Asymptotics for High Dimension, Low Sample Size data and Analysis of Data on Manifolds

Sungkyu Jung

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research (Statistics).

Chapel Hill
2011

Approved by:

Advisor: : Dr. J. S. Marron

Reader: : Dr. Mark Foskey

Reader: : Dr. Jason Fine

Reader: : Dr. Yufeng Liu

Reader: : Dr. Jan Hannig

ABSTRACT

SUNGKYU JUNG: Asymptotics for High Dimension, Low Sample Size data and Analysis of Data on Manifolds.

(Under the direction of Dr. J. S. Marron.)

The dissertation consists of two research topics regarding modern non-standard data analytic situations. In particular, data under the High Dimension, Low Sample Size (HDLSS) situation and data lying on manifolds are analyzed. These situations are related to the statistical image and shape analysis.

The first topic is an asymptotic study of the high dimensional covariance matrix. In particular, the behavior of eigenvalues and eigenvectors of the covariance matrix is analyzed, which is closely related to the method of Principal Component Analysis (PCA). The asymptotic behavior of the Principal Component (PC) directions, when the dimension tends to infinity with the sample size fixed, is investigated. We have found mathematical conditions which characterize the consistency and the strong inconsistency of the empirical PC direction vectors. Moreover, the conditions where the empirical PC direction vectors are neither consistent nor strongly inconsistent are revealed, and the limiting distributions of the angle formed by the empirical PC direction and the population counterpart are presented. These findings help to understand the use of PCA in the HDLSS context, which is justified when the conditions for the consistency occur.

The second part of the dissertation studies data analysis methods for data lying in curved manifolds that are the features from shapes or images. A common goal in statistical shape analysis is to understand variation of shapes. As a means of dimension reduction and visualization, there is a need to develop PCA-like methods for manifold data. We propose flexible extensions of PCA to manifold data: Principal Arc Analysis and Analysis of Principal Nested Spheres. The methods are implemented to two important types of manifolds. The sample space of the medial representation of shapes, frequently used in image analysis to parameterize the shape of human organs, naturally forms curved manifolds, which we characterize as direct product manifolds. Another type of manifolds we consider is the landmark-based shape space, proposed by Kendall. The proposed methods in the dissertation capture major variations along non-geodesic paths. The benefits of the methods are illustrated by several data examples from image and shape analysis.

ACKNOWLEDGEMENTS

I thank Professor Steve Marron for his advice and support. His support enabled me to enjoy my research and to complete the dissertation work successfully. He helped me explore the subject with his keen insight, and provided many opportunities for various kinds of collaborative research. Every result in the dissertation is indepted to his guidance.

This dissertation work would have not been completed without valuable comments and suggestions of coauthors. Professor Arush Sen lead me to take another leap in the theoretical part of the dissertation in Chapter 3. Dr. Mark Foskey proposed to use small circles, and investigations since then results in the work presented in Chapter 6. Professor Ian Dryden introduced statistical shape analysis in which Chapter 7 is rooted. Professor Steve Pizer introduced the broad area of image analysis and motivated to develop Part II of the dissertation.

I wish to express my sincere appreciation to senior researchers for their valuable advice. Professor Steve Pizer has been a great mentor. Conversations with him on topics ranging from personal life decisions to deep mathematical questions have helped me a lot during the Ph.D. program. Professor Jason Fine and Professor Yufeng Liu gave me advices in making difficult decisions. I also appreciate valuable academic discussions and conversations with Professor Stephan Huckemann, Professor Anuj Srivastava, Professor Hongtu Zhu, Dr. Xiaoxiao Liu, Dr. Ja-Yeon Jeong and the community of MIDAG.

My life at Chapel Hill during the program was happy and fulfilling. I can not imagin it without friends: Classmates, gangs who flocked at Baity Hill and Finley Forest, and basketball and tennis players I loved to play with. I must thank my family—parents, my sister and her family—who have been very supportive in planning and achieving my life goal.

Last, my deepest gratitude goes to my wife, Seo Young, for her kindness as family and for valuable opinions as a fellow statistician. She always believes in me and helps to reach greater goals. For your patience and understanding, I will always love you, Seo Young.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This dissertation demonstrates development, theoretical study, and implementation of statistical methods for non-standard data, where conventional statistical methods are sometimes not directly applicable. The non-standard data are increasingly emerging, examples of which are the data with High Dimension, Low Sample Size (HDLSS, or the large p, small n) and the data that naturally lie on curved manifolds. Specifically, datasets with those properties are frequently observed in image and shape analysis, genomics, and functional data. These application areas together span a new statistical field in which traditional concepts need to be re-considered and development of new methodologies is required. The dissertation addresses two important aspects of this broad field: The HDLSS asymptotics, where the limiting operation has the dimension growing with the sample size fixed, to understand the use of Principal Component Analysis (PCA) in the HDLSS context; Development of PCA-like methods for manifold data, where the proposed methods intuitively capture major non-linear variations in lower dimension.

1.1 Motivations and Problems

The work is mainly motivated by statistical problems arising in image and shape analysis. The image analysis concerns extraction of information from an image, or set of images, usually in a digital form. The field of image analysis is broad, where mathematician, computer scientists, engineers, medical researchers and statisticians have contributed in different aspects of the area, separately or in collaborative fashion. Statistical image analysis puts a distributional assumption on images (either on the population of images or an image itself), develops statistical procedures to obtain meaningful information from the images. A subcategory of the broad field concerns an object found in images (e.g. human brain from an image of MRI), where the set of objects is under investigation. The statistical shape analysis focuses on the object *shape* in the image.

A shape can be represented by many different methods, examples of which are the medial representations (Siddiqi and Pizer (2008)), a point correspondence model and spherical harmonic representations (Gerig et al. (2004)), the traditional landmark-based shapes (Bookstein (1991), Dryden and Mardia (1998)) and a newly developed functional representation for shapes (Srivastava et al. (2010)). A challenge in shape analysis is that natural sample spaces of those shape representations are usually not Euclidean spaces but curved manifolds. In particular, the statistical methods that benefit from Euclidean geometry are not directly applicable for the manifold-valued objects. This motivates statistical study of manifold-valued random variables.

Another challenge in the image analysis is that the dimensionality of the representations is very high. Due to advances in modern technology, obtaining high resolution images became easier. However, the sample size (i.e. the number of observations) stays low especially in the medical image analysis because of the high cost in obtaining medical images such as MRI and CT. Hall et al. (2005) termed the situation as High Dimension, Low Sample Size (HDLSS) context, which becomes increasingly common and is observed not only in medical imaging but also in genomics, chemometrics, and functional analysis.

In the broad field of image analysis, we focus on data analyses in the sample space of the representations, called *feature space.* Specifically, we wish to give insight into the following basic statistical tasks:

1. Exploratory statistics and visualization of important structure: Given a random object or a set of observations of those in a non-Euclidean and multi-dimensional feature space, the first task for a statistician is to provide exploratory statistics. This includes finding a central location (mean), measure of variability (variance), and patterns in the data set.

2. Dimension reduction: When the dimension of the Euclidean feature space is high, we

wish to find an affine subspace of smaller dimension in a way that we do not lose important variation. If the feature space is a non-Euclidean curved manifold, we wish to find a sub-manifold of smaller dimension with great variance contained.

3. Estimation of probability distribution: A probability distribution can be assumed with fewer parameters if we choose to work with the subspace (or the sub-manifold) found in the dimension reduction.

The Principal Component Analysis (PCA) plays an important role in all of those problems.

PCA is often used in multivariate analysis for dimension reduction and visualization of important modes of variation. It is also understood as an estimation of the population principal components by the sample principal components. PCA is commonly credited to Hotelling (1933) and its common references are Muirhead (1982) and Jolliffe (2002). The objective of PCA is to capture most of the variability, of a multidimensional random vector or a set of observations consisting of a large number of measurements, in a low dimensional subspace. When the sample space is a Euclidean (linear) space, PCA finds a sequence of affine subspaces or a set of orthogonal direction vectors. Optimal choice of this sequence can be characterized in two equivalent ways: maximization of the variance captured by the affine subspace or minimization of the variance of the residuals. In the Euclidean space, PCA is usually computed through the eigen-decomposition of covariance matrices.

We discuss two different views of the PCA, namely the forward and backward stepwise views, later in Part II of the dissertation. These views are first introduced in Marron et al. (2010), which stems from the work in this dissertation.

Now, consider using PCA in the image or shape analysis, which in fact is a common routine for data analysts. While the traditional PCA is well defined for random vectors, we have two concerns; First of all, the sample size is usually too low but the dimension of the sample space is very high; Second, the natural sample space of the features of images or shapes is not the flat Euclidean space, but a curved manifold. A simple application of Euclidean PCA usually fails in the case. There are two different but related research questions:

1. What is the behavior of PCA in the high dimension low sample size context?

2. How to generalize PCA for manifold-valued objects?

Part I and II of the dissertation answer the first and second questions, respectively. A short summary is given in the next section.

1.2 Summary and Contributions

In Part I of the dissertation, Euclidean PCA is asymptotically studied when the dimension d grows, while the sample size n is fixed. In particular, we are interested in the situations when the PCA works in the HDLSS situation and when it fails. The success and failure of PCA are well described by the consistency of the empirical Principal Component (PC) directions $\hat{u}_i$ (or the ith eigenvector of the covariance matrix, see Section 2.1) with its population counterpart u_i, under the limiting operation $d \to \infty$ and n fixed. Since the size of the direction vectors $u_i, \hat{u}_i$ are increasing as d grows, the discrepancy between directions is measured by the *angle* formed by the directions. We say $\hat{u}_i$ is :

- *consistent* with u_i if $\text{Angle}(\hat{u}_i, u_i) \to 0$ as $d \to \infty$;
- *strongly inconsistent* with u_i if $\text{Angle}(\hat{u}_i, u_i) \to \pi/2$ as $d \to \infty$.

While the consistency is the case we desire, the strong inconsistency is somewhat counterintuitive since the angle $\pi/2$ is indeed the largest possible value of the angle. In other words, the estimate $\hat{u}_i$ of u_i loses its connection to the population structure and becomes completely arbitrary. We see that in HDLSS asymptotics when $\hat{u}_i$ is not consistent, it is often not just inconsistent but strongly inconsistent. The mathematical mechanism of these situations is the main topic of Chapter 2. In short, when the population PC variance is large such as of order d^α, $\alpha > 1$, the corresponding PC direction estimate is consistent. If not so large, i.e. $\alpha < 1$, then the corresponding PC direction estimate is strongly inconsistent.

The distributional assumptions in the study are general and intuitive. Gaussianity is relaxed into a ρ-mixing under some permutation and a moment condition. While the ρ-mixing assumption is most appropriate for a time series data set (because of the ordering of the variables), we relax the assumption by allowing any permutation of variable ordering. Thus the assumption makes sense for gene expression microarray data and image or shape data. See Section 2.1.1 for detailed assumptions.

We also considered a case where PC directions are not distinguishable, which occurs when some eigenvalues of the covariance matrix are identical. In Chapter 2, we consider a subspace spanned by those indistinguishable PC directions, and develop a notion of subspace consistency.

A natural question arises after we see that the exponent α in the order of PC variance d^{α} is the main driver between consistency ($\alpha > 1$) and strong inconsistency ($\alpha < 1$). What can we expect at the boundary between two cases? Chapter 3 discusses topics related to this question. We find that the angles do not degenerate (either to 0 or $\pi/2$) but weakly converge to distributions with support in $(0, \pi/2)$. We provide explicit forms of the limiting distributions in Sections 3.2 and 3.3. We further extend the HDLSS geometric representation (i.e. modulo rotation, n independent samples converge to vertices of a regular n simplex) found in Hall et al. (2005) and Ahn et al. (2007) into *three* different HDLSS representations, which gives an intuition in understanding a transition between consistency and strong inconsistency.

Part I of the dissertation is concluded with a discussion on open problems and a literature review; see Chapter 4.

The second part of the dissertation introduces some developments in generalization of PCA to manifold-valued data. We begin by introducing examples of specific manifolds, that are feature spaces in image analysis (Section 5.1). This gives a motivation to focus on specific types of manifolds. We briefly describe some basic Riemannian geometry used in the analysis and some specific manifolds we focus in Section 5.2. In particular, we focus on spheres and direct product of those with Euclidean space.

A common goal in analyzing these manifold data is to understand the variability. PCA, when appropriately modified and implemented for manifolds, provides a very effective means of analyzing the major modes of variation. The main contribution of Part II is to provide flexible extensions of PCA to manifold data. Briefly, the main idea in improving PCA is to make the method adaptive to certain non-linear structures of the data while the desirable properties of PCA are inherited.

To elaborate the work intuitively, let us assume that the manifold is the usual unit sphere $S^2 := \{x \in \mathbb{R}^3 : \|x\| = 1\}$. An analogue of a straight line on a sphere is a great circle, also referred to as a geodesic. A principal component direction on the sphere is represented by a great circle. As in the standard PCA, one may find a mean first then fit a great circle as the first PC direction. The second and higher PC directions are found among all great circles orthogonal to the first PC (geodesic). However, we found out that this standard approach may lead to an ineffective representation of the data, as discussed in Section 6.2. To resolve this issue, we took a reverse viewpoint in generalizing PCA, called *backward* approach, where the dimensionality of approximating subspaces (or submanifolds) is successively reduced. That is, analogous to the backward variable selection in regression, we remove the least important component first. This backward approach agrees with the standard PCA in Euclidean space but not in manifolds. The backward stepwise viewpoint gives a common root in generalization of PCA to non-linear spaces. In Section 5.4 the two different viewpoints of the usual PCA is discussed, and we point out that the backward approach is suitable for extension of PCA to manifold data.

The backward approach not only works well when the standard approach fails, but also gives a flexibility in capturing variations. The generalized PCA by the backward approach can be relaxed to find a small circle (e.g. the Tropic of Cancer) of the sphere, adaptive to the data. The resulting principal arcs capture the variation in the data set more succinctly and serve as principal components for visualization and dimension reduction.

The two methods proposed in Part II are generalizations of PCA to manifold-valued variables. Chapter 6 discusses *Principal Arc Analysis* for direct product manifolds; Chapter 7 discusses *Analysis of Principal Nested Spheres* for hyperspheres and the shape spaces of Kendall (1984).

The direct product manifold is a class of specially structured manifolds. The sample spaces of medial shape representations, diffeomorphisms, point distribution models, points and normals models and size-and-shapes can all be considered as a special form of direct product manifolds. We propose to use a transformation of non-Euclidean data into Euclidean coordinates, then a composite space of Euclidean data and the transformed non-Euclidean data gives a space suitable for further analysis including non-linear dimension reduction, visualization, and modeling with fewer parameters. Principal Arc Analysis consists of *1)* estimation of the principal circle, which is the basis for transformations, *2)* choices of the transformations, and *3)* ad-hoc tests to suppress overfitting in principal arcs. Chapter 6 also discusses an application of Principal Arc Analysis to medial shape representations data set.

Analysis of Principal Nested Spheres can be viewed as an extension of Principal Arc Analysis to higher dimensional spheres S^d and Kendall's shape space. Since the dimension of the sphere is now higher than the simple 2, the geometry involved is a bit more complicated than the usual sphere. We define nested spheres, which are in a special form of sub-manifolds of S^d, as a basis for decomposition of the hypersphere. Geometry of nested spheres are discussed in depth in Section 7.8. Simply put, Principal Nested Spheres are fitted nested spheres that capture most of the variances in the data. We discuss application of Analysis of Principal Nested Spheres to Kendall's shape space in Section 7.6, as well as several real data examples in Section 7.7. We exemplify that the proposed method results in a succinct description of real data set in fewer dimension than alternative methods in literature.

Part I

PCA in High Dimension, Low Sample Size Context

Chapter 2

PCA consistency in HDLSS context

Since traditional tools of multivariate statistical analysis were designed for the $n > d$ case, it should not be surprising that they often do not work properly for the HDLSS data. However, the method of PCA has been seen to sometimes be very effective in the HDLSS context, and is widely used as a dimension reduction technique. In this chapter we investigate the asymptotic behaviors of PCA when d tends to infinity while n is fixed, which provide the appropriate analysis to study the HDLSS context. In short, the asymptotics characterizes when Euclidean PCA works and when it fails. The work presented in this chapter is based on Jung and Marron (2009).

2.1 Introduction and summary

The High Dimension, Low Sample Size (HDLSS) data situation occurs in many areas of modern science and the asymptotic studies of this type of data are becoming increasingly relevant. We will focus on the case that the dimension d increases while the sample size n is fixed as done in Hall et al. (2005) and Ahn et al. (2007). The d-dimensional covariance matrix is challenging to analyze in general since the number of parameters is $\frac{d(d+1)}{2}$, which increases even faster than d. Instead of assessing all of the parameter estimates, the covariance matrix is usually analyzed by Principal Component Analysis (PCA). PCA is often used to visualize important structure in the data, as shown in Figure 2.1. The data in Figure 1, described in detail in Bhattacharjee et al. (2001) and Liu et al. (2008), are from a microarray study of lung cancer. Different symbols correspond to cancer subtypes and Figure 2.1 shows the projections of the data onto the subspaces generated by PC1 and PC2 (left panel) and PC1 and PC3

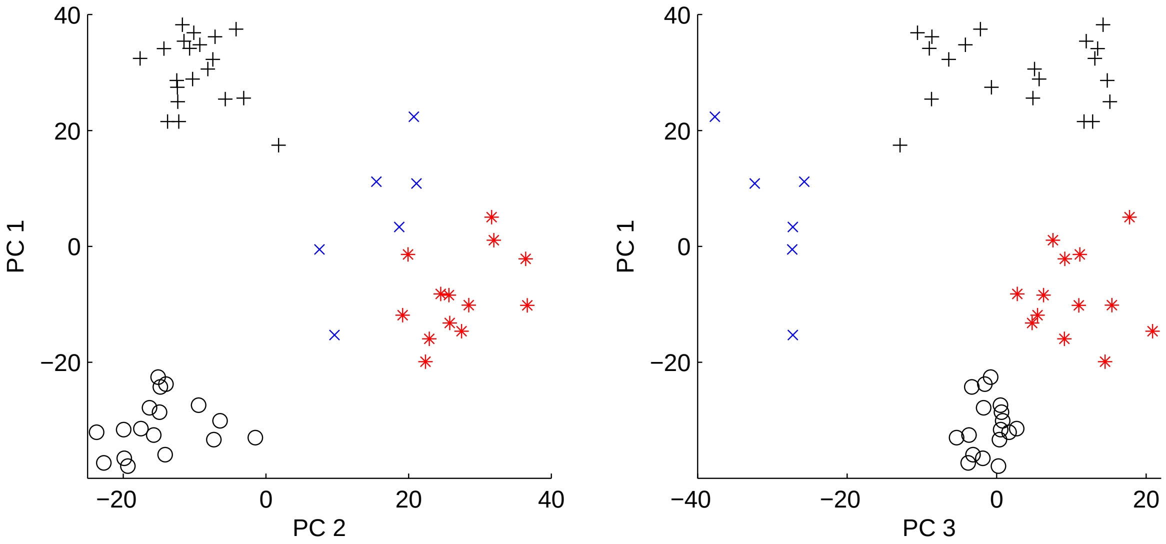


Figure 2.1: Scatterplots of data projected on the first three PC directions. The dataset contains 56 patients with 2530 genes. There are 20 Pulmonary Carcinoid (plotted as $+$), 13 Colon Cancer Metastases ($*$), 17 Normal Lung ($\circ$), and 6 Small Cell Carcinoma ($\times$). In spite of the high dimensionality, PCA reveals important structure in the data.

(right panel, respectively) directions. This shows the difference between subtypes is so strong that it drives the first three principal components. This illustrates common occurrence: the data have important underlying structure which is revealed by the first few PC directions.

PCA is also used to reduce dimensionality by approximating the data with the first few principal components.

