm{L}_1^*} - P_{\mathrm{L}_2^*})P_{\mathrm{L}_0^{*\perp}}) =& -\mathrm{tr}(P_{\mathrm{L}_0^*}(P_{\mathrm{L}_1^*} - P_{\mathrm{L}_2^*})P_{\mathrm{L}_0^*}) = -t\cdot\mathrm{tr}(P_{\mathrm{L}_0^*}(P_{\mathrm{L}_1^*} - P_{\hat{\mathrm{L}}_2})P_{\mathrm{L}_0^*})\\ =& t\cdot\mathrm{tr}(P_{\mathrm{L}_0^{*\perp}}(P_{\mathrm{L}_1^*} - P_{\hat{\mathrm{L}}_2})P_{\mathrm{L}_0^{*\perp}}).\end{aligned} \tag{A.30}$$

Since $P_{\mathrm{L}_0^{*\perp}}$ is at most one-dimensional, (A.30) can be rewritten as

$$P_{\mathrm{L}_0^{*\perp}}(P_{\mathrm{L}_1^*} - P_{\mathrm{L}_2^*})P_{\mathrm{L}_0^{*\perp}} = t\cdot(P_{\mathrm{L}_0^{*\perp}}(P_{\mathrm{L}_1^*} - P_{\hat{\mathrm{L}}_2})P_{\mathrm{L}_0^{*\perp}}). \tag{A.31}$$

Combining (A.28), (A.29) and (A.31), we obtain that $(P_{\mathrm{L}_1^*} - P_{\hat{\mathrm{L}}_2}) = t\,(P_{\mathrm{L}_1^*} - P_{\mathrm{L}_2^*})$, equivalently,

$$P_{\hat{\mathrm{L}}_2} = (1-t)\,P_{\mathrm{L}_1^*} + t\,P_{\mathrm{L}_2^*}. \tag{A.32}$$

We conclude the desired contradiction in two different cases. Assume first that $t < 1$ and let $\mathbf{v}_0$ be an arbitrary unit vector in L_2^*. We note that $\mathbf{v}_0^T P_{\hat{\mathrm{L}}_2}\mathbf{v}_0 = 1$ as well as $(1-t)\,\mathbf{v}_0^T P_{\mathrm{L}_1^*}\mathbf{v}_0 =$

$1 - t\,\mathbf{v}_0^T P_{\mathrm{L}_2^*}\mathbf{v}_0 \geq 1 - t$. Consequently, $\mathbf{v}_0^T P_{\mathrm{L}_1^*}\mathbf{v}_0 = 1$, i.e., $\mathbf{v}_0 \in \mathrm{L}_1^*$ and thus we obtain the following contradiction with (3.97): $\mathrm{L}_1^* = \hat{\mathrm{L}}_2$ (in view of (A.32) this is equivalent with $\hat{\mathrm{L}}_2 = \mathrm{L}_2^*$). Next, assume that $t \geq 1$ and as before $\mathbf{v}_0$ is an arbitrary unit vector in $\mathrm{L}_2^{*\perp}$. In this case: $\mathbf{v}_0^T P_{\hat{\mathrm{L}}_2}\mathbf{v}_0 = (1-t)\,\mathbf{v}_0^T P_{\mathrm{L}_1^*}\mathbf{v}_0 + t\,\mathbf{v}_0^T P_{\mathrm{L}_2^*}\mathbf{v}_0 \leq 0 + 0 = 0$. Therefore, $\mathbf{v}_0 \in \hat{\mathrm{L}}_2^{\perp}$ and we obtain the following contradiction with (3.97): $\mathrm{L}_2^* = \hat{\mathrm{L}}_2$. Equation (3.99) is thus proved.