For both visualization and data reduction, it is critical that the PCA empirical eigenvectors reflect true underlying distributional structure. Hence our focus is on the underlying mechanism which determines when the sample PC directions converge to their population counterparts as $d \to \infty$. For example, we quantify situations where the dominant sample eigenvectors are revealing important underlying structure in the data as in Figure 2.1. In general we assume $d > n$. Since the size of the covariance matrix depends on d, the population covariance matrix is denoted as Σ_d and similarly the sample covariance matrix, S_d, so that their dependency on the dimension is emphasized. PCA is done by eigen-decomposition of a covariance matrix. The eigen-decomposition of Σ_d is

$$\Sigma_d = U_d \Lambda_d U_d',$$

where Λ_d is a diagonal matrix of eigenvalues $\lambda_{1,d} \geqslant \lambda_{2,d} \geqslant \cdots \geqslant \lambda_{d,d}$ and U_d is a matrix of corresponding eigenvectors so that $U_d = [u_{1,d}, u_{2,d}, \ldots, u_{d,d}]$. S_d is similarly decomposed as

$$S_d = \hat{U}_d \hat{\Lambda}_d \hat{U}_d'.$$

Ahn et al. (2007) developed the concept of HDLSS consistency which was the first investigation of when PCA could be expected to find important structure in HDLSS data. Our main results are formulated in terms of three related concepts:

1. *consistency*: The direction $\hat{u}_{i,d}$ is *consistent* with its population counterpart $u_{i,d}$ if $\text{Angle}(u_{i,d}, \hat{u}_{i,d}) \longrightarrow 0$ as $d \to \infty$. The growth of dimension can be understood as adding more variation. The consistency of sample eigenvectors occurs when the added variation supports the existing structure in the covariance or is small enough to be ignored.

2. *strong inconsistency*: In situations where $\hat{u}_{i,d}$ is not consistent, a perhaps counter-intuitive HDLSS phenomenon frequently occurs. In particular, $\hat{u}_{i,d}$ is said to be *strongly inconsistent* with its population counterpart $u_{i,d}$ in the sense that it tends to be as far away from $u_{i,d}$ as possible, that is, $\text{Angle}(u_{i,d}, \hat{u}_{i,d}) \longrightarrow \frac{\pi}{2}$ as $d \to \infty$. Strong inconsistency occurs when the added variation obscures the underlying structure of the population covariance matrix.

3. *subspace consistency*: When several population eigenvalues indexed by $j \in J$ are similar, the corresponding sample eigenvectors may not be distinguishable. In this case, $\hat{u}_{j,d}$ will not be consistent for $u_{j,d}$ but will tend to lie in the linear span, $\text{span}\{u_{j,d} : j \in J\}$. This motivates the definition of convergence of a direction $\hat{u}_{i,d}$ to a subspace, called *subspace consistency*;

   $$\text{Angle}(\hat{u}_{i,d}, \text{span}\{u_{j,d} : j \in J\}) \longrightarrow 0$$

   as $d \to \infty$. This definition essentially comes from the theory of *canonical angles* discussed by Gaydos (2008). That theory also gives a notion of convergence of subspaces, that could be developed here.

In this chapter, a broad and general set of conditions for consistency and strong inconsistency are provided. Section 2.2 develops conditions that guarantee the non-zero eigenvalues of the sample covariance matrix tend to a increasing constant, which are much more general than those of Hall et al. (2005) and Ahn et al. (2007). This asymptotic behavior of the sample covariance matrix is the basis of the geometric representation of HDLSS data. Our result gives broad new insight into this representation as discussed in section 2.3. The central issue of consistency and strong inconsistency is developed in section 2.4, as a series of theorems. For a fixed number κ, we assume the first κ eigenvalues are much larger than the others. We show that when $\kappa = 1$, the first sample eigenvector is consistent and the others are strongly inconsistent. We also generalize to the $\kappa > 1$ case, featuring two different types of results (consistency and subspace consistency) according to the asymptotic behaviors of the first κ eigenvalues. All results are combined and generalized in the main theorem (Theorem 2.5). Proofs of theorems are given in section 2.5.

Further discussions on the *boundary* case where the PC directions are neither consistent nor strongly inconsistent are deferred to Chapter 3. Relevant literatures are reviewed and discussed later at Chapter 4.

2.1.1 General setting

Suppose we have a $d \times n$ data matrix $X_{(d)} = [X_{1,(d)}, \ldots, X_{n,(d)}]$ with $d > n$, where the d-dimensional random vectors $X_{1,(d)}, \ldots, X_{n,(d)}$ are independent and identically distributed. We assume that each $X_{i,(d)}$ follows a multivariate distribution (which does not have to be Gaussian) with mean zero and covariance matrix Σ_d. Define the sphered data matrix $Z_{(d)} = \Lambda_d^{-1/2} U_d' X_{(d)}$. Then the components of the $d \times n$ matrix $Z_{(d)}$ have unit variances, and are uncorrelated with each other. We shall regulate the dependency (recall for non-Gaussian data, uncorrelated variables can still be dependent) of the random variables in $Z_{(d)}$ by a ρ-mixing condition. This allows serious weakening of the assumptions of Gaussianity while still enabling the law of large numbers that lie behind the geometric representation results of Hall et al. (2005).

The concept of ρ-mixing was first developed by Kolmogorov and Rozanov (1960). See Bradley (2005) for a clear and insightful discussion. For $-\infty \leqslant J \leqslant L \leqslant \infty$, let $\mathcal{F}_J^L$ denote the σ-field of events generated by the random variables $(Z_i,\ J \leqslant i \leqslant L)$. For any σ-field $\mathcal{A}$, let

$L_2(\mathcal{A})$ denote the space of square-integrable, $\mathcal{A}$ measurable (real-valued) random variables. For each $m \geqslant 1$, define the maximal correlation coefficient

$$\rho(m) := \sup |\mathrm{corr}(f, g)|, \quad f \in L_2(\mathcal{F}_{-\infty}^{j}),\ g \in L_2(\mathcal{F}_{j+m}^{\infty}),$$

where sup is over all f, g and $j \in \mathbf{Z}$. The sequence $\{Z_i\}$ is said to be ρ-mixing if $\rho(m) \to 0$ as $m \to \infty$.

While the concept of ρ-mixing is useful as a mild condition for the development of laws of large numbers, its formulation is critically dependent on the ordering of variables. For many interesting data types, such as microarray data, there is clear dependence but no natural ordering of the variables. Hence we assume that there is some permutation of the data which is ρ-mixing. In particular, let $\{Z_{ij,(d)}\}_{i=1}^{d}$ be the components of the jth column vector of $Z_{(d)}$. We assume that for each d, there exists a permutation $\pi_d : \{1, \ldots, d\} \longmapsto \{1, \ldots, d\}$ so that the sequence $\{Z_{\pi_d(i)j,(d)} : i = 1, \ldots, d\}$ is ρ-mixing.

In the following, all the quantities depend on d, but the subscript d will be omitted for the sake of simplicity when it does not cause any confusion. The sample covariance matrix is defined as $S = n^{-1}XX'$. We do not subtract the sample mean vector because the population mean is assumed to be 0. Since the dimension of the sample covariance matrix S grows, it is challenging to deal with S directly. A useful approach is to work with the *dual* of S. The dual approach switches the role of columns and rows of the data matrix, by replacing X by X'. The $n \times n$ *dual sample covariance matrix* is defined as $S_D = n^{-1}X'X$. An advantage of this dual approach is that S_D and S share non-zero eigenvalues. If we write X as $U\Lambda^{\frac{1}{2}}Z$ and use the fact that U is a unitary matrix,

$$nS_D = (Z'\Lambda^{\frac{1}{2}}U')(U\Lambda^{\frac{1}{2}}Z) = Z'\Lambda Z = \sum_{i=1}^{d} \lambda_{i,d} z_i' z_i, \tag{2.1}$$

where the z_i's, $i = 1, \ldots, d$, are the row vectors of the matrix Z. Note that nS_D is commonly referred to as the *Gram matrix*, consisting of inner products between observations.

2.2 HDLSS asymptotic behavior of the sample covariance matrix

In this section, we investigate the behavior of the sample covariance matrix S when $d \to \infty$ and n is fixed. Under mild and broad conditions, the eigenvalues of S, or the dual S_D, behave asymptotically as if they are from the identity matrix. That is, the set of sample eigenvectors tends to be an arbitrary choice. This lies at the heart of the geometric representation results of Hall et al. (2005) and Ahn et al. (2007) which are studied more deeply in section 2.3. We will see that this condition readily implies the strong inconsistency of sample eigenvectors, see Theorem 2.5.

The conditions for the theorem are conveniently formulated in terms of a measure of sphericity

$$\epsilon \equiv \frac{\mathrm{tr}^2(\Sigma)}{d\,\mathrm{tr}(\Sigma^2)} = \frac{(\sum_{i=1}^d \lambda_{i,d})^2}{d\sum_{i=1}^d \lambda_{i,d}^2},$$

proposed and used by John (1971, 1972) as the basis of a hypothesis test for equality of eigenvalues. Note that these inequalities always hold:

$$\frac{1}{d} \leqslant \epsilon \leqslant 1.$$

Also note that perfect sphericity of the distribution (i.e. equality of eigenvalues) occurs only when $\epsilon = 1$. The other end of the ϵ range is the most singular case where in the limit as the first eigenvalue dominates all others.

Ahn et al. (2007) claimed that if $\epsilon \gg \frac{1}{d}$, in the sense that $\epsilon^{-1} = o(d)$, then the eigenvalues of S_D tend to be identical in probability as $d \to \infty$. However, they needed an additional assumption (e.g. a Gaussian assumption on $X_{(d)}$) to have independence among components of $Z_{(d)}$, as described in example 2.1. We extend this result to the case of arbitrary distributions with dependency regulated by the ρ-mixing condition as in section 2.1.1, which is much more general than either a Gaussian or an independence assumption. We also explore convergence in the almost sure sense with stronger assumptions. Our results use a measure of sphericity for part of the eigenvalues for conditions of a.s. convergence and also for later use in section 2.4.

In particular, define the measure of sphericity for $\{\lambda_{k,d}, \dots, \lambda_{d,d}\}$ as

$$\epsilon_k \equiv \frac{(\sum_{i=k}^d \lambda_{i,d})^2}{d \sum_{i=k}^d \lambda_{i,d}^2}.$$

For convenience, we name several assumptions used in this chapter made about the measure of sphericity ϵ:

- *The ϵ-condition*: $\epsilon \gg \frac{1}{d}$, i.e.

$$(d\epsilon)^{-1} = \frac{\sum_{i=1}^d \lambda_{i,d}^2}{(\sum_{i=1}^d \lambda_{i,d})^2} \to 0 \text{ as } d \to \infty. \tag{2.2}$$

- *The ϵ_k-condition*: $\epsilon_k \gg \frac{1}{d}$, i.e.

$$(d\epsilon_k)^{-1} = \frac{\sum_{i=k}^d \lambda_{i,d}^2}{(\sum_{i=k}^d \lambda_{i,d})^2} \to 0 \text{ as } d \to \infty. \tag{2.3}$$

- *The strong ϵ_k-condition*: For some fixed $l \geqslant k$, $\epsilon_l \gg \frac{1}{\sqrt{d}}$, i.e.

$$d^{-\frac{1}{2}} \epsilon_l^{-1} = \frac{d^{\frac{1}{2}} \sum_{i=l}^d \lambda_{i,d}^2}{(\sum_{i=l}^d \lambda_{i,d})^2} \to 0 \text{ as } d \to \infty. \tag{2.4}$$

Remark 2.1. Note that the ϵ_k-condition is identical to the ϵ-condition when $k = 1$. Similarly, the strong ϵ_k-condition is also called *the strong ϵ-condition* when $k = 1$. The strong ϵ_k-condition is stronger than the ϵ_k condition if the minimum of l's which satisfy (2.4), l_o, is as small as k. But, if $l_o > k$, then this is not necessarily true. We will use the strong ϵ_k-condition combined with the ϵ_k-condition.

Note that the ϵ-condition is quite broad in the spectrum of possible values of ϵ: It only avoids the most singular case. The strong ϵ-condition further restricts ϵ_l to essentially in the range $(\frac{1}{\sqrt{d}}, 1]$.

The following theorem states that if the (strong) ϵ-condition holds for Σ_d, then the sample eigenvalues behave as if they are from a scaled identity matrix. It uses the notation I_n for the $n \times n$ identity matrix.

Theorem 2.1. *For a fixed* n*, let* $\Sigma_d = U_d \Lambda_d U_d'$*,* $d = n+1, n+2, \ldots$ *be a sequence of covariance matrices. Let* $X_{(d)}$ *be a* $d \times n$ *data matrix from a* d*-variate distribution with mean zero and covariance matrix* Σ_d*. Let* $S_d = \hat{U}_d \hat{\Lambda}_d \hat{U}_d'$ *be the sample covariance matrix estimated from* $X_{(d)}$ *for each* d *and let* $S_{D,d}$ *be its dual.*

(1) Assume that the components of $Z_{(d)} = \Lambda_d^{-\frac{1}{2}} U_d' X_{(d)}$ *have uniformly bounded fourth moments and are* ρ*-mixing under some permutation. If (2.2) holds, then*

$$c_d^{-1} S_{D,d} \longrightarrow I_n, \tag{2.5}$$

in probability as $d \to \infty$*, where* $c_d = n^{-1} \sum_{i=1}^{d} \lambda_{i,d}$*.*

(2) Assume that the components of $Z_{(d)} = \Lambda_d^{-\frac{1}{2}} U_d' X_{(d)}$ *have uniformly bounded eighth moments and are independent to each other. If both (2.2) and (2.4) hold, then* $c_d^{-1} S_{D,d} \to I_n$ *almost surely as* $d \to \infty$*.*

The (strong) ϵ-condition holds for quite general settings. The strong ϵ-condition combined with the ϵ-condition holds under;

(a) Null case: All eigenvalues are the same.

(b) Mild spiked model: The first m eigenvalues are moderately larger than the others, for example, $\lambda_{1,d} = \cdots = \lambda_{m,d} = C_1 \cdot d^{\alpha}$ and $\lambda_{m+1,d} = \cdots = \lambda_{d,d} = C_2$, where $m < d$, $\alpha < 1$ and $C_1, C_2 > 0$.

The ϵ-condition fails when;

(c) Singular case: Only the first few eigenvalues are non-zero.

(d) Exponential decrease: $\lambda_{i,d} = c^{-i}$ for some $c > 1$.

(e) Sharp spiked model: The first m eigenvalues are much larger than the others. One example is the same as (b) but $\alpha \geqslant 1$.

The polynomially decreasing case, $\lambda_{i,d} = i^{-\beta}$, is interesting because it depends on the power β;

(f-1) The strong ϵ-condition holds when $0 \leqslant \beta < \frac{3}{4}$.

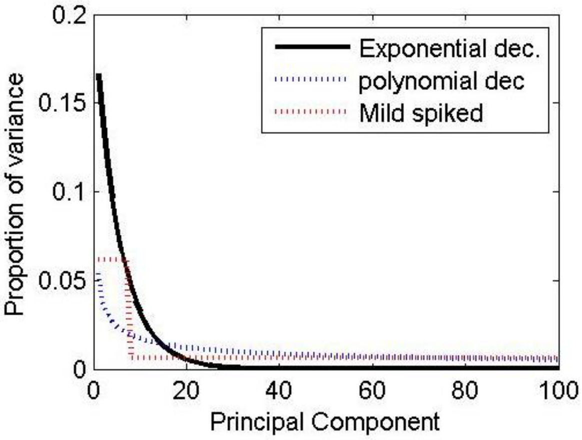


Figure 2.2: The ϵ-condition is satisfied by the mild spike and polynomial decrease case, but is satisfied by the exponential decrease.

(f-2) The ϵ-condition holds but the strong ϵ-condition fails when $\frac{3}{4} \leqslant \beta \leqslant 1$.

(f-3) The ϵ-condition fails when $\beta > 1$.

Another family of examples that includes all three cases is the spiked model with the number of spikes increasing, for example, $\lambda_{1,d} = \cdots = \lambda_{m,d} = C_1 \cdot d^\alpha$ and $\lambda_{m+1,d} = \cdots = \lambda_{d,d} = C_2$, where $m = \lfloor d^\beta \rfloor$, $0 < \beta < 1$ and $C_1, C_2 > 0$;

(g-1) The strong ϵ-condition holds when $0 \leqslant 2\alpha + \beta < \frac{3}{2}$;

(g-2) The ϵ-condition holds but the strong ϵ-condition fails when $\frac{3}{2} \leqslant 2\alpha + \beta < 2$;

(g-3) The ϵ-condition fails when $2\alpha + \beta \geqslant 2$.

Figure 2.2 shows some examples depicting when the ϵ-condition holds and fails.

2.3 Geometric representation of HDLSS data

Suppose $X \sim \mathcal{N}_d(0, I_d)$. When the dimension d is small, most of the mass of the data lies near origin. However with a large d, Hall et al. (2005) showed that Euclidean distance of X to the origin is described as

$$\|X\| = \sqrt{d} + o_p(\sqrt{d}). \tag{2.6}$$

Moreover the distance between two samples is also rather deterministic, i.e.

$$\|X_1 - X_2\| = \sqrt{2d} + o_p(\sqrt{d}). \tag{2.7}$$

These results can be derived by the law of large numbers. Hall et al. (2005) generalized those results under the assumptions that $d^{-1}\sum_{i=1}^{d} \mathrm{Var}(X_i) \to 1$ and $\{X_i\}$ is ρ-mixing.

Application of part (1) of Theorem 2.1 generalizes these results. Let $X_{1,(d)}$, $X_{2,(d)}$ be two samples that satisfy the assumptions of Theorem 2.1 part (1). Assume without loss of generality that $\lim_{d\to\infty} d^{-1}\sum_{i=1}^{d} \lambda_{i,d} = 1$. The scaled squared distance between two data points is

$$\frac{\|X_{1,(d)} - X_{2,(d)}\|^2}{\sum_{i=1}^{d} \lambda_{i,d}} = \sum_{i=1}^{d} \tilde{\lambda}_{i,d} z_{i1}^2 + \sum_{i=1}^{d} \tilde{\lambda}_{i,d} z_{i2}^2 - 2\sum_{i=1}^{d} \tilde{\lambda}_{i,d} z_{i1} z_{i2},$$

where $\tilde{\lambda}_{i,d} = \frac{\lambda_{i,d}}{\sum_{i=1}^{d} \lambda_{i,d}}$. Note that by (2.1), the first two terms are diagonal elements of $c_d^{-1} S_{D,d}$ in Theorem 2.1 and the third term is an off-diagonal element. Since $c_d^{-1} S_{D,d} \to I_n$, we have (2.7). (2.6) is derived similarly.

Remark 2.2. If $\lim_{d\to\infty} d^{-1}\sum_{i=1}^{d} \lambda_{i,d} = 1$, then the conclusion (2.5) of Theorem 2.1 part (1) holds if and only if the representations (2.6) and (2.7) hold under the same assumptions in the theorem.

In this representation, the ρ-mixing assumption plays a very important role. The following example, due to John Kent, shows that some type of mixing condition is important.

Example 2.1 (Strong dependency via a scale mixture of Gaussian). Let $X = Y_1 U + \sigma Y_2 (1-U)$, where Y_1, Y_2 are two independent $\mathcal{N}_d(0, I_d)$ random variables, $U = 0$ or 1 with probability $\frac{1}{2}$ and independent of Y_1, Y_2, and $\sigma > 1$. Then,

$$\|X\| = \begin{cases} d^{\frac{1}{2}} + O_p(1) & \text{w.p. } \frac{1}{2} \\ \sigma d^{\frac{1}{2}} + O_p(1) & \text{w.p. } \frac{1}{2} \end{cases}$$

Thus, (2.6) does not hold. Note that since $\mathrm{Cov}(X) = \frac{1+\sigma^2}{2} I_d$, the ϵ-condition holds and the variables are uncorrelated. However, there is strong dependency, i.e. $\mathrm{Cov}(z_i^2, z_j^2) = (\frac{1+\sigma^2}{2})^{-2}\mathrm{Cov}(x_i^2, x_j^2) = (\frac{1-\sigma^2}{1+\sigma^2})^2$ for all $i \neq j$ which implies that $\rho(m) > c$ for some $c > 0$, for all m. Thus, the ρ-mixing condition does not hold for all permutation. Note that, however, under Gaussian assumption, given any covariance matrix Σ, $Z = \Sigma^{-\frac{1}{2}}X$ has independent components.

Note that in the case $X = (X_1, \ldots, X_d)$ is a sequence of i.i.d. random variables, the results (2.6) and (2.7) can be considerably strengthened to $\|X\| = \sqrt{d} + O_p(1)$, and $\|X_1 - X_2\| = \sqrt{2d} + O_p(1)$. The following example shows that strong results are beyond the reach of reasonable assumption.

Example 2.2 (Varying sphericity). Let $X \sim \mathcal{N}_d(0, \Sigma_d)$, where $\Sigma_d = \mathrm{diag}(d^\alpha, 1, \ldots, 1)$ and $\alpha \in (0, 1)$. Define $Z = \Sigma_d^{-\frac{1}{2}}X$. Then the components of Z, z_i's, are independent standard Gaussian random variables. We get $\|X\|^2 = d^\alpha z_1^2 + \sum_{i=2}^d z_i^2$. Now for $0 < \alpha < \frac{1}{2}$, $d^{-\frac{1}{2}}(\|X\|^2 - d) \Rightarrow \mathcal{N}(0,1)$ and for $\frac{1}{2} < \alpha < 1$, $d^{-\alpha}(\|X\|^2 - d) \Rightarrow z_1^2$, where $\Rightarrow$ denotes convergence in distribution. Thus by the delta-method, we get

$$\|X\| = \begin{cases} \sqrt{d} + O_p(1), & \text{if } 0 < \alpha < \frac{1}{2}, \\ \sqrt{d} + O_p(d^{\alpha - \frac{1}{2}}), & \text{if } \frac{1}{2} < \alpha < 1. \end{cases}$$

In both cases, the representation (2.6) holds.

2.4 Consistency and strong inconsistency of PC directions

In this section, conditions for consistency or strong inconsistency of the sample PC direction vectors are investigated, in the general setting of section 2.1.1. The generic eigen-structure of the covariance matrix that we assume is the following. For a fixed number κ, we assume the first κ eigenvalues are much larger than others. (The precise meaning of *large* will be addressed shortly.) The rest of eigenvalues are assumed to satisfy the ϵ-condition, which is very broad in the range of sphericity. We begin with the case $\kappa = 1$ and generalize the result for $\kappa > 1$ in two distinct ways. The main theorem (Theorem 2.5) contains and combines those previous results and also embraces various cases according to the magnitude of the first κ eigenvalues. We also investigate the sufficient conditions for a stronger result, i.e. almost sure convergence, which involves use of the strong ϵ-condition.

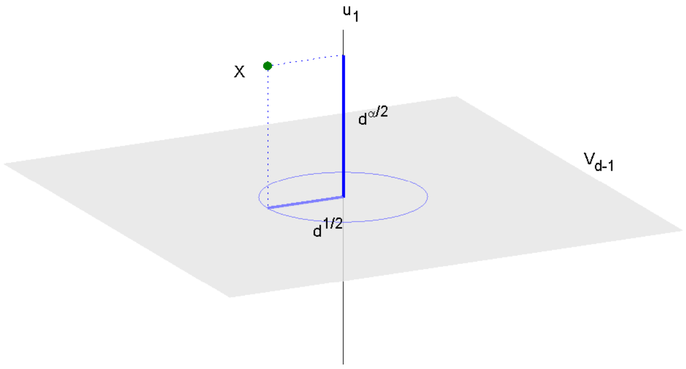


Figure 2.3: Projection of a d-dimensional random variable X onto u_1 and V_{d-1}. If $\alpha > 1$, then the subspace V_{d-1} becomes negligible compared to u_1 when $d \to \infty$

2.4.1 Criteria for consistency or strong inconsistency of the first PC direction

Consider the simplest case that only the first PC direction of S is of interest. Section 2.3 gives some preliminary indication of this. As an illustration, consider a spiked model as in Example 2.2 but now let $\alpha > 1$. Let $\{u_i\}$ be the set of eigenvectors of Σ_d and V_{d-1} be the subspace of all eigenvectors except the first one. Then the projection of X onto u_1 has a norm $\|\text{Proj}_{u_1} X\| = \|X_1\| = O_p(d^{\frac{\alpha}{2}})$. The projection of X onto V_{d-1} has a norm $\sqrt{d} + o_p(\sqrt{d})$ by (2.6). Thus when $\alpha > 1$, if we scale the whole data space $\mathbb{R}^d$ by dividing by $d^{\frac{\alpha}{2}}$, then $\text{Proj}_{V_{d-1}} X$ becomes negligible compared to $\text{Proj}_{u_1} X$. (See Figure 2.3.) Thus for a large d, $\Sigma_d \approx \lambda_1 u_1 u_1'$ and the variation of X is mostly along u_1. Therefore the sample eigenvector corresponding to the largest eigenvalue, $\hat{u}_1$, will be similar to u_1.

To generalize this, suppose the ϵ_2 condition holds. The following proposition states that under the general setting in section 2.1.1, the first sample eigenvector $\hat{u}_1$ converges to its population counterpart u_1 (consistency) or tends to be perpendicular to u_1 (strong inconsistency) according to the magnitude of the first eigenvalue λ_1, while all the other sample eigenvectors are strongly inconsistent regardless of the magnitude λ_1.

Proposition 2.2. *For a fixed n, let $\Sigma_d = U_d \Lambda_d U_d'$, $d = n+1, n+2, \dots$ be a sequence of covariance matrices. Let $X_{(d)}$ be a $d \times n$ data matrix from a d-variate distribution with mean zero and covariance matrix Σ_d. Let $S_d = \hat{U}_d \hat{\Lambda}_d \hat{U}_d'$ be the sample covariance matrix estimated from $X_{(d)}$ for each d. Assume the following:*

(a) The components of $Z_{(d)} = \Lambda_d^{-\frac{1}{2}} U_d' X_{(d)}$ *have uniformly bounded fourth moments and are* ρ*-mixing for some permutation.*

For an $\alpha_1 > 0$*,*

(b) $\dfrac{\lambda_{1,d}}{d^{\alpha_1}} \longrightarrow c_1$ *for some* $c_1 > 0$*,*

(c) The ϵ_2*-condition holds and* $\sum_{i=2}^{d} \lambda_{i,d} = O(d)$*.*

If $\alpha_1 > 1$*, then the first sample eigenvector is consistent and the others are strongly inconsistent in the sense that*

$$Angle(\hat{u}_1, u_1) \xrightarrow{p} 0 \text{ as } d \to \infty,$$

$$Angle(\hat{u}_i, u_i) \xrightarrow{p} \frac{\pi}{2} \text{ as } d \to \infty \ \forall i = 2, \ldots, n.$$

If $\alpha_1 \in (0, 1)$*, then all sample eigenvectors are strongly inconsistent, i.e.*

$$Angle(\hat{u}_i, u_i) \xrightarrow{p} \frac{\pi}{2} \text{ as } d \to \infty \ \forall i = 1, \ldots, n.$$

Note that the gap between consistency and strong inconsistency is very thin, i.e. if we avoid $\alpha_1 = 1$, then we have either consistency or strong inconsistency. Thus in the HDLSS context, asymptotic behavior of PC directions is mostly captured by consistency and strong inconsistency. Now it makes sense to say λ_1 is much larger than the others when $\alpha_1 > 1$, which results in consistency. Also note that if $\alpha_1 < 1$, then the ϵ-condition holds, which is in fact the condition for Theorem 2.1.

2.4.2 Generalizations

In this section, we generalize Proposition 2.2 to the case that multiple eigenvalues are much larger than the others. This leads to two different types of result.

First is the case that the first p eigenvectors are each consistent. Consider a covariance structure with multiple spikes, that is, p eigenvalues, $p > 1$, which are much larger than the others. In order to have consistency of the first p eigenvectors, we require that each of p eigenvalues has a distinct order of magnitude, for example, $\lambda_{1,d} = d^3$, $\lambda_{2,d} = d^2$ and sum of the rest is order of d.

Proposition 2.3. *For a fixed n, let Σ_d, $X_{(d)}$, and S_d be as before. Assume (a) of Proposition 2.2. Let $\alpha_1 > \alpha_2 > \cdots > \alpha_p > 1$ for some $p < n$. Suppose the following conditions hold:*

(b) $\dfrac{\lambda_{i,d}}{d^{\alpha_i}} \longrightarrow c_i$ *for some* $c_i > 0, \ \forall i = 1, \dots, p$

(c) The ϵ_{p+1}-condition holds and $\sum_{i=p+1}^{d} \lambda_{i,d} = O(d)$.

Then, the first p sample eigenvectors are consistent and the others are strongly inconsistent in the sense that

$$Angle(\hat{u}_i, u_i) \xrightarrow{p} 0 \text{ as } d \to \infty \ \forall i = 1, \dots, p,$$

$$Angle(\hat{u}_i, u_i) \xrightarrow{p} \frac{\pi}{2} \text{ as } d \to \infty \ \forall i = p+1, \dots, n.$$

Consider now a distribution having a covariance structure with multiple spikes as before. Let k be the number of spikes. An interesting phenomenon happens when the first k eigenvalues are of the same order of magnitude, i.e. $\lim_{d\to\infty} \frac{\lambda_{1,d}}{\lambda_{k,d}} = c > 1$ for some constant c. Then the first k sample eigenvectors are neither consistent nor strongly inconsistent. However, all of those random directions converge to the subspace spanned by the first k population eigenvectors. Essentially, when eigenvalues are of the same order, the eigen-directions can not be separated but are subspace consistent with the proper subspace.

Proposition 2.4. *For a fixed n, let Σ_d, $X_{(d)}$, and S_d be as before. Assume (a) of Proposition 2.2. Let $\alpha_1 > 1$ and $k < n$. Suppose the following conditions hold:*

(b) $\dfrac{\lambda_{i,d}}{d^{\alpha_1}} \longrightarrow c_i$ *for some* $c_i > 0, \ \forall i = 1, \dots, k$

(c) The ϵ_{k+1}-condition holds and $\sum_{i=k+1}^{d} \lambda_{i,d} = O(d)$.

Then, the first k sample eigenvectors are subspace-consistent with the subspace spanned by the first k population eigenvectors and the others are strongly inconsistent in the sense that

$$Angle(\hat{u}_i, span\{u_1, \dots, u_k\}) \xrightarrow{p} 0 \text{ as } d \to \infty \ \forall i = 1, \dots, k,$$

$$Angle(\hat{u}_i, u_i) \xrightarrow{p} \frac{\pi}{2} \text{ as } d \to \infty \ \forall i = k+1, \dots, n.$$

2.4.3 Main theorem

Propositions 2.2 - 2.4 are combined and generalized in the main theorem. Consider p groups of eigenvalues, which grow at the same rate within each group as in Proposition 2.4. Each group has a finite number of eigenvalues and the number of eigenvalues in all groups, κ, does not exceed n. Also similar to Proposition 2.3, let the orders of magnitude of the p groups be different to each other. We require that the $\epsilon_{\kappa+1}$-condition holds. The following theorem states that a sample eigenvector of a group converges to the subspace of population eigenvectors of the group.

Theorem 2.5 (Main theorem)**.** *For a fixed n, let Σ_d, $X_{(d)}$, and S_d be as before. Assume (a) of Proposition 2.2. Let $\alpha_1, \ldots, \alpha_p$ be such that $\alpha_1 > \alpha_2 > \cdots > \alpha_p > 1$ for some $p < n$. Let $k_1, \ldots, k_p$ be nonnegative integers such that $\sum_{j=1}^{p} k_j \doteq \kappa < n$. Let $k_0 = 0$ and $k_{p+1} = d - \kappa$. Let $J_1, \ldots, J_{p+1}$ be sets of indices such that*

$$J_l = \left\{ \sum_{j=0}^{l-1} k_j + 1, \sum_{j=0}^{l-1} k_j + 2, \ldots, \sum_{j=0}^{l-1} k_j + k_l \right\}, \quad l = 1, \ldots, p+1.$$

Suppose the following conditions hold:

(b) $\dfrac{\lambda_{i,d}}{d^{\alpha_l}} \longrightarrow c_i$ *for some* $c_i > 0, \ \forall i \in J_l, \ \forall l = 1, \ldots, p$

(c) The $\epsilon_{\kappa+1}$-condition holds and $\sum_{i \in J_{p+1}} \lambda_{i,d} = O(d)$.

Then, the sample eigenvectors whose label is in the group J_l, for $l = 1, \ldots, p$, are subspace-consistent with the space spanned by the population eigenvectors whose labels are in J_l and the others are strongly inconsistent in the sense that

$$Angle(\hat{u}_i, span\{u_j : j \in J_l\}) \xrightarrow{p} 0 \ as \ d \to \infty \ \forall i \in J_l, \ \forall l = 1, \ldots p, \tag{2.8}$$

and

$$Angle(\hat{u}_i, u_i) \xrightarrow{p} \frac{\pi}{2} \ as \ d \to \infty \ \forall i = \kappa + 1, \ldots, n. \tag{2.9}$$

Remark 2.3. If the cardinality of J_l, k_l, is 1, then (2.8) implies $\hat{u}_i$ is consistent for $i \in J_l$.

Remark 2.4. The strongly inconsistent eigenvectors whose labels are in J_{p+1} can be considered to be subspace-consistent. Let Γ_d be the subspace spanned by the population eigenvectors whose labels are in J_{p+1} for each d, i.e. $\Gamma_d = span\{u_j : j \in J_{p+1}\} = span\{u_{\kappa+1}, \ldots, u_d\}$. Then

$$\text{Angle}(\hat{u}_{i,d}, \Gamma_d) \xrightarrow{p} 0 \text{ as } d \to \infty,$$

for all $i \in J_{p+1}$.

Note that the formulation of the theorem is similar to the spiked covariance model but much more general. The uniform assumption on the underlying eigenvalues, i.e. $\lambda_i = 1$ for all $i > \kappa$, is relaxed to the ϵ-condition. We also have catalogued a large collection of specific results according to the various sizes of spikes.

These results are now illustrated for some classes of covariance matrices that are of special interest. These covariance matrices are easily represented in *factor form*, i.e. in terms of $F_d = \Sigma_d^{\frac{1}{2}}$.

Example 2.3. Consider a series of covariance matrices $\{\Sigma_d\}_d$. Let $\Sigma_d = F_d F_d'$, where F_d is a $d \times d$ symmetric matrix such that

$$F_d = (1-\rho_d)I_d + \rho_d J_d = \begin{pmatrix} 1 & \rho_d & \cdots & \rho_d \\ \rho_d & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_d \\ \rho_d & \cdots & \rho_d & 1 \end{pmatrix},$$

where J_d is the $d \times d$ matrix of ones and $\rho_d \in (0,1)$ depends on d. The eigenvalues of Σ_d are $\lambda_{1,d} = (d\rho_d + 1 - \rho_d)^2, \lambda_{2,d} = \cdots = \lambda_{d,d} = (1-\rho_d)^2$. The first eigenvector is $u_1 = \frac{1}{\sqrt{d}}(1, 1, \ldots, 1)'$, while $\{u_2, \ldots, u_d\}$ are any orthogonal sets of direction vectors perpendicular to u_1. Note that $\sum_{i=2}^{d} \lambda_{i,d} = d(1-\rho_d)^2 = O(d)$ and the ϵ_2-condition holds. Let $X_d \sim \mathcal{N}_d(0, \Sigma_d)$. By Theorem 2.5, if $\rho_d \in (0,1)$ is a fixed constant or decreases to 0 slowly so that $\rho_d \gg d^{-\frac{1}{2}}$, then the first PC direction $\hat{u}_1$ is consistent. Else if ρ_d decreases to 0 so quickly that $\rho_d \ll d^{-\frac{1}{2}}$, then $\hat{u}_1$ is strongly inconsistent. In both cases all the other sample PC directions are strongly inconsistent.

Example 2.4. Consider now a $2d \times 2d$ covariance matrix $\Sigma_d = F_d F_d'$, where F_d is a block diagonal matrix such that

$$F_d = \begin{pmatrix} F_{1,d} & O \\ O & F_{2,d} \end{pmatrix},$$

where $F_{1,d} = (1-\rho_{1,d})I_d + \rho_{1,d}J_d$ and $F_{2,d} = (1-\rho_{2,d})I_d + \rho_{2,d}J_d$. Suppose $0 < \rho_{2,d} \leqslant \rho_{1,d} < 1$. Note that $\lambda_{1,d} = (d\rho_{1,d} + 1 - \rho_{1,d})^2$, $\lambda_{2,d} = (d\rho_{2,d} + 1 - \rho_{2,d})^2$ and the ϵ_3-condition holds. Let $X_{2d} \sim \mathcal{N}_{2d}(0, \Sigma_d)$. Application of Theorem 2.5 for various conditions on $\rho_{1,d}$, $\rho_{2,d}$ is summarized as follows. Denote, for two non-increasing sequences μ_d, $\nu_d \in (0,1)$, $\mu_d \gg \nu_d$ for $\nu_d = o(\mu_d)$ and $\mu_d \succeq \nu_d$ for $\lim_{d\to\infty} \frac{\mu_d}{\nu_d} = c \in [1, \infty)$.

1. $\rho_{1,d} \gg \rho_{2,d} \gg d^{-\frac{1}{2}}$: Both $\hat{u}_1$,$\hat{u}_2$ consistent.
2. $\rho_{1,d} \succeq \rho_{2,d} \gg d^{-\frac{1}{2}}$: Both $\hat{u}_1$,$\hat{u}_2$ subspace-consistent to span$\{u_1, u_2\}$.
3. $\rho_{1,d} \gg d^{-\frac{1}{2}} \gg \rho_{2,d}$: $\hat{u}_1$ consistent, $\hat{u}_2$ strongly inconsistent.
4. $d^{-\frac{1}{2}} \gg \rho_{1,d} \gg \rho_{2,d}$: Both $\hat{u}_1$,$\hat{u}_2$ strongly inconsistent.

2.4.4 Corollaries to the main theorem

The result can be extended for special cases.

First of all, consider constructing $X_{(d)}$ from Z_d by $X_{(d)} \equiv U_d \Lambda_d^{\frac{1}{2}} Z_d$ where Z_d is a truncated set from an infinite sequence of independent random variables with mean zero and variance 1. This assumption makes it possible to have convergence in the almost sure sense. This is mainly because the triangular array $\{Z_{1i,(d)}\}_{i,d}$ becomes the single sequence $\{Z_{1i}\}_i$.

Corollary 2.6. *Suppose all the assumptions in Theorem 2.5, with the assumption (a) replaced by the following:*

(a′) The components of $Z_{(d)} = \Lambda_d^{-\frac{1}{2}} U_d' X_{(d)}$ have uniformly bounded eighth moments and are independent to each other. Let $Z_{1i,(d)} \equiv Z_{1i}$ for all i, d.

If the strong $\epsilon_{\kappa+1}$-condition (2.4) holds, then the mode of convergence of (2.8) and (2.9) is almost sure.

Second, consider the case that both d, n tend to infinity. Under the setting of Theorem 2.5, we can separate PC directions better when the eigenvalues are distinct. When $d \to \infty$, we have subspace consistency of $\hat{u}_i$ with the proper subspace, which includes u_i. Now letting $n \to \infty$ makes it possible for $\hat{u}_i$ to be consistent.

Corollary 2.7. *Let Σ_d, $X_{(d)}$, and S_d be as before. Under the assumptions (a), (b) and (c) in Theorem 2.5, assume further for (b) that the first κ eigenvalues are distinct, i.e. $c_i > c_j$ for $i > j$ and $i, j \in J_l$ for $l = 1, \ldots, p$. Then for all $i \leqslant \kappa$,*

$$Angle(\hat{u}_i, u_i) \xrightarrow{p} 0 \text{ as } d \to \infty,\ n \to \infty, \tag{2.10}$$

where the limits are applied successively.

If the assumption (a) is replaced by the assumption (a′) of Corollary 2.6, then the mode of convergence of (2.10) is almost sure.

This corollary can be viewed as the case when d, n tend to infinity together, but d increases at a much faster rate than n, i.e. $d \gg n$. When n also increases in the particular setting of the corollary, the sample eigenvectors, which were only subspace-consistent in the $d \to \infty$ case, tend to be distinguishable and each of the eigenvectors is consistent. We conjecture that the inconsistent sample eigenvalues are still strongly inconsistent when $d, n \to \infty$ and $d \gg n$.

2.4.5 Limiting distributions of corresponding eigenvalues

The study of asymptotic behavior of the sample eigenvalues is an important part in the proof of Theorem 2.5, and also could be of independent interest. The following lemma states that the large sample eigenvalues increase at the same speed as their population counterpart and the relatively small eigenvalues tend to be of order of d as d tends to infinity. Let $\varphi_i(A)$ denote the ith largest eigenvalue of the symmetric matrix A and $\varphi_{i,l}(A) = \varphi_{i^*}(A)$ where $i^* = i - \sum_{j=1}^{l-1} k_j$.

Lemma 2.8. *If the assumptions of Theorem 2.5 hold, and let Z_l be a $k_l \times n$ matrix from blocks of Z as defined in (2.12), then*

$$\begin{aligned} \hat{\lambda}_i / d^{\alpha_l} &\Longrightarrow \eta_i \qquad \text{as } d \to \infty \text{ if } i \in J_l,\ \forall l = 1, \ldots, p, \\ \hat{\lambda}_i / d &\xrightarrow{p} K \qquad \text{as } d \to \infty \text{ if } i = \kappa + 1, \ldots, n, \end{aligned}$$

where each η_i *is a random variable whose support is* $(0, \infty)$ *almost surely and indeed* $\eta_i = \varphi_{i,l}(n^{-1} C_l^{\frac{1}{2}} Z_l Z_l' C^{\frac{1}{2}})$ *for each* $i \in J_l$*, where* $C_l = diag\{c_j : j \in J_l\}$ *and* $K = \lim_{d \to \infty} (dn)^{-1} \sum_{i \in J_{p+1}} \lambda_{i,d}$*.*

If the data matrix $X_{(d)}$ is Gaussian, then the first κ sample eigenvalues converge in distribution to some quantities, which have known distributions.

Corollary 2.9. *Under all the assumptions of Theorem 2.5, assume further that* $X_{(d)} \sim \mathcal{N}_d(0, \Sigma_d)$ *for each* d*. Then, for* $i \in J_l$*,* $l = 1, \ldots p$

$$\frac{\hat{\lambda}_i}{d^{\alpha_l}} \Longrightarrow \varphi_{i,l}(n^{-1} \mathcal{W}_{k_l}(n, C_l)) \text{ as } d \to \infty,$$

where $\mathcal{W}_{k_l}(n, C_l)$ *denotes a* $k_l \times k_l$ *random matrix distributed as the Wishart distribution with degree of freedom* n *and covariance* C_l*.*

If $k_l = 1$ *for some* l*, then for* $i \in J_l$

$$\frac{\hat{\lambda}_i}{\lambda_i} \Longrightarrow \frac{\chi_n^2}{n} \text{ as } d \to \infty,$$

where χ_n^2 *denotes a random variable distributed as the* χ^2 *distribution with degree of freedom* n*.*

This generalizes the results in section 4.2 of Ahn et al. (2007).

2.5 Proofs

This section contains the proofs of theorems in this chapter.

Proof of Theorem 2.1. First we give the proof of part (1). By (2.1), the mth diagonal entry of nS_D can be expressed as $\sum_{i=1}^{d} \lambda_{i,d} z_{im,d}^2$ where $z_{im,d}$ is the (i, m)th entry of the matrix $Z_{(d)}$. Define the relative eigenvalues $\tilde{\lambda}_{i,d}$ as $\tilde{\lambda}_{i,d} \equiv \frac{\lambda_{i,d}}{\sum_{i=1}^{d} \lambda_{i,d}}$. Let π_d denote the given permutation for each d and let $Y_i = z_{\pi_d(i)m,d}^2 - 1$. Then the Y_i's are ρ-mixing, $\mathrm{E}(Y_i) = 0$ and $\mathrm{E}(Y_i^2) \leqslant B$ for all i for some $B < \infty$. Let $\rho(m) = \sup |\mathrm{corr}(Y_i, Y_{i+m})|$ where the sup is over all i. We shall use the following lemma.

Lemma 2.10. *For any permutation* π_d^*,

$$\lim_{d\to\infty}\sum_{i=1}^{d}\tilde{\lambda}_{\pi_d^*(i),d}\rho(i)=0.$$

Proof. For any $\delta>0$, since $\lim_{i\to\infty}\rho(i)=0$, we can choose N such that $\rho(i)<\frac{\delta}{2}$ for all $i>N$. Since $\lim_{d\to\infty}\sum_{i=1}^{d}\tilde{\lambda}^2_{\pi_d^*(i),d}=0$, we get $\lim_{d\to\infty}\sum_{i=1}^{N}\tilde{\lambda}_{\pi_d^*(i),d}=0$. Thus we can choose d_0 satisfying $\sum_{i=1}^{N}\tilde{\lambda}_{\pi_d^*(i),d}<\frac{\delta}{2}$ for all $d>d_0$. With the fact $\sum_{i=1}^{d}\tilde{\lambda}_{i,d}=1$ for all d and $\rho(i)<1$, we get for all $d>d_0$,

$$\sum_{i=1}^{d}\tilde{\lambda}_{\pi_d^*(i),d}\rho(i)=\sum_{i=1}^{N}\tilde{\lambda}_{\pi_d^*(i),d}\rho(i)+\sum_{i=N+1}^{d}\tilde{\lambda}_{\pi_d^*(i),d}\rho(i)<\delta. \qquad \square$$

Now let π_d^{-1} be the inverse permutation of π_d. Then by Lemma 2.10 and the ϵ-condition, there exists a permutation π_d^* such that

$$\begin{aligned}\mathrm{E}(\sum_{i=1}^{d}\tilde{\lambda}_{\pi_d^{-1}(i),d}Y_i)^2 &= \sum_{i=1}^{d}\tilde{\lambda}^2_{\pi_d^{-1}(i),d}\mathrm{E}Y_i^2+2\sum_{i=1}^{d}\tilde{\lambda}_{\pi_d^{-1}(i),d}\sum_{j=i+1}^{d}\tilde{\lambda}_{\pi_d^{-1}(j),d}\mathrm{E}Y_iY_j\\ &\leqslant \sum_{i=1}^{d}\tilde{\lambda}^2_{i,d}B+2\sum_{i=1}^{d}\tilde{\lambda}_{i,d}\sum_{j=1}^{d}\tilde{\lambda}_{\pi_d^*(j),d}\rho(j)B^2\to 0,\end{aligned}$$

as $d\to\infty$. Then Chebyshev's inequality gives us, for any $\tau>0$,

$$P\left[\left|\sum_{i=1}^{d}\tilde{\lambda}_{i,d}z_{im}^2-1\right|>\tau\right]\leqslant\frac{\mathrm{E}\left(\sum_{i=1}^{d}\tilde{\lambda}_{\pi_d^{-1}(i),d}Y_i\right)^2}{\tau^2}\to 0,$$

as $d\to\infty$. Thus we conclude that the diagonal elements of nS_D converge to 1 in probability.

The off-diagonal elements of nS_D can be expressed as $\sum_{i=1}^{d}\lambda_{i,d}z_{im}z_{il}$. Similar arguments to those used in the diagonal case, together with the fact that z_{im} and z_{il} are independent, gives that

$$\mathrm{E}(\sum_{i=1}^{d}\tilde{\lambda}_{i,d}z_{im}z_{il})^2\leqslant\sum_{i=1}^{d}\tilde{\lambda}^2_{i,d}+2\sum_{i=1}^{d}\tilde{\lambda}_{i,d}\sum_{j=i+1}^{d}\tilde{\lambda}_{\pi_d^{-1}(j),d}\rho^2(j-i)\to 0,$$

as $d\to\infty$. Thus by Chebyshev's inequality, the off-diagonal elements of nS_D converge to 0 in probability.

Now, we give the proof for part (2). We begin with the mth diagonal entry of nS_D, $\sum_{i=1}^{d} \lambda_{i,d} z_{im}^2$. Note that since $\sum_{i=1}^{k-1} \tilde{\lambda}_{i,d} \to 0$ by the ϵ-condition, we assume $k = 1$ in (2.4) without loss of generality.

Let $Y_i = z_{im}^2 - 1$. Note that the Y_i's are independent, $\mathrm{E}(Y_i) = 0$ and $\mathrm{E}(Y_i^4) \leqslant B$ for all i for some $B < \infty$. Now

$$\mathrm{E}\left(\sum_{i=1}^{d} \tilde{\lambda}_{i,d} Y_i\right)^4 = \mathrm{E} \sum_{i,j,k,l=1}^{d} \tilde{\lambda}_{i,d}\tilde{\lambda}_{j,d}\tilde{\lambda}_{k,d}\tilde{\lambda}_{l,d} Y_i Y_j Y_k Y_l. \tag{2.11}$$

Note that terms in the sum of the form $\mathrm{E}Y_iY_jY_kY_l$, $\mathrm{E}Y_i^2Y_jY_k$, and $\mathrm{E}Y_i^3Y_j$ are 0 if i, j, k, l are distinct. The only terms that do not vanish are those of the form $\mathrm{E}Y_i^4$, $\mathrm{E}Y_i^2Y_j^2$, both of which are bounded by B. Note that $\tilde{\lambda}_{i,d}^2$'s are non-negative and hence the sum of squares is less than the square of sum, we have $\sum_{i=1}^{d} \tilde{\lambda}_{i,d}^4 \leqslant (\sum_{i=1}^{d} \tilde{\lambda}_{i,d}^2)^2$. Also note that by the strong ϵ-condition, $\sum_{i=1}^{d} \tilde{\lambda}_{i,d}^2 = (d\epsilon)^{-1} = o(d^{-\frac{1}{2}})$. Thus (2.11) is bounded as

$$\begin{aligned}
\mathrm{E}\left(\sum_{i=1}^{d} \tilde{\lambda}_{i,d} Y_i\right)^4 &\leqslant \sum_{i=1}^{d} \tilde{\lambda}_{i,d}^4 B + \sum_{i=j\neq k=l} \tilde{\lambda}_{i,d}^2 \tilde{\lambda}_{k,d}^2 B \\
&\leqslant (\sum_{i=1}^{d} \tilde{\lambda}_{i,d}^2)^2 B + \binom{4}{2} (\sum_{i=1}^{d} \tilde{\lambda}_{i,d}^2)^2 B \\
&= o(d^{-1}).
\end{aligned}$$

Then Chebyshev's inequality gives us, for any $\tau > 0$,

$$P\left[\left|\sum_{i=1}^{d} \tilde{\lambda}_{i,d} z_{im}^2 - 1\right| > \tau\right] \leqslant \frac{\mathrm{E}\left(\sum_{i=1}^{d} \tilde{\lambda}_{i,d} Y_i\right)^4}{\tau^4} \leqslant \frac{o(d^{-1})}{\tau^4}.$$

Summing over d gives $\sum_{d=1}^{\infty} P\left[\left|\sum_{i=1}^{d} \tilde{\lambda}_{i,d} z_{im}^2 - 1\right| > \tau\right] < \infty$ and by Borel-Cantelli Lemma, we conclude that a diagonal element $\sum_{i=1}^{d} \tilde{\lambda}_{i,d} z_{ij}^2$ converges to 1 almost surely.

The off-diagonal elements of nS_D can be expressed as $\sum_{i=1}^{d} \lambda_{i,d} z_{im} z_{il}$. Using similar arguments to those used in the diagonal case, we have

$$P\left[\left|\sum_{i=1}^{d}\tilde{\lambda}_{i,d}z_{im}z_{il}\right|>\tau\right]\leqslant\frac{\mathrm{E}\left(\sum_{i=1}^{d}\tilde{\lambda}_{i,d}z_{im}z_{il}\right)^{4}}{\tau^{4}}\leqslant\frac{o(d^{-1})}{\tau^{4}},$$

and again by the Borel-Cantelli Lemma, the off-diagonal elements converge to 0 almost surely.

□ □

The proof of Theorem 2.5 is divided in two parts. Since eigenvectors are associated to eigenvalues, at first, we focus on asymptotic behavior of sample eigenvalues (proof of Lemma 2.8) and then investigate consistency or strong inconsistency of sample eigenvectors (proof of Theorem 2.5).

Proof of Lemma 2.8. We summarize a few definitions and lemmas that are useful to prove this lemma. Let $\mathcal{S}_m$ be the set of all $m \times m$ real symmetric matrices. Let $\varphi(A)$ be a vector of eigenvalues of A for $A \in \mathcal{S}_m$ arranged in non-increasing order and let $\varphi_i(A)$ be the ith largest eigenvalue of A. Let $\|\bullet\|_2$ be the usual 2-norm of vectors, and $\|\bullet\|_F$ be the Frobenius norm of matrices defined by $\|A\|_F = (\sum_{i,j} A_{ij}^2)^{1/2}$.

Lemma 2.11 (Wielandt-Hoffman inequality)**.** *If A, $B \in \mathcal{S}_m$, then*

$$\|\varphi(A+B)-\varphi(A)\|_2 \leqslant \|\varphi(B)\|_2 = \|B\|_F.$$

This inequality is known as Wielandt-Hoffman inequality. See Wilkinson (1988) for detailed discussion and proof.

Corollary 2.12 (Continuity of eigenvalues)**.** *The mapping of eigenvalues $\varphi : \mathcal{S}_m \longmapsto \mathbf{R}_m$ is uniformly continuous.*

Proof. By Lemma 2.11, $\forall\epsilon > 0$, $\forall A, B \in \mathcal{S}_m$, ${}^{\exists}\delta = \epsilon$ such that $\|A - B\|_F \leqslant \delta$, then

$$\|\varphi(A)-\varphi(B)\|_2 \leqslant \|\varphi(A-B)\|_2 \leqslant \delta = \epsilon.$$ □

The proof relies heavily on the following lemma.

Lemma 2.13 (Weyl's inequality)**.** *If A, B are $m \times m$ real symmetric matrices, then for all $k = 1, \dots, m$,*

$$\left.\begin{array}{c} \varphi_k(A) + \varphi_m(B) \\ \varphi_{k+1}(A) + \varphi_{m-1}(B) \\ \vdots \\ \varphi_m(A) + \varphi_k(B) \end{array}\right\} \leqslant \varphi_k(A+B) \leqslant \left\{\begin{array}{c} \varphi_k(A) + \varphi_1(B) \\ \varphi_{k-1}(A) + \varphi_2(B) \\ \vdots \\ \varphi_1(A) + \varphi_k(B) \end{array}\right.$$

This inequality is discussed in Rao (1973) and its use on asymptotic studies of eigenvalues of a random matrix appeared in Eaton and Tyler (1991).

Since S and its dual S_D share nonzero eigenvalues, one of the main ideas of the proof is working with S_D. By our decomposition (2.1), $nS_D = Z'\Lambda Z$. We also write Z and Λ as block matrices such that

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_{p+1} \end{pmatrix}, \quad \Lambda = \begin{pmatrix} \Lambda_1 & O & \cdots & O \\ O & \Lambda_2 & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & \Lambda_{p+1} \end{pmatrix}, \tag{2.12}$$

where Z_l is a $k_l \times n$ matrix for each $l = 1, \dots, p+1$ and $\Lambda_l (\equiv \Lambda_{l,d})$ is a $k_l \times k_l$ diagonal matrix for each $l = 1, \dots, p+1$ and O denotes a matrix where all elements are zeros. Now we can write

$$nS_D = Z'\Lambda Z = \sum_{l=1}^{p+1} Z_l' \Lambda_l Z_l. \tag{2.13}$$

Note that Z_l depends on d. We will, however, simplify notation Z_l for representing for all $d = 1, \dots, \infty$.

Note that Theorem 2.1 implies that when the last term in equation (2.13) is divided by d, it converges to an identity matrix, namely,

$$d^{-1} Z_{p+1}' \Lambda_{p+1} Z_{p+1} \xrightarrow{p} nK \cdot I_n, \tag{2.14}$$

where $K \in (0,\infty)$ is such that $(dn)^{-1}\sum_{i \in J_{p+1}} \lambda_{i,d} \to K$. Moreover dividing by d^{α_1} gives us

$$nd^{-\alpha_1} S_D = d^{-\alpha_1} Z_1' \Lambda_1 Z_1 + d^{-\alpha_1} \sum_{l=2}^{p} Z_l' \Lambda_l Z_l + d^{1-\alpha_1} d^{-1} Z_{p+1}' \Lambda_{p+1} Z_{p+1}.$$

By the assumption (b), the first term on the right hand side converges to $Z_1' C_1 Z_1$ where C_1 is the $k_1 \times k_1$ diagonal matrix such that $C_1 = \mathrm{diag}\{c_j; j \in J_1\}$ and the other terms tend to a zero matrix. Thus, we get

$$nd^{-\alpha_1} S_D \Longrightarrow Z_1' C_1 Z_1 \text{ as } d \to \infty.$$

Note that the non-zero eigenvalues of $Z_1' C_1 Z_1$ are the same as the nonzero eigenvalues of $C_1^{\frac{1}{2}} Z_1 Z_1' C_1^{\frac{1}{2}}$ which is a $k_1 \times k_1$ random matrix with full rank almost surely. Since eigenvalues are continuous, we have for $i \in J_1$,

$$\begin{aligned} \varphi_i(nd^{-\alpha_1} S_D) &\Longrightarrow \varphi_i(Z_1' C_1 Z_1) \text{ as } d \to \infty \\ &= \varphi_i(C_1^{\frac{1}{2}} Z_1 Z_1' C_1^{\frac{1}{2}}). \end{aligned}$$

Thus, we conclude that for the sample eigenvalues in the group J_1, $\hat{\lambda}_i / d^{\alpha_1} = \varphi_i(d^{-\alpha_1} S_D)$ converges in distribution to $\varphi_i(n^{-1} C_1^{\frac{1}{2}} Z_1 Z_1' C_1^{\frac{1}{2}})$ for $i \in J_1$.

Let us focus on eigenvalues whose indices are in the group $J_2, \ldots, J_p$. Suppose we have $\hat{\lambda}_i = O_p(d^{\alpha_j})$ for all $i \in J_j$, for $j = 1, \ldots, l-1$. Pick any $i \in J_l$. We will provide upper and lower bounds on $\hat{\lambda}_i$ by Weyl's inequality (Lemma 2.13). Dividing both sides of (2.13) by d^{α_l}, we get

$$nd^{-\alpha_l} S_D = d^{-\alpha_l} \sum_{j=1}^{l-1} Z_j' \Lambda_j Z_j + d^{-\alpha_l} \sum_{j=l}^{p+1} Z_j' \Lambda_j Z_j$$

and apply Weyl's inequality for the upper bound,

$$\begin{aligned} \varphi_i(nd^{-\alpha_l} S_D) &\leqslant \varphi_{1+\sum_{j=1}^{l-1} k_j}(d^{-\alpha_l} \sum_{j=1}^{l-1} Z_j' \Lambda_j Z_j) + \varphi_{i-\sum_{j=1}^{l-1} k_j}(d^{-\alpha_l} \sum_{j=l}^{p+1} Z_j' \Lambda_j Z_j) \\ &= \varphi_{i-\sum_{j=1}^{l-1} k_j}(d^{-\alpha_l} \sum_{j=l}^{p+1} Z_j' \Lambda_j Z_j). \end{aligned} \tag{2.15}$$

Note that the first term vanishes since the rank of $d^{-\alpha_l}\sum_{j=1}^{l-1} Z_j'\Lambda_j Z_j$ is at most $\sum_{j=1}^{l-1} k_j$. Also note that the matrix in the upper bound (2.15) converges to a simple form

$$\begin{aligned} d^{-\alpha_l}\sum_{j=l}^{p+1} Z_j'\Lambda_j Z_j &= d^{-\alpha_l} Z_l'\Lambda_l Z_l + d^{-\alpha_l}\sum_{j=l+1}^{p+1} Z_j'\Lambda_j Z_j \\ &\Longrightarrow Z_l' C_l Z_l \text{ as } d \to \infty, \end{aligned}$$

where C_l is the $k_l \times k_l$ diagonal matrix such that $C_l = \operatorname{diag}\{c_j; j \in J_l\}$.

In order to have a lower bound of $\hat{\lambda}_i$, Weyl's inequality is applied to the expression

$$d^{-\alpha_l}\sum_{j=1}^{l} Z_j'\Lambda_j Z_j + d^{-\alpha_l}\sum_{j=l+1}^{p+1} Z_j'\Lambda_j Z_j = n d^{-\alpha_l} S_D,$$

so that

$$\varphi_i(d^{-\alpha_l}\sum_{j=1}^{l} Z_j'\Lambda_j Z_j) + \varphi_n(d^{-\alpha_l}\sum_{j=l+1}^{p+1} Z_j'\Lambda_j Z_j) \leqslant \varphi_i(n d^{-\alpha_l} S_D). \tag{2.16}$$

It turns out that the first term of the left hand side is not easy to manage, so we again use Weyl's inequality to get

$$\varphi_{\sum_{j=1}^{l} k_j}(d^{-\alpha_l}\sum_{j=1}^{l-1} Z_j'\Lambda_j Z_j) \leqslant \varphi_i(d^{-\alpha_l}\sum_{j=1}^{l} Z_j'\Lambda_j Z_j) + \varphi_{1-i+\sum_{j=1}^{l-1} k_j}(-d^{-\alpha_l} Z_l'\Lambda_l Z_l), \tag{2.17}$$

where the left hand side is 0 since the rank of the matrix inside is at most $\sum_{j=1}^{l-1} k_j$. Note that since $d^{-\alpha_l} Z_l'\Lambda_l Z_l$ and $d^{-\alpha_l}\Lambda_l^{\frac{1}{2}} Z_l Z_l' \Lambda_l^{\frac{1}{2}}$ share non-zero eigenvalues, we get

$$\begin{aligned} \varphi_{1-i+\sum_{j=1}^{l} k_j}(-d^{-\alpha_l} Z_l'\Lambda_l Z_l) &= \varphi_{1-i+\sum_{j=1}^{l} k_j}(-d^{-\alpha_l}\Lambda_l^{\frac{1}{2}} Z_l Z_l'\Lambda_l^{\frac{1}{2}}) \\ &= \varphi_{k_l-i+1+\sum_{j=1}^{l-1} k_j}(-d^{-\alpha_l}\Lambda_l^{\frac{1}{2}} Z_l Z_l'\Lambda_l^{\frac{1}{2}}) \\ &= -\varphi_{i-\sum_{j=1}^{l-1} k_j}(d^{-\alpha_l}\Lambda_l^{\frac{1}{2}} Z_l Z_l'\Lambda_l^{\frac{1}{2}}) \\ &= -\varphi_{i-\sum_{j=1}^{l-1} k_j}(d^{-\alpha_l} Z_l'\Lambda_l Z_l). \end{aligned} \tag{2.18}$$

Here we use the fact that for any $m \times m$ real symmetric matrix A, $\varphi_i(A) = -\varphi_{m-i+1}(-A)$ for all $i = 1, \ldots, m$.

Combining (2.16), (2.17), and (2.18) gives the lower bound

$$\varphi_{i-\sum_{j=1}^{l-1}k_j}(d^{-\alpha_l}Z_l'\Lambda_l Z_l)+\varphi_n(d^{-\alpha_l}\sum_{j=l+1}^{p+1}Z_j'\Lambda_j Z_j)\leqslant \varphi_i(nd^{-\alpha_l}S_D). \tag{2.19}$$

Note that the matrix inside of the first term of the lower bound (2.19) converges to $Z_l'C_lZ_l$ in distribution. The second term converges to 0 since the matrix inside converges to a zero matrix.

The difference between the upper and lower bounds of $\varphi_i(nd^{-\alpha_l}S_D)$ converges to 0 since

$$\varphi_{i-\sum_{j=1}^{l-1}k_j}(d^{-\alpha_l}\sum_{j=l}^{p+1}Z_j'\Lambda_j Z_j)-\varphi_{i-\sum_{j=1}^{l-1}k_j}(d^{-\alpha_l}Z_l'\Lambda_l Z_l)\to 0,$$

as $d\to\infty$. This is because φ is a continuous function and the difference between the two matrices converges to zero matrix. Therefore $\varphi_i(nd^{-\alpha_l}S_D)$ converges to the upper or lower bound as $d\to\infty$.

Now since both upper and lower bound of $\varphi_i(nd^{-\alpha_l}S_D)$ converge in distribution to same quantity, we have

$$\begin{aligned}\varphi_i(nd^{-\alpha_l}S_D) &\implies \varphi_{i-\sum_{j=1}^{l-1}k_j}(Z_l'C_lZ_l) \text{ as } d\to\infty.\\ &= \varphi_{i-\sum_{j=1}^{l-1}k_j}(C_l^{\frac{1}{2}}Z_lZ_l'C_l^{\frac{1}{2}}). \end{aligned}\tag{2.20}$$

Thus, by induction, we have the scaled ith sample eigenvalue $\hat{\lambda}_i/d^{\alpha_l}$ converges in distribution to $\varphi_{i-\sum_{j=1}^{l-1}k_j}(n^{-1}C_l^{\frac{1}{2}}Z_lZ_l'C_l^{\frac{1}{2}})$ for $i\in J_l$, $l=1,\ldots,p$ as desired.

Now let us focus on the rest of the sample eigenvalues $\hat{\lambda}_i$, $i=\kappa+1,\ldots,n$. For any i, again by Weyl's upper bound inequality we get

$$\begin{aligned}\varphi_i(nd^{-1}S_D) &\leqslant \varphi_{i-\kappa}(d^{-1}Z_{p+1}'\Lambda_{p+1}Z_{p+1})+\varphi_{\kappa+1}(d^{-1}\sum_{j=1}^{p}Z_j'\Lambda_jZ_j)\\ &= \varphi_{i-\kappa}(d^{-1}Z_{p+1}'\Lambda_{p+1}Z_{p+1}),\end{aligned}$$

where the second term on the right hand side vanishes since the matrix inside is of rank at most κ. Also for lower bound, we have

$$\begin{aligned}\varphi_i(nd^{-1}S_D) &\geqslant \varphi_i(d^{-1}Z'_{p+1}\Lambda_{p+1}Z_{p+1}) + \varphi_n(d^{-1}\sum_{j=1}^{p} Z'_j\Lambda_j Z_j) \\ &= \varphi_i(d^{-1}Z'_{p+1}\Lambda_{p+1}Z_{p+1}),\end{aligned}$$

where the second term vanishes since $\kappa < n$. Thus we have complete bounds for $\varphi_i(nd^{-1}S_D)$ such that

$$\varphi_i(d^{-1}Z'_{p+1}\Lambda_{p+1}Z_{p+1}) \leqslant \varphi_i(nd^{-1}S_D) \leqslant \varphi_{i-\kappa}(d^{-1}Z'_{p+1}\Lambda_{p+1}Z_{p+1}),$$

for all $i = \kappa+1, \ldots, n$. However, by (2.14), the matrix in both bounds converges to $nK \cdot I_n$ in probability. Thus lower and upper bounds of $\varphi_i(d^{-1}S_D)$ converge to K in probability for $i = \kappa+1, \ldots, n$, which completes the proof. □

Proof of Theorem 2.5. We begin by defining a standardized version of the sample covariance matrix, not to be confused with the dual S_D, as

$$\begin{aligned}\tilde{S} &= \Lambda^{-\frac{1}{2}}U'SU\Lambda^{-\frac{1}{2}} \\ &= \Lambda^{-\frac{1}{2}}U'(\hat{U}\hat{\Lambda}\hat{U}')U\Lambda^{-\frac{1}{2}} \\ &= \Lambda^{-\frac{1}{2}}P\hat{\Lambda}P'\Lambda^{-\frac{1}{2}}, \end{aligned} \tag{2.21}$$

where $P = U'\hat{U} = \{u'_i\hat{u}_j\}_{ij} \equiv \{p_{ij}\}_{ij}$. Note that elements of P are inner products between population eigenvectors and sample eigenvectors. Since $\tilde{S}$ is standardized, we have by $S = n^{-1}XX'$ and $X = U\Lambda^{\frac{1}{2}}Z$,

$$\tilde{S} = n^{-1}ZZ'. \tag{2.22}$$

Note that the angle between two directions can be formulated as an inner product of the two direction vectors. Thus we will investigate the behavior of the inner product matrix P as $d \to \infty$, by showing that

$$\sum_{j\in J_l} p_{ji}^2 \xrightarrow{p} 1 \text{ as } d \to \infty, \tag{2.23}$$

for all $i \in J_l$, $l = 1, \ldots, p$, and

$$p_{ii}^2 \xrightarrow{p} 0 \text{ as } d \to \infty, \tag{2.24}$$

for all $i = \kappa + 1, \ldots, n$.

Suppose for now we have the result of (2.23) and (2.24). Then for any $i \in J_l$, $l = 1, \ldots, p$,

$$\begin{aligned}
\text{Angle}(\hat{u}_i, \text{span}\{u_j : j \in J_l\}) &= \arccos\left(\frac{\hat{u}_i'[\text{Proj}_{\text{span}\{u_j : j \in J_l\}}\hat{u}_i]}{\|\hat{u}_i\|_2 \cdot \|[\text{Proj}_{\text{span}\{u_j : j \in J_l\}}\hat{u}_i]\|_2}\right) \\
&= \arccos\left(\frac{\hat{u}_i'(\sum_{j \in J_l}(u_j'\hat{u}_i)u_j)}{\|\hat{u}_i\|_2 \cdot \|\sum_{j \in J_l}(u_j'\hat{u}_i)u_j\|_2}\right) \\
&= \arccos\left(\frac{\sum_{j \in J_l}(u_j'\hat{u}_i)^2}{1 \cdot (\sum_{j \in J_l}(u_j'\hat{u}_i)^2)^{\frac{1}{2}}}\right) \\
&= \arccos\left((\sum_{j \in J_l} p_{ji}^2)^{\frac{1}{2}}\right) \\
&\xrightarrow{p} 0 \text{ as } d \to \infty,
\end{aligned}$$

by (2.23) and for $i = \kappa + 1, \ldots, n$,

$$\begin{aligned}
\text{Angle}(\hat{u}_i, u_i) &= \arccos(|u_i'\hat{u}_i|) \\
&= \arccos(|p_{ii}|) \\
&\xrightarrow{p} \frac{\pi}{2} \text{ as } d \to \infty,
\end{aligned}$$

by (2.24), as desired.

Therefore, it is enough to show (2.23) and (2.24). We begin with taking jth diagonal entry of $\tilde{S}$, $\tilde{s}_{jj}$, from (2.21) and (2.22),

$$\tilde{s}_{jj} = \lambda_j^{-1} \sum_{i=1}^{n} \hat{\lambda}_i p_{ji}^2 = n^{-1} z_j z_j',$$

where z_j denotes the jth row vector of Z. Since

$$\lambda_j^{-1} \hat{\lambda}_i p_{ji}^2 \leqslant n^{-1} z_j z_j', \tag{2.25}$$

we have at most

$$p_{ji}^2 = O_p(\frac{\lambda_j}{\hat{\lambda}_i}),$$

for all $i = 1, \ldots, n$, $j = 1, \ldots, d$. Note that by Lemma 2.8, we have for $i \in J_{l_1}$, $j \in J_{l_2}$ where $1 \leqslant l_1 < l_2 \leqslant p + 1$,

$$p_{ji}^2 = O_p(\frac{\lambda_j}{\hat{\lambda}_i}) = \begin{cases} O_p(d^{\alpha_{l_2} - \alpha_{l_1}}), & \text{if } l_2 \leqslant p, \\ O_p(d^{1-\alpha_{l_1}}), & \text{if } l_2 = p + 1, \end{cases} \tag{2.26}$$

so that $p_{ji}^2 \xrightarrow{p} 0$ as $d \to \infty$ in both cases.

Note that the inner product matrix P is also a unitary matrix. The norm of the ith column vector of P must be 1 for all d, i.e. $\sum_{j=1}^{d} p_{ji}^2 = 1$. Thus (2.23) is equivalent to $\sum_{j \in \{1,\ldots,d\}/J_l} p_{ji}^2 \xrightarrow{p} 0$ as $d \to \infty$.

Now for any $i \in J_1$,

$$\sum_{j \in \{1,\ldots,d\}/J_1} p_{ji}^2 = \sum_{j \in J_2 \cup \cdots \cup J_p} p_{ji}^2 + \sum_{j \in J_{p+1}} p_{ji}^2.$$

Since the first term on the right hand side is a finite sum of quantities converging to 0, it converges to 0 almost surely as d tends to infinity. By (2.25), we have an upper bound for the second term,

$$\begin{aligned} \sum_{j \in J_{p+1}} p_{ji}^2 &= \sum_{j \in J_{p+1}} \lambda_j^{-1} \hat{\lambda}_i p_{ji}^2 \frac{\lambda_j}{\hat{\lambda}_i} \\ &\leqslant \frac{\sum_{j \in J_{p+1}} n^{-1} z_j z_j' \lambda_j}{d} \frac{d}{\hat{\lambda}_i} = \frac{\sum_{k=1}^{n} \sum_{j=\kappa+1}^{d} z_{j,k}^2 \lambda_j}{nd} \frac{d}{\hat{\lambda}_i}, \end{aligned}$$

where the $z_{j,k}$'s are the entries of a row random vector z_j. Note that by applying Theorem 2.1 with $\Sigma_d = \text{diag}\{\lambda_{\kappa+1}, \ldots, \lambda_d\}$, we have $\sum_{j=\kappa+1}^{d} z_{j,k}^2 \lambda_j / d \xrightarrow{p} 1$ as $d \to \infty$. Also by Lemma 2.8, the upper bound converges to 0 in probability. Thus we get

$$\sum_{j \in \{1,\ldots,d\}/J_1} p_{ji}^2 \xrightarrow{p} 0 \text{ as } d \to \infty,$$

which is equivalent to

$$\sum_{j\in J_1} p_{ji}^2 \xrightarrow{p} 1 \text{ as } d\to\infty. \tag{2.27}$$

Let us focus on the group $J_2,\ldots,J_p$. For any $l=2,\ldots,p$, suppose we have $\sum_{j\in J_m} p_{ji}^2 \xrightarrow{p} 1$ as $d\to\infty$ for all $i\in J_m$, $m=1,\ldots,l-1$. Note that it implies that for any $j\in J_m$,, $m=1,\ldots,l-1$,

$$\sum_{i\in\{1,\ldots,d\}/J_m} p_{ji}^2 \xrightarrow{p} 0 \text{ as } d\to\infty, \tag{2.28}$$

since

$$\sum_{j\in J_m}\sum_{i\in\{1,\ldots,d\}/J_m} p_{ji}^2 = \sum_{j\in J_m}\sum_{i=1}^{d} p_{ji}^2 - \sum_{j\in J_m}\sum_{i\in J_m} p_{ji}^2 \xrightarrow{p} \sum_{j\in J_m} 1 - \sum_{i\in J_m} 1 = 0,$$

as $d\to\infty$.

Now pick $i\in J_l$. We have

$$\sum_{j\in\{1,\ldots,d\}/J_l} p_{ji}^2 = \sum_{j\in J_1\cup\cdots\cup J_{l-1}} p_{ji}^2 + \sum_{j\in J_{l+1}\cup\cdots\cup J_p} p_{ji}^2 + \sum_{j\in J_{p+1}} p_{ji}^2.$$

Note that the first term is bounded as

$$\sum_{j\in J_1\cup\cdots\cup J_{l-1}} p_{ji}^2 \leqslant \sum_{i\in J_l}\sum_{j\in J_1\cup\cdots\cup J_{l-1}} p_{ji}^2 \leqslant \sum_{m=1}^{l-1}\sum_{j\in J_m}\left(\sum_{i\in\{1,\ldots,d\}/J_m} p_{ji}^2\right) \xrightarrow{p} 0$$

by (2.28). The second term also converges to 0 by (2.26). The last term is also bounded as

$$\sum_{j\in J_{p+1}} p_{ji}^2 = \sum_{j\in J_{p+1}} \lambda_j^{-1}\hat{\lambda}_i p_{ji}^2 \frac{\lambda_j}{\hat{\lambda}_i} \leqslant \frac{\sum_{j\in J_{p+1}} n^{-1} z_j z_j' \lambda_j}{d}\frac{d}{\hat{\lambda}_i},$$

so that it also converges to 0 in probability. Thus, we have $\sum_{j\in\{1,\ldots,d\}/J_l} p_{ji}^2 \xrightarrow{p} 0$ as $d\to\infty$ which implies that

$$\sum_{j\in J_l} p_{ji}^2 \xrightarrow{p} 1 \text{ as } d\to\infty.$$

Thus, by induction, (2.23) is proved.

For $i=\kappa+1,\ldots,n$, We have $\lambda_i^{-1}\hat{\lambda}_i p_{ii}^2 \leqslant n^{-1} z_i z_i'$, and so

$$p_{ii}^2 \leqslant \hat{\lambda}_i^{-1}\lambda_i n^{-1} z_i z_i' = O_p(\hat{\lambda}_i^{-1}\lambda_i),$$

which implies (2.24) by the assumption (c) and Lemma 2.8, and the proof is completed. ☐

Proof of Corollary 2.6. The proof follows the same lines as the proof of Theorem 2.5, with convergence in probability replaced by almost sure convergence. ☐ ☐

Proof of Corollary 2.7. From the proof of Theorem 2.5, write the inner product matrix P of (2.21) as a block matrix such that

$$P = \begin{pmatrix} P_{11} & \cdots & P_{1p} & P_{1,p+1} \\ \vdots & \ddots & \vdots & \vdots \\ P_{p1} & \cdots & P_{pp} & P_{p,p+1} \\ P_{p+1,1} & \cdots & P_{p+1,p} & P_{p+1,p+1} \end{pmatrix},$$

where each P_{ij} is a $k_i \times k_j$ random matrix. In the proof of theorem 2.5 we have shown that P_{ii}, $i = 1, \ldots, p$, tends to be a unitary matrix and P_{ij}, $i \neq j$, tends to be a zero matrix as $d \to \infty$. Likewise, Λ and $\hat{\Lambda}$ can be blocked similarly as $\Lambda = \mathrm{diag}\{\Lambda_i : i = 1, \ldots, p+1\}$ and $\hat{\Lambda} = \mathrm{diag}\{\hat{\Lambda}_i : i = 1, \ldots, p+1\}$.

Now pick $l \in \{1, \ldots, p\}$. The lth block diagonal of $\tilde{S}$, $\tilde{S}_{ll}$, is expressed as $\tilde{S}_{ll} = \sum_{j=1}^{p+1} \Lambda_l^{-\frac{1}{2}} P_{lj} \hat{\Lambda}_l P'_{lj} \Lambda_l^{-\frac{1}{2}}$. Since $P_{ij} \to 0$, $i \neq j$, we get

$$\|\tilde{S}_{ll} - \Lambda_l^{-\frac{1}{2}} P_{ll} \hat{\Lambda}_l P'_{ll} \Lambda_l^{-\frac{1}{2}}\|_F \xrightarrow{p} 0$$

as $d \to \infty$.

Note that by (2.22), $\tilde{S}_{ll}$ can be replaced by $n^{-1} Z_l Z'_l$. We also have $d^{-\alpha_l} \Lambda_l \to C_l$ by the assumption (b) and $d^{-\alpha_l} \hat{\Lambda}_l \xrightarrow{p} \mathrm{diag}\{\varphi(n^{-1} C_l^{\frac{1}{2}} Z_l Z'_l C_l^{\frac{1}{2}})\}$ by (2.20). Thus we get

$$\|n^{-1} Z_l Z'_l - C_l^{-\frac{1}{2}} P_{ll} \mathrm{diag}\{\varphi(n^{-1} C_l^{\frac{1}{2}} Z_l Z'_l C_l^{\frac{1}{2}})\} P'_{ll} C_l^{-\frac{1}{2}}\|_F \xrightarrow{p} 0$$

as $d \to \infty$.

Also note that since $n^{-1} Z_l Z'_l \to I_{k_l}$ almost surely as $n \to \infty$, we get $n^{-1} C_l^{\frac{1}{2}} Z_l Z'_l C_l^{\frac{1}{2}} \to C_l$ and $\mathrm{diag}\{\varphi(n^{-1} C_l^{\frac{1}{2}} Z_l Z'_l C_l^{\frac{1}{2}})\} \to C_l$ almost surely as $n \to \infty$. Using the fact that the Frobenius norm is unitarily invariant and $\|AB\|_F \leqslant \|A\|_F \|B\|_F$ for any square matrices A and B, we get

$$\begin{aligned}\|P_{ll}'C_lP_{ll}-C_l\|_F &\leqslant \|P_{ll}'C_lP_{ll}-\mathrm{diag}\{\varphi(n^{-1}C_l^{\frac{1}{2}}Z_lZ_l'C_l^{\frac{1}{2}})\}\|_F+o_p(1) \qquad (2.29)\\ &=\|C_l-P_{ll}\mathrm{diag}\{\varphi(n^{-1}C_l^{\frac{1}{2}}Z_lZ_l'C_l^{\frac{1}{2}})\}P_{ll}'\|_F+o_p(1)\\ &\leqslant \|n^{-1}C_l^{\frac{1}{2}}Z_lZ_l'C_l^{\frac{1}{2}}-P_{ll}\mathrm{diag}\{\varphi(n^{-1}C_l^{\frac{1}{2}}Z_lZ_l'C_l^{\frac{1}{2}})\}P_{ll}'\|_F+o_p(1)\\ &\leqslant \|C_l^{\frac{1}{2}}\|_F^2\|n^{-1}Z_lZ_l'-C_l^{-\frac{1}{2}}P_{ll}\mathrm{diag}\{\varphi(n^{-1}C_l^{\frac{1}{2}}Z_lZ_l'C_l^{\frac{1}{2}})\}P_{ll}'C_l^{-\frac{1}{2}}\|_F+o_p(1)\\ &\xrightarrow{p} 0 \text{ as } d,n\to\infty.\end{aligned}$$

Note that in order to have (2.29), P_{ll} must converge to $\mathrm{diag}\{\pm 1,\pm 1,\ldots,\pm 1\}$ since diagonal entries of C_l are distinct and a spectral decomposition is unique up to sign changes. Let $l=1$ for simplicity. Now suppose for any $\delta>0$, $\lim_{d,n}P(p_{m1}^2>\delta)>0$ for $m=2,\ldots,k_1$. Then for any $m=2,\ldots,k_1$,

$$\|P_{11}'C_1P_{11}-C_1\|_F\geqslant\sum_{j=1}^{k_1}(c_1-c_j)p_{j1}^2\geqslant(c_1-c_m)p_{m1}^2,$$

which contradicts (2.29) since $c_1-c_m>0$. Thus $p_{m1}^2\xrightarrow{p}0$ for all $m=2,\ldots,k_l$ which implies $p_{11}^2\xrightarrow{p}1$ as $d,n\to\infty$. Now by induction, $p_{ii}^2\xrightarrow{p}1$ for all $i\in J_l, l=1,\ldots,p$. Therefore $\mathrm{Angle}(\hat{u}_i,u_i)=\arccos(|p_{ii}|)\xrightarrow{p}0$ as $d,n\to\infty$.

If the assumptions of Corollary 2.6 also hold, then every convergence in the proof is replaced by almost sure convergence, which completes the proof. □ □

Proof of Corollary 2.9. With Gaussian assumption, noticing $C_l^{\frac{1}{2}}Z_lZ_l'C_l^{\frac{1}{2}}\sim\mathcal{W}_{k_l}(n,C_l)$ gives the first result. When $k_l=1$, The assumption (b) and that $C_l^{\frac{1}{2}}Z_lZ_l'C_l^{\frac{1}{2}}\sim c_i\chi_n^2$ imply that

$$\frac{\hat{\lambda}_i}{\lambda_i}=\frac{\hat{\lambda}_i}{c_id^{\alpha_l}}\cdot\frac{c_id^{\alpha_l}}{\lambda_i}\Longrightarrow\frac{\chi_n^2}{n}\text{ as }d\to\infty.\quad\square$$

□

Chapter 3

Boundary Behavior in HDLSS asymptotics of PCA

The work presented here is based on a paper Jung, Sen and Marron (2011c).

3.1 Introduction

In this chapter, we continue to answer theoretical questions regarding the principal component analysis (PCA) in the High Dimension, Low Sample Size (HDLSS) context.

A central question is whether the sample principal components reflect true underlying distributional structure in the HDLSS context. This is well described in a spiked covariance model, originally proposed by Johnstone (2001). A simplified version of the framework we used in the previous chapter falls into this category. A spiked covariance model assumes that the first few eigenvalues are distinctively larger than the others. We focus on a generalized version of the spike model, as described in Section 3.3, which is different from that of Johnstone (2001) and Paul (2007). Let $\Sigma_{(d)}$ denote the population covariance matrix and $S_{(d)}$ denote the sample covariance matrix. The eigen-decomposition of $\Sigma_{(d)}$ is $\Sigma_{(d)} = U_d \Lambda_d U_d'$, where Λ_d is a diagonal matrix of eigenvalues $\lambda_{1,d} \geqslant \lambda_{2,d} \geqslant \cdots \geqslant \lambda_{d,d}$ in non-increasing order, U_d is a matrix of corresponding eigenvectors so that $U_d = [u_{1,d}, \ldots, u_{d,d}]$, and $'$ denotes the transpose of the preceding matrix. The eigen-decomposition of $S_{(d)}$ is similarly defined as $S_{(d)} = \hat{U}_d \hat{\Lambda}_d \hat{U}_d'$. As a simple example of the spiked model, consider $\lambda_{1,d} = \sigma^2 d^\alpha$, $\lambda_{2,d} = \cdots = \lambda_{d,d} = \tau^2$, for $\alpha, \sigma^2, \tau^2 > 0$ fixed. The first eigenvector of $S_{(d)}$ corresponding to the largest eigenvalue is of interest, as it contains the most important variation of the data. The first sample eigenvector $\hat{u}_{1,d}$ is assessed with the angle formed by itself and its population counterpart $u_{1,d}$. The direction $\hat{u}_{1,d}$ is said to be *consistent* with $u_{1,d}$ if $\text{Angle}(\hat{u}_{1,d}, u_{1,d}) \to 0$ as $d \to \infty$. However in the HDLSS context, a perhaps counter-intuitive phenomenon frequently occurs, where the two directions tend to be as far away as possible. We say the direction $\hat{u}_{1,d}$ is *strongly inconsistent* with $u_{1,d}$ if $\text{Angle}(\hat{u}_{1,d}, u_{1,d}) \to \frac{\pi}{2}$ as $d \to \infty$. In the one spike model above, the order of magnitude α of the first eigenvalue is the key condition for these two limiting phenomena. In Chapter 2, we have shown that

$$\text{Angle}(\hat{u}_{1,d}, u_{1,d}) \to \begin{cases} 0, & \alpha > 1; \\ \frac{\pi}{2}, & \alpha < 1, \end{cases} \tag{3.1}$$

in probability (or almost surely) under some conditions. Although the gap between consistency and strong inconsistency is relatively thin, the case $\alpha = 1$ has not been investigated, and is a main focus of this chapter.

It is natural to conjecture from (3.1) that when $\alpha = 1$, the angle does not degenerate but converges to a random quantity in $(0, \frac{\pi}{2})$. This claim is established in the simple one spike model in the next section, where we describe a range of limits of the eigenvalues and eigenvectors, depending on the order of magnitude α of $\lambda_{1,d}$. In Section 3.3, the claim is generalized for multiple spike cases, and is proved in a much more general distributional setting. In a multiple spike model with $m > 1$ spikes, where the first m principal components contain the important *signal* of the distribution, the sample PCA can be assessed by simultaneously comparing the first m principal components. In particular, we investigate the limits of *distance* between two subspaces: the subspace generated by the first m sample PC directions $\hat{u}_{1,d}, \ldots, \hat{u}_{m,d}$ and the subspace by the first m population PC directions. The distance can be measured by canonical angles and metrics between subspaces, the limiting distributions of which will be investigated for the $\alpha = 1$ case, as well as the cases $\alpha \neq 1$, in Section 3.3.2. The probability density functions of the limiting distributions for $\alpha = 1$ are also derived and illustrated under Gaussian assumption, to show the effect of parameters in the distributions.

The HDLSS data set has an interesting geometric representation in the limit $d \to \infty$, as shown in Section 2.3. In Section 3.4, we extend the result and show that there are *three different geometric representations*, which coincide with the range of limits depending on α. The three different representations may be understood in connection to the phase transition phenomenon in random matrix theories (Baik et al. (2005), Nadler (2008) and Rao et al. (2008)). This is further discussed in Section 3.4.

3.2 Range of limits in the single spike model

Suppose we have a data matrix $X_{(d)} = [X_{1,d}, \ldots, X_{n,d}]$, with $d > n$, where the d dimensional random vectors $X_{i,d}$ are independent and identically distributed. We assume for now that $X_{i,d}$ is normally distributed with mean zero and covariance matrix $\mathbf{\Sigma}_{(d)}$, but the Gaussian assumption will be relaxed in the next section. The population covariance matrix $\mathbf{\Sigma}_{(d)}$ is assumed to have one spike, that is, the eigenvalues of $\mathbf{\Sigma}_{(d)}$ are $\lambda_{1,d} = \sigma^2 d^\alpha$, $\lambda_{2,d} = \cdots = \lambda_{d,d} = \tau^2$. The corresponding eigenvectors of $\mathbf{\Sigma}_{(d)}$ are denoted by $u_{i,d}$. The sample covariance matrix is defined as $S_{(d)} = \frac{1}{n} X_{(d)} X'_{(d)}$ with its ith eigenvalue and eigenvector denoted by $\hat{\lambda}_{i,d}$ and $\hat{u}_{i,d}$, respectively.

The following theorem summarizes the spectrum of the limiting distributions of the eigenvalues and eigenvectors of $S_{(d)}$, depending on the different order α of $\lambda_{1,d}$. Note that the angle between the two vectors $u, \hat{u}$ is represented by the inner product through $\mathrm{Angle}(u, \hat{u}) = \cos^{-1}(u'\hat{u})$. We also assume that the eigenvectors $u_{i,d}$, $i \geqslant 2$, are fixed.

Theorem 3.1. *Under the Gaussian assumption and the one spike case above, (i) the limit of the first eigenvalue depends on* α*:*

$$\frac{\hat{\lambda}_{1,d}}{\max(d^\alpha, d)} \longrightarrow \begin{cases} \sigma^2 \frac{\chi_n^2}{n}, & \alpha > 1; \\ \sigma^2 \frac{\chi_n^2}{n} + \frac{\tau^2}{n}, & \alpha = 1; \\ \frac{\tau^2}{n}, & \alpha < 1, \end{cases}$$

as $d \to \infty$*, where* $\longrightarrow$ *denotes the convergence in distribution, and* χ_n^2 *denotes a random variable with the* χ^2 *distribution with degree of freedom* n*. The rest of the eigenvalues converge to the same quantity when scaled, that is for any* $\alpha \in [0, \infty)$, $j = 2, \ldots, n$,

$$\frac{\hat{\lambda}_{j,d}}{d} \to \frac{\tau^2}{n}, \text{ as } d \to \infty,$$

in probability.

(ii) The limit of the first eigenvector depends on α:

$$u'_{1,d}\hat{u}_{1,d} \longrightarrow \begin{cases} 1 & \alpha > 1; \\ \left(1 + \frac{\tau^2}{\sigma^2 \chi_n^2}\right)^{-\frac{1}{2}} & \alpha = 1; \\ 0, & \alpha < 1, \end{cases}$$

as $d \to \infty$. The rest of the eigenvectors are strongly inconsistent with their population counterpart, for any $\alpha \in [0, \infty)$, $j = 2, \ldots, n$,

$$u'_{j,d}\hat{u}_{j,d} \to 0, \text{ as } d \to \infty,$$

in probability.

The case $\alpha = 1$ bridges the other two cases. In particular, the ratio of the sample and population eigenvalue $\hat{\lambda}_{1,d}/\lambda_{1,d}$ is asymptotically unbiased to 1 when $\alpha > 1$. It is asymptotically biased when $\alpha = 1$, and becomes completely deterministic in the case $\alpha < 1$, where the effect of σ^2 on $\hat{\lambda}_{1,d}$ becomes negligible. Moreover, the angle $\mathrm{Angle}(u_{1,d}, \hat{u}_{1,d})$ to the optimal direction converges to a random quantity which is defined on $(0, \pi/2)$ and depends on σ^2, τ^2, and n. The effect of those parameters on the limiting distribution of $\mathrm{Angle}(u_{1,d}, \hat{u}_{1,d})$ is illustrated in Fig. 3.1. The ratio σ^2/τ^2 can be understood as a *signal to noise* ratio. A high value of σ^2/τ^2 means that the major variation along the first PC direction is strong. Therefore, for larger values of σ^2/τ^2, $\mathrm{Angle}(u_{1,d}, \hat{u}_{1,d})$ should be closer to zero than smaller values of the ratio, as depicted in the upper panel of Fig. 3.1. Moreover, the sample PCA with larger sample size n should perform better than with smaller sample size. The sample size n becomes the degrees of freedom of the χ^2 distribution in the limit, and the bottom panel of Fig. 3.1 shows that the $\hat{u}_{i,d}$ is closer to $u_{i,d}$ for larger values of n.

The Gaussian assumption in the previous theorem appears as a driver of the limiting χ^2 distributions. Under the general non-Gaussian assumption we state in the next section, the χ^2 will be replaced by a distribution that depends heavily on the distribution of the population principal component scores, which may not be Gaussian in general.

Remark 3.1. The results in Theorem 3.1 can be used to estimate the parameters σ^2 and τ^2 in the model with $\alpha = 1$. As a simple example, one can set $\hat{\tau}^2 = \frac{n}{n-1}\sum_{j=2}^{n} \frac{\hat{\lambda}_{j,d}}{d}$ and $\hat{\sigma}^2 = \hat{\lambda}_{1,d}/d-$

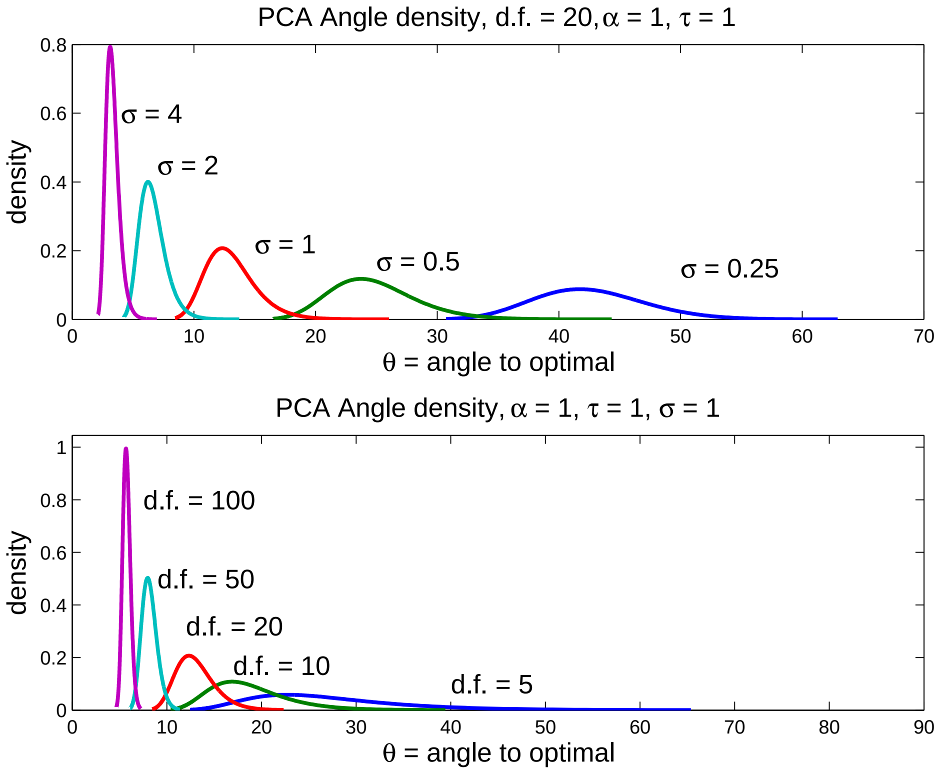


Figure 3.1: Angle densities for the one spike case. The top panel shows an overlay of the densities with different σ^2, with other parameters fixed. The bottom panel shows an overlay of the densities with different degrees of freedom n of the χ^2 distribution. For a larger signal to noise ratio $\frac{\sigma^2}{\tau^2}$, and for a larger n, the angle to optimal is smaller.

$\hat{\tau}^2/n$. Then $\hat{\tau}^2 \to \tau^2$ and $\hat{\sigma}^2 \longrightarrow \sigma^2 \frac{\chi^2_n}{n}$ as $d \to \infty$ by Theorem 1 and Slutsky's theorem. The estimator $\hat{\sigma}^2$ is not consistent but asymptotically unbiased. This can be extended to provide an asymptotically unbiased prediction of principal component scores. A similar estimation scheme can be found in a related but different setting where $d, n \to \infty$ together; see for example Paul (2007) and Lee et al. (2010). These papers do not discuss eigenvector estimation. The sample eigenvector, $\hat{u}_{1,d}$ is difficult to improve upon mainly because the direction of deviation $\hat{u}_{1,d}$ from $u_{1,d}$ is quite random and difficult to estimate unless a more restrictive assumption (e.g. sparsity) is made.

3.3 Limits under generalized spiked covariance model

The results in the previous section will be generalized to much broader situations, including a generalized spiked covariance model and a relaxation of the Gaussian assumption. We focus on the $\alpha = 1$ case, and describe the limiting distributions for eigenvalues and eigenvectors. We also point out that the subspace generated by the first few principal components is an important notion for assessing PCA, discuss measures of closeness between two subspaces, and present limit theory for measures of this type.

3.3.1 Eigenvalues and eigenvectors

In the following, all the quantities depend on the dimension d, but the subscript d is omitted if it does not cause any confusion. We first describe some elementary facts from matrix algebra, that are useful throughout this chapter. The dimension of the sample covariance matrix S increases as d grows, so it is challenging to deal with S directly. A useful approach is to use the dual of S, defined as the $n \times n$ symmetric matrix

$$S_D = \frac{1}{n} X' X,$$

by switching the role of rows and columns of X. The (i, j)th element of S_D is $\frac{1}{n} X_i' X_j$. An advantage of working with S_D is that for large d, the finite dimensional matrix S_D is positive definite with probability one, and its n eigenvalues are the same as the non-zero eigenvalues of S. Moreover, the sample eigenvectors $\hat{u}_i$ are related to the eigen-decomposition of S_D, as shown next. Let $S_D = \hat{V}_n \hat{\Lambda}_n \hat{V}_n$, where $\hat{\Lambda}_n = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_n)$ and $\hat{V}_n$ is the $n \times n$ orthogonal matrix of eigenvectors $\hat{v}_i$ corresponding to $\hat{\lambda}_i$. Recall $S = \hat{U} \hat{\Lambda} \hat{U}'$. Since S is at most rank n, we can write $S = \hat{U}_n \hat{\Lambda}_n \hat{U}_n'$, where $\hat{U}_n = [\hat{u}_1, \ldots, \hat{u}_n]$ consists of the first n columns of $\hat{U}$. The singular value decomposition of X is given by

$$X = \hat{U}_n \hat{\Lambda}_n \hat{V}_n' = \sum_{i=1}^{n} (n\hat{\lambda}_i)^{-\frac{1}{2}} \hat{u}_i \hat{v}_i'.$$

Then the kth sample principal component direction $\hat{u}_k$ for $k \leqslant n$ is proportional to $X\hat{v}_k$,

$$\hat{u}_k = (n\hat{\lambda}_k)^{-\frac{1}{2}} X \hat{v}_k. \tag{3.2}$$

Therefore the asymptotic properties of the eigen-decomposition of S, as $d \to \infty$, can be studied via those of the finite dimensional matrix S_D.

It is also useful to represent S_D in terms of the population principal components. Let $Z_{(d)}$ be the standardized principal components of X, defined by

$$Z_{(d)} = \begin{pmatrix} Z_1' \\ \vdots \\ Z_d' \end{pmatrix} = \Lambda_d^{-1/2} U_d' X,$$

where $Z_i' = (Z_{i1}, \ldots, Z_{in})$ is the ith row of $Z_{(d)}$, so that

$$Z_i' = \lambda_i^{-\frac{1}{2}} u_i' X. \tag{3.3}$$

Under the Gaussian assumption of the previous section, each element of $Z_{(d)}$ is independently distributed as the standard normal distribution. By $X = U_d \Lambda^{1/2} Z_{(d)}$,

$$S_D = \frac{1}{n} X'X = \frac{1}{n} Z_{(d)}' \Lambda Z_{(d)} = \frac{1}{n} \sum_{i=1}^{d} \lambda_i Z_i Z_i'.$$

The Gaussian assumption on X is relaxed as done in the previous chapter. We assume that each column X_i of X follows a d dimensional multivariate distribution with mean zero and covariance matrix Σ. Each entry of the standardized principal components, or the sphered variables $Z_{(d)}$ is assumed to have finite fourth moments, and is uncorrelated but in general dependent with each other. The dependency of the principal components are regulated by the ρ-mixing condition under some permutation as done Section 2.1.1. We denote these distributional assumptions as $(c1)$.

We then define a generalized spiked covariance model, which is a simplified version of the model used in Chapter 2. Recall that a simple one spike model was defined on the eigenvalues of the population covariance matrix Σ, for example, $\lambda_1 = \sigma^2 d^\alpha$, $\lambda_2 = \cdots = \lambda_d = \tau^2$. This is generalized by allowing multiple spikes, and by relaxing the uniform eigenvalue assumption in the tail to a decreasing sequence. The tail eigenvalues are regulated by a measure of sphericity ϵ_k in the limit $d \to \infty$. The ϵ_k condition holds for quite general settings (see Section 2.2). In the generalized spiked model, the eigenvalues are assumed to be of the form $\lambda_1 = \sigma_1^2 d^\alpha, \ldots, \lambda_m = \sigma_m^2 d^\alpha$, for $\sigma_1^2 \geqslant \cdots \geqslant \sigma_m^2 > 0$ for some $m \geqslant 1$, and the ϵ_{m+1} condition holds for $\lambda_{m+1}, \ldots, \lambda_d$. Also assume that $\frac{1}{d}\sum_{i=m+1}^{d} \lambda_i \to \tau^2$ as $d \to \infty$. These conditions for spike models are denoted by $(c2)$.

The following theorem gives the limits of the sample eigenvalues and eigenvectors under the general assumptions in this section. We use the following notations. Let $\varphi(A)$ be a vector of eigenvalues of a real symmetric matrix A arranged in non-increasing order and let $\varphi_i(A)$ be the ith largest eigenvalue of A. Let $v_i(A)$ denote the ith eigenvector of the matrix A corresponding to the eigenvalue $\varphi_i(A)$ and $v_{ij}(A)$ be the jth loading of $v_i(A)$. Also note that there are many choices of eigenvectors of S including the sign changes. We use the convention that the sign of $\hat{u}_i$ will be chosen so that $\hat{u}_i' u_i \geqslant 0$. Recall that the vector of the ith standardized principal component scores is $Z_i = (Z_{1i}, \ldots, Z_{ni})'$. Denote the $n \times m$ matrix of the first m principal component scores as $\mathbf{W} = [\sigma_1 Z_1, \cdots, \sigma_m Z_m]$. The limiting distributions heavily depend on the finite dimensional random matrix $\mathbf{W}$.

Theorem 3.2. *Under the assumptions $(c1)$ and $(c2)$ with fixed $n \geqslant m \geqslant 1$, if $\alpha = 1$, then (i) the sample eigenvalues*

$$d^{-1} n \hat{\lambda}_{i,d} \longrightarrow \begin{cases} \varphi_i(\mathbf{W}'\mathbf{W}) + \tau^2, & i = 1, \ldots, m; \\ \tau^2, & i = m+1, \ldots, n, \end{cases}$$

as $d \to \infty$ jointly for all i.

(ii) The inner products between the sample and population eigenvectors have limiting distributions:

$$\hat{u}_{i,d}' u_{j,d} \longrightarrow \frac{v_{ij}(\mathbf{W}'\mathbf{W})}{\sqrt{1 + \tau^2/\varphi_i(\mathbf{W}'\mathbf{W})}} \text{ as } d \to \infty \text{ jointly for } i, j = 1, \ldots, m.$$

The rest of eigenvectors are strongly inconsistent with their population counterpart, i.e.

$$\hat{u}_{i,d}' u_{i,d} \to 0 \text{ as } d \to \infty \text{ for } i = m+1, \ldots, n,$$

in probability.

The theorem shows that the first m eigenvectors are neither consistent nor strongly inconsistent to the population counterparts. The limiting distributions of angles $\text{Angle}(\hat{u}_{i,d}, u_{i,d})$ to optimal directions are supported on $(0, \pi/2)$ and depend on the magnitude of the noise τ^2 and the distribution of $\mathbf{W}'\mathbf{W}$. Note that the $m \times m$ symmetric matrix $\mathbf{W}'\mathbf{W}$ is the scaled covariance matrix of the principal component scores in the first m directions. When the underlying distribution of $\mathbf{X}$ is assumed to be Gaussian, then $\mathbf{W}'\mathbf{W}$ is the Wishart matrix $\mathcal{W}_m(n, \Lambda_m)$, where $\Lambda_m = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2)$. If $m = 1$, the matrix becomes a scalar random variable, and is χ_n^2 under the Gaussian assumption, which leads to Theorem 3.1.

The limiting distributions of the cases $\alpha \neq 1$ can be found in a similar manner, which we only state the result for the case $\alpha > 1$ and for the first m components. For more general results, see Chapter 2. For $i, j = 1, \ldots, m$, the eigenvalues $d^{-\alpha} n \hat{\lambda}_{i,d} \longrightarrow \phi_i(\mathbf{W}'\mathbf{W})$ and the inner products $\hat{u}_{i,d}' u_{j,d} \longrightarrow v_{ij}(\mathbf{W}'\mathbf{W})$ as $d \to \infty$. In comparison to the $\alpha = 1$ case, if we set τ^2 to be zero, the result becomes identical for all $\alpha \geqslant 1$.

Remark 3.2. When the sample size n also grows, consistency of sample eigenvalues and eigenvectors can be achieved. In particular, for $i = 1, \ldots, m$, we have as d grows

$$\frac{\hat{\lambda}_{i,d}}{\lambda_{i,d}} = \frac{d}{n} \frac{d^{-1} n \hat{\lambda}_{i,d}}{\sigma_i^2 d} \longrightarrow \frac{\varphi_i(\mathbf{W}'\mathbf{W}/n)}{\sigma_i^2} + \frac{\tau^2}{n\sigma_i^2}$$

by Theorem 3.2. Since $\mathbf{W}'\mathbf{W}/n \to \mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2)$ by a law of large numbers, we get the consistency of eigenvalues, i.e. $\hat{\lambda}_{i,d}/\lambda_{i,d} \to 1$ as $d, n \to \infty$, where the limits are applied successively. For the sample eigenvectors, from Theorem 3.2(ii) and because $v_{ij}(\mathbf{W}^T\mathbf{W}) \to \delta_{ij}$ as $n \to \infty$ and $\varphi_i(\mathbf{W}'\mathbf{W}) = O(n)$, we get $\hat{u}_{i,d}' u_{j,d} \to 1$ if $i = j$, 0 otherwise as $d, n \to \infty$. Therefore, the sample PC directions are consistent to the corresponding population PC directions, i.e. $\mathrm{Angle}(\hat{u}_{i,d}, u_{i,d}) \to 0$ as $d, n \to \infty$, for $i \leqslant m$.

Proof of Theorem 3.2. The following lemma shows a version of the law of large numbers for matrices, that is useful in the proof. Note that the lemma is obtained by rephrasing Theorem 2.1. Recall that $Z_i' \equiv (Z_{1i}, \ldots, Z_{ni})$ is the ith row of $Z_{(d)}$.

Lemma 3.3. *If the assumption (c1) and the ϵ_k-condition holds, then*

$$c_d^{-1} \sum_{i=k}^{d} \lambda_{i,d} Z_i Z_i' \to I_n, \text{ as } d \to \infty$$

in probability, where $c_d = n^{-1} \sum_{i=1}^{d} \lambda_{i,d}$ and I_n denotes the $n \times n$ identity matrix. In particular, *if* $d^{-1}\sum_{i=k}^{d}\lambda_{i,d}\to\tau^2$, *then*

$$\frac{1}{d}\sum_{i=k}^{d}\lambda_{i,d}Z_iZ_i'\to\tau^2 I_n,\ \textit{as } d\to\infty$$

in probability.

This lemma is used to show that the spectral decomposition of $d^{-1}nS_D$,

$$d^{-1}nS_D=d^{-1}\sum_{i=1}^{m}\sigma_i^2Z_iZ_i'+d^{-1}\sum_{i=m+1}^{d}\lambda_iZ_iZ_i',$$

can be divided into two parts, and the latter converges to a deterministic part. Applying Lemma 3.3, we have $d^{-1}nS_D\longrightarrow S_0$ as $d\to\infty$, where

$$S_0=\mathbf{W}\mathbf{W}'+\tau^2I_n.$$

Then since the eigenvalues of a symmetric matrix A are a continuous function of elements of A, we have

$$\varphi(d^{-1}nS_D)\longrightarrow\varphi(S_0),$$

as $d\to\infty$. Noticing that for $i=1,\ldots,m$,

$$\varphi_i(S_0)=\varphi_i(\mathbf{W}\mathbf{W}')+\tau^2=\varphi_i(\mathbf{W}'\mathbf{W})+\tau^2,$$

and for $i=m+1,\ldots,n$, $\varphi_i(S_0)=\tau^2$ gives the result.

For the eigenvectors, note that the eigenvectors $\hat{v}_i$ of $d^{-1}nS_D$ can be chosen so that they are continuous (Acker (1974)). Therefore, we also have that $\hat{v}_i=v_i(d^{-1}nS_D)\longrightarrow v_i(S_0)$ as $d\to\infty$, for all i. Also note that $v_i(S_0)=v_i(\mathbf{W}\mathbf{W}')$ for $i\leqslant m$.

Similar to the dual approach for covariance matrices, the eigenvectors of the $n\times n$ matrix $\mathbf{W}\mathbf{W}'$ can be evaluated from the dual of the matrix. In particular, let $\mathbf{W}=U_w\Lambda_wV_w'=\sum_{i=1}^{m}\lambda_{iw}u_{iw}v_{iw}'$, where $\lambda_{iw}^2=\varphi_i(\mathbf{W}'\mathbf{W})$ and $v_{iw}=v_i(\mathbf{W}'\mathbf{W})$. Then

$$v_1(S_0)=u_{iw}=\frac{\mathbf{W}v_{iw}}{\lambda_{iw}}=\frac{\mathbf{W}v_i(\mathbf{W}'\mathbf{W})}{\sqrt{\varphi_i(\mathbf{W}'\mathbf{W})}}.$$

Now from (3.2),(3.3) and the previous equation, for $1 \leqslant i, j \leqslant m$,

$$\begin{aligned} u_j'\hat{u}_i &= u_j' \frac{X\hat{v}_i}{\sqrt{n\hat{\lambda}_i}} = \frac{u_j' X \hat{v}_i}{\sqrt{n\hat{\lambda}_i}} = \frac{\sqrt{\sigma_j^2 d} Z_j' \hat{v}_i}{\sqrt{n\hat{\lambda}_i}} = \frac{\sigma_j Z_j' \hat{v}_i}{\sqrt{d^{-1} n \hat{\lambda}_i}} \\ &\longrightarrow \frac{\sigma_j Z_j' \mathbf{W} v_i(\mathbf{W}'\mathbf{W})}{\sqrt{\varphi_i(\mathbf{W}'\mathbf{W}) + \tau^2} \sqrt{\varphi_i(\mathbf{W}'\mathbf{W})}}. \end{aligned} \tag{3.4}$$

Note that $\sigma_j Z_j' \mathbf{W} = [\sigma_j \sigma_1 Z_j' Z_1 \cdots \sigma_j \sigma_m Z_j' Z_m]$ is the jth row of $\mathbf{W}'\mathbf{W}$ and $\mathbf{W}'\mathbf{W} v_i(\mathbf{W}'\mathbf{W}) = \varphi_i(\mathbf{W}'\mathbf{W}) v_i(\mathbf{W}'\mathbf{W})$. Therefore, the limiting form (3.4) becomes

$$\frac{\varphi_i(\mathbf{W}'\mathbf{W}) v_{ij}(\mathbf{W}'\mathbf{W})}{\sqrt{\varphi_i(\mathbf{W}'\mathbf{W}) + \tau^2} \sqrt{\varphi_i(\mathbf{W}'\mathbf{W})}}.$$

For $i = m+1, \ldots, n$, again from (3.2) and (3.3), we get

$$u_i'\hat{u}_i = \frac{\sqrt{\lambda_i} Z_i' \hat{v}_i}{\sqrt{n\hat{\lambda}_i}} = d^{-\frac{1}{2}} \frac{\tau Z_i' \hat{v}_i}{\sqrt{n\hat{\lambda}_i / d}} = O(d^{-\frac{1}{2}}).$$

□

3.3.2 Angles between principal component spaces

Under the generalized spiked covariance model with $m > 1$, the first m population principal directions provide a basis of the most important variation. Therefore, it would be more informative to investigate the deviation of each $\hat{u}_i$ from the subspace $\mathcal{L}_1^m(d)$ spanned by $\{u_{1,d}, \ldots, u_{m,d}\}$. Also, denote the subspace spanned by the first m sample principal directions as $\hat{\mathcal{L}}_1^m(d) \equiv \text{span}\{\hat{u}_{1,d}, \ldots, \hat{u}_{m,d}\}$. When performing dimension reduction, it is critical for the sample PC space $\hat{\mathcal{L}}_1^m$ to be close to the population PC space $\mathcal{L}_1^m$. The closeness of two subspaces can be measured in terms of *canonical angles.*

We briefly introduce the notion of canonical angles and metrics between subspaces, detailed discussions of which can be found in Stewart and Sun (1990) and Gaydos (2008). As a simple case, the canonical angle between a 1-dimensional subspace and an m-dimensional subspace is defined as follows. Let $\hat{\mathcal{L}}_i$ be the 1-dimensional linear space with basis $\hat{u}_i$. Infinitely many angles can be formed between $\hat{\mathcal{L}}_i$ and $\mathcal{L}_1^m$ with $m > 1$. The canonical angle, denoted by $\text{Angle}(\hat{\mathcal{L}}_i, \mathcal{L}_1^m)$, is defined by the smallest angle formed, that is the angle between $\hat{u}_i$ and its projection $\hat{u}_i^P$ onto $\mathcal{L}_1^m$. This angle is represented in terms of an inner product as

$$\text{Angle}(\hat{\mathcal{L}}_i, \mathcal{L}_1^m) = \cos^{-1}\left(\frac{\hat{u}_i' \hat{u}_i^P}{\left\|\hat{u}_i^P\right\| \left\|\hat{u}_i\right\|}\right). \tag{3.5}$$

When two multi-dimensional subspaces are considered, multiple canonical angles are defined. Among angles between $\hat{\mathcal{L}}_1^m$ and $\mathcal{L}_1^m$, the first canonical angle is geometrically defined as

$$\theta_1(\hat{\mathcal{L}}_1^m, \mathcal{L}_1^m) = \max_{\mathbf{x}\in\hat{\mathcal{L}}_1^m} \min_{\mathbf{y}\in\mathcal{L}_1^m} \text{Angle}(\mathbf{x}, \mathbf{y}) \text{ for } \|\mathbf{x}\|, \|\mathbf{y}\| > 0, \tag{3.6}$$

where $\text{Angle}(\mathbf{x}, \mathbf{y})$ is the angle formed by the two vectors $\mathbf{x}$, $\mathbf{y}$. One can show that the second canonical angle is defined by the same geometric relation as above with $\hat{\mathcal{L}}_{-\mathbf{x}}$ and $\mathcal{L}_{-\mathbf{y}}$ for $\mathbf{x}$, $\mathbf{y}$ from (3.6), where $\hat{\mathcal{L}}_{-\mathbf{x}}$ is the orthogonal complement of $\mathbf{x}$ in $\hat{\mathcal{L}}_1^m$. In practice, the canonical angles are found by the singular value decomposition of a matrix. Let $\hat{U}_m$ and U_m be orthonormal bases for $\hat{\mathcal{L}}_1^m$, $\mathcal{L}_1^m$ and γ_i's be the singular values of $\hat{U}_m' U_m$. Then the canonical angles are

$$\theta_i(\hat{\mathcal{L}}_1^m, \mathcal{L}_1^m) = \cos^{-1}(\gamma_i)$$

in descending order.

Distances between two subspaces can be defined using the canonical angles. We point out two metrics from Chapter II.4 of Stewart and Sun (1990):

1. *gap metric* $\rho_g(\hat{\mathcal{L}}_1^m, \mathcal{L}_1^m) = \sin(\theta_1)$,

2. *Euclidean sine metric* $\rho_s(\hat{\mathcal{L}}_1^m, \mathcal{L}_1^m) = \{\sum_{i=1}^m \sin^2(\theta_i)\}^{\frac{1}{2}}$.

The gap metric is simple and only involves the largest canonical angle. The Euclidean sine metric makes use of all canonical angles and thus gives more comprehensive understanding of the closeness between the two subspaces. We will use both metrics in the following discussion.

At first, we examine the limiting distribution of the angle between the sample PC direction $\hat{u}_i$ and $\mathcal{L}_1^m$.

Theorem 3.4. *Under the assumptions (c1) and (c2) with fixed $n \geqslant m \geqslant 1$, if $\alpha = 1$, then for*

$i = 1, \ldots, m$, the canonical angle converges in distribution:

$$\cos\left(Angle(\hat{\mathcal{L}}_i, \mathcal{L}_1^m)\right) \longrightarrow \frac{1}{\sqrt{1 + \tau^2/\varphi_i(\mathbf{W}'\mathbf{W})}} \text{ as } d \to \infty.$$

Proof. Since $\hat{u}_i^P = \sum_{j=1}^m (\hat{u}_i' u_j) u_j$,

$$\frac{\hat{u}_i' \hat{u}_i^P}{\|\hat{u}_i^P\| \, \|\hat{u}_i\|} = \|\hat{u}_i^P\| = \sqrt{(\hat{u}_i' u_1)^2 + \cdots + (\hat{u}_i' u_m)^2}.$$

The result follows from (3.5), Theorem 3.2(ii) and the fact that $\sum_{j=1}^m (v_{ij}(\mathbf{W}'\mathbf{W}))^2 = \|v_i(\mathbf{W}'\mathbf{W})\|^2 = 1$. $\square$

We then investigate the limiting behavior of the distances between $\hat{\mathcal{L}}_1^m$ and $\mathcal{L}_1^m$, in terms of either the canonical angles or the distances. From the fact that $\hat{U}_m = [\hat{u}_1, \ldots, \hat{u}_m]$ and $U_m = [u_1, \ldots, u_m]$ are orthonormal bases of $\hat{\mathcal{L}}_1^m$ and $\mathcal{L}_1^m$ respectively, cosines of the canonical angles are the singular values of $\hat{U}_m' U_m$. Since the (i,j)th element of $\hat{U}_m' U_m$ is $\hat{u}_i' u_j$, Theorem 3.2 leads to

$$\hat{U}_m' U_m \longrightarrow \left[v_1(\mathbf{W}'\mathbf{W}) \cdots v_m(\mathbf{W}'\mathbf{W})\right] \begin{pmatrix} \left(1 + \frac{\tau^2}{\varphi_1(\mathbf{W}'\mathbf{W})}\right)^{-\frac{1}{2}} & & 0 \\ & \ddots & \\ 0 & & \left(1 + \frac{\tau^2}{\varphi_m(\mathbf{W}'\mathbf{W})}\right)^{-\frac{1}{2}} \end{pmatrix},$$

as $d \to \infty$. Therefore the canonical angles $(\theta_1, \ldots, \theta_m)$ between $\hat{\mathcal{L}}_1^m$ and $\mathcal{L}_1^m$ converge to the arccosines of

$$(\left(1 + \tau^2/\varphi_m(\mathbf{W}'\mathbf{W})\right)^{-\frac{1}{2}}, \ldots, \left(1 + \tau^2/\varphi_1(\mathbf{W}'\mathbf{W})\right)^{-\frac{1}{2}}), \tag{3.7}$$

as $d \to \infty$. Notice that these canonical angles between subspaces converge to the same limit as in Theorem 3.4, except that the order is reversed. In particular, the limiting distribution of the largest canonical angle θ_1 is the same as that of the angle between $\hat{u}_m$ and $\mathcal{L}_1^m$, and the smallest canonical angle θ_m corresponds to the angle between the first sample PC direction $\hat{u}_1$ and the population PC space $\mathcal{L}_1^m$.

The limiting distributions of the distances between two subspaces are readily derived by the discussions so far. When using the gap metric,

$$\rho_g(\hat{\mathcal{L}}_1^m, \mathcal{L}_1^m) \longrightarrow \left(1 + \varphi_m(\mathbf{W}'\mathbf{W})/\tau^2\right)^{-\frac{1}{2}} \text{ as } d \to \infty.$$

And by using the Euclidean sine metric,

$$\rho_s(\hat{\mathcal{L}}_1^m, \mathcal{L}_1^m) \longrightarrow \left(\sum_{i=1}^{m} \frac{1}{1 + \varphi_i(\mathbf{W}'\mathbf{W})/\tau^2}\right)^{\frac{1}{2}} \text{ as } d \to \infty. \tag{3.8}$$

Remark 3.3. The convergence of the canonical angles for the case $\alpha > 1$ has been shown earlier in the previous chapter. Recall the notion of *subspace consistency*, where the direction $\hat{u}_i$ may not be consistent to u_i but will tend to lie in $\mathcal{L}_1^m$, i.e. $\text{Angle}(\hat{\mathcal{L}}_i, \mathcal{L}_1^m) \to 0$ as $d \to \infty$, for $i \leqslant m$. In this case, the canonical angles between $\hat{\mathcal{L}}_i^m$ and $\mathcal{L}_1^m$ and the distances will converge to 0 as d grows. In that sense, the empirical PC space $\hat{\mathcal{L}}_i^m$ is consistent to $\mathcal{L}_1^m$. On the other hand, when $\alpha < 1$, all directions $\hat{u}_i$ tend to behave as if they were from the eigen-decomposition of the identity matrix. Therefore, all angles tend to be $\pi/2$ and the distances will converge to their largest possible values, leading to the strong inconsistency.

We now focus back on the $\alpha = 1$ case, and illustrate the limiting distributions of the canonical angles and the Euclidean sine distance, to see the effect of parameters in the distribution. For simplicity and clear presentation, the results corresponding to $m = 2$ are presented under the Gaussian assumption. Note that the limiting distributions depend on the marginal distributions of the first few principal component scores. Therefore no common distribution is evaluated in the limit.

Denote σ^2 $(= \sigma_1^2 + \sigma_2^2)$ for the (scaled) total variance of the first two principal components. The ratio $\frac{\sigma^2}{\tau^2}$ is understood as a signal to noise ratio, similar to the single spike case. Since the ratio of σ_1^2 and σ_2^2 affects the limiting distributions, we use (λ_1, λ_2) with $\lambda_1 + \lambda_2 = 1$ so that $\sigma^2(\lambda_1, \lambda_2) = (\sigma_1^2, \sigma_2^2)$. Note that for $\mathbf{W}'\mathbf{W} \sim \mathcal{W}_2(n, \text{diag}(\sigma_1^2, \sigma_2^2))$, $\varphi(\mathbf{W}'\mathbf{W})$ has the same law as $\sigma^2\varphi(\mathcal{W}_2(n, \text{diag}(\lambda_1, \lambda_2)))$.

The joint limiting distribution of the two canonical angles in (3.7), also in Theorem 3.4, is illustrated in Fig. 3.2, with various values of $(\sigma^2/\tau^2, \lambda_1/\lambda_2)$ and fixed n. Note that for large d, the first canonical angle $\theta_1 \approx \text{Angle}(\hat{u}_{2,d}, \mathcal{L}_1^m)$ and $\theta_2 \approx \text{Angle}(\hat{u}_{1,d}, \mathcal{L}_1^m)$, and that $\theta_1 \geqslant \theta_2$.

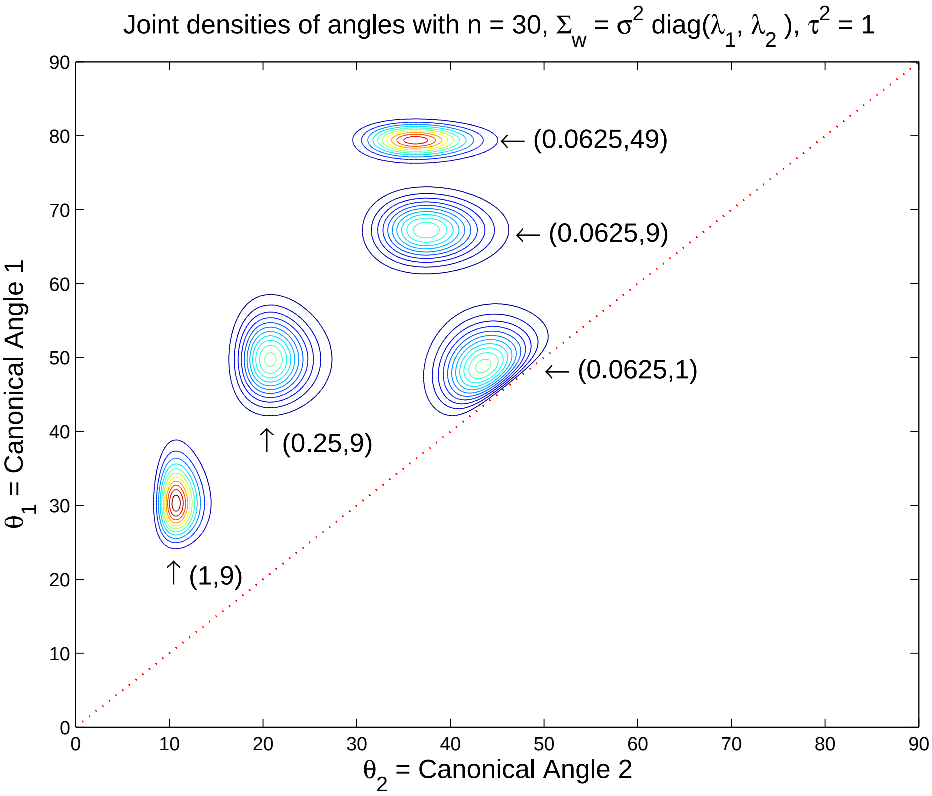


Figure 3.2: Overlay of contours of densities of canonical angles for the $m = 2$ case, corresponding to different $(\sigma^2/\tau^2, \lambda_1/\lambda_2)$. Larger signal to noise ratios σ^2/τ^2 lead to the diagonal shift of the density function, and both canonical angles will have smaller values. For a fixed σ^2/τ^2, the ratio λ_1/λ_2 between the first and second PC variances is a driver for different distributions, depicted as the vertical shift of the density function.

- The diagonal shift of the joint densities in Fig. 3.2 is driven by different σ^2s with other parameters fixed. Both θ_1 and θ_2 are smaller for larger signal to noise ratios.
- For fixed σ^2/τ^2, several values of the ratio between the first and second variances (λ_1/λ_2) are considered, and the overlay of densities according to different λ_1/λ_2 is illustrated as the vertical shift in Fig. 3.2. When the variation along the first PC direction is much stronger than that along the second, i.e. when λ_1/λ_2 is large, θ_2 becomes smaller but θ_1 tends to be much larger. In other words, $\hat{u}_1$ is a reasonable estimate of u_1, but $\hat{u}_2$ becomes a poor estimate of u_2.

See (3.15) in the appendix for the probability density function of the canonical angles.

The limiting distribution of the Euclidean sine distance between $\hat{\mathcal{L}}_1^m$ and $\mathcal{L}_1^m$ is also depicted in Fig. 3.3, again with various values of $(\sigma^2, \lambda_1, \lambda_2)$. It can be checked from the top panel of Fig. 3.3 that the distance to the optimal subspace is smaller when the signal to noise

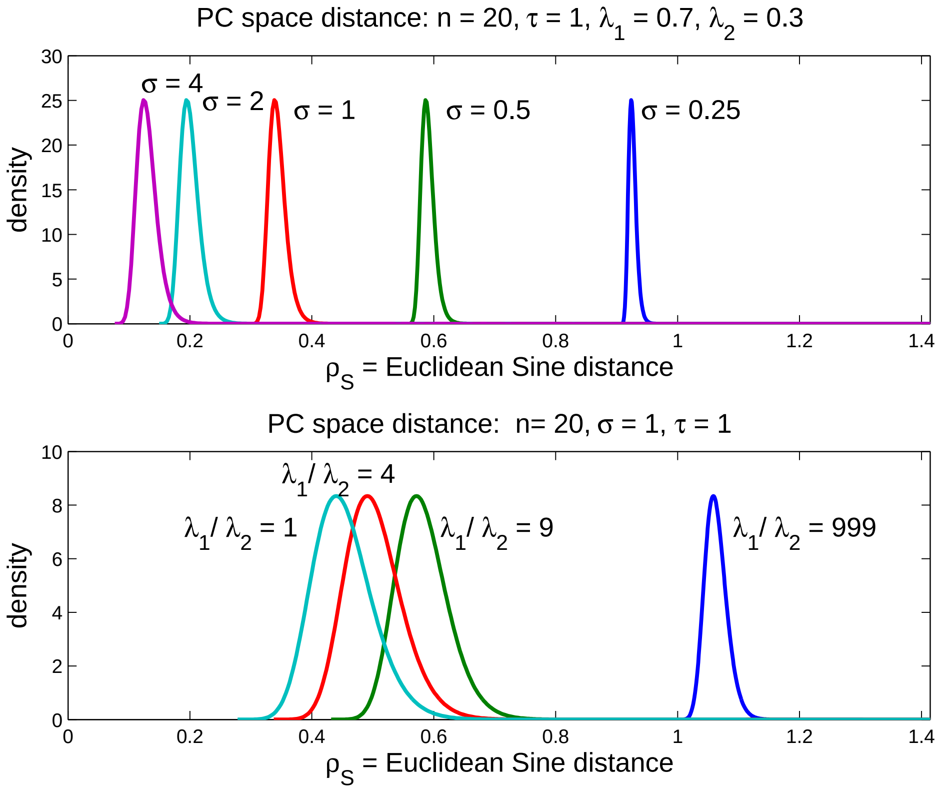


Figure 3.3: Overlay of densities of the distance between the sample and population PC spaces, measured by ρ_s for the $m = 2$ case. The top panel shows a transition of the density function corresponding to different signal to noise ratios. The bottom panel illustrates the effect of the ratio λ_1/λ_2 between the first two eigenvalues. For a larger signal to noise ratio σ^2/τ^2, and for a smaller value of λ_1/λ_2, the Euclidean sine distance is smaller.

ratio is larger. The bottom panel illustrates the densities corresponding to different ratios of λ_1/λ_2. The effect of λ_1/λ_2 is relatively small compared to the effect of different σ^2s, unless λ_2 is too small.

3.4 Geometric representations of the HDLSS data

Hall et al. (2005) first showed that the HDLSS data has an interesting *geometric representation* in the limit $d \to \infty$. In particular, for large d, the data tend to appear at vertices of a regular simplex and the variability is contained in the random rotation of the simplex. In the spike model we consider, this geometric representation of the HDLSS data holds when $\alpha < 1$, as shown earlier in Section 2.3. The representation in mathematical terms is

$$\|X_i\| = \tau\sqrt{d} + o_p(\sqrt{d}), \|X_i - X_j\| = \tau\sqrt{2d} + o_p(\sqrt{d}), \tag{3.9}$$

for d dimensional X_i, $i = 1, \dots, n$. This simplified representation has been used to show some high dimensional limit theory for discriminant analysis, see Ahn et al. (2007), Qiao et al. (2010) and Huang et al. (2010).

Similar types of representation can be derived in the $\alpha \geqslant 1$ case. When $\alpha > 1$, where consistency of PC directions happens, we have

$$\|X_i\| / d^{\alpha/2} \longrightarrow \|Y_i\|, \|X_i - X_j\| / d^{\alpha/2} \longrightarrow \|Y_i - Y_j\| \tag{3.10}$$

where $Y_i = (\sigma_1 Z_{1i}, \dots \sigma_m Z_{mi})'$s are m-dimensional independent random vectors with mean zero and covariance matrix $\mathrm{diag}(\sigma_1^2, \dots, \sigma_m^2)$. To understand Y_i, let X_i^P be the projection of X_i onto the true PC space $\mathcal{L}_1^m$. It can be checked from $d^{-1/2} X_i^P = \sum_{j=1}^m (\sigma_j Z_{ji}) u_j$ that Y_i is the vector of loadings of the scaled X_i^P in the first m principal component coordinates. When $\alpha = 1$, a deterministic term is added:

$$\|X_i\|^2 / d^1 \longrightarrow \|Y_i\|^2 + \tau^2, \|X_i - X_j\|^2 / d \longrightarrow \|Y_i - Y_j\|^2 + 2\tau^2, \tag{3.11}$$

These results can be understood geometrically, as summarized and discussed in the following;

$\alpha > 1$: The variability of samples is restricted to the true PC space $\mathcal{L}_1^m$ for large d, which coincides with the notion of subspace consistency discussed in Remark 3.3. The d-dimensional probability distribution degenerates to the m-dimensional subspace $\mathcal{L}_1^m$.

$\alpha = 1$: (3.11) is understood with a help of Pythagorean theorem, that is, the norm of X_i is asymptotically decomposed into orthogonal random and deterministic parts. Thus, data tend to be $\tau\sqrt{d}$ away from $\mathcal{L}_1^m$, and X_is projected on $\mathcal{L}_1^{m\perp} = \mathrm{span}\{u_{m+1}, \dots, u_d\}$, the orthogonal complement of $\mathcal{L}_1^m$, will follow the representation similar to $\alpha < 1$ case.

$\alpha < 1$: The geometric representation (3.9) holds.

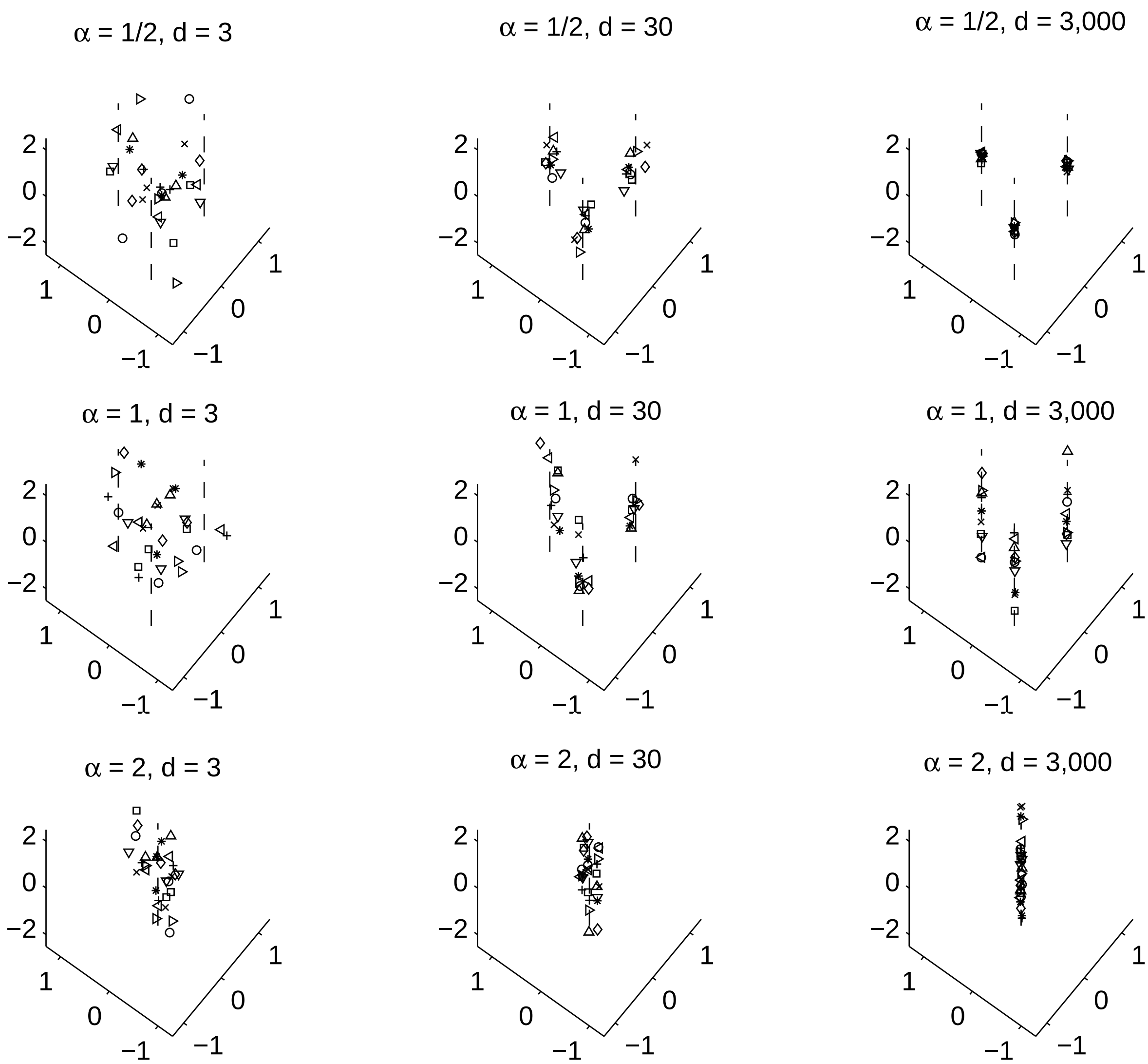
Note that the case $\alpha = 1$ smoothly bridges the others.

An example elucidating these ideas is shown in Fig. 3.4. Each panel shows scatter of 10 different samples (shown as different symbols) of $n = 3$ Gaussian random vectors in dimensions $d = 3, 30, 3000$ (shown in respective columns of Fig. 3.4. In the spiked model, we take $\sigma =$

$\tau = 1$ and $m = 1$ for simplicity and investigate three different orders of magnitude $\alpha = \frac{1}{2}, 1, 2$ of the first eigenvalue $\lambda_1 = d^\alpha$. For each pair of (d, α), each sample X_i is projected onto the first true PC direction u_1, shown as the vertical axis. In the orthogonal $d-1$ dimensional subspace, the 2-dimensional hyperplane that is generated by the data is found, and the data are projected onto that. Within the hyperplane, variation due to rotation is removed by optimally rotating the data onto edges of a regular triangle by a Procrustes method, to give the horizontal axes in each part of Fig. 3.4. These axes are scaled by dividing by $\max(d^\alpha, d)$ and the 10 samples give an impression of the various types of convergence as a function of d, for each α.

The asymptotic geometric representations summarized above can be confirmed by investigation of the figure. For $\alpha = \frac{1}{2}$, it is expected from (3.9) that the data are close to the vertices of the regular triangle, with edge length $\sqrt{2d}$. The vertices of the triangle (in the horizontal plane) with vertical rays (representing u_1, the first PC direction) are shown as the dashed lines in the first two rows of Fig. 3.4. Note that for $d = 3$, in the top row, the points appear to be quite random, but for $d = 30$, there already is reasonable convergence to the vertices with notable variation along u_1. The case $d = 3000$ shows more rigid representation with much less variation along u_1. On the other hand, for the case $\alpha = 2$ in the last row of Fig. 3.4, most of the variation in the data is found along u_1, shown as the vertical dotted line, and the variation perpendicular to u_1 becomes negligible as d grows, which confirms the degeneracy to $\mathcal{L}_1^1$ in (3.10). From these examples, conditions for consistency and strong inconsistency can be checked heuristically. The sample eigenvector $\hat{u}_1$ is consistent with u_1 when $\alpha > 1$, since the variation along u_1 is so strong that $\hat{u}_1$ should be close to that. $\hat{u}_1$ is inconsistent with u_1 when $\alpha < 1$, since the variation along u_1 becomes negligible so that $\hat{u}_1$ will not be near u_1.

For the $\alpha = 1$ case, in the middle row of Fig. 3.4, it is expected from (3.11) that each data point will be asymptotically decomposed into a random and a deterministic part. This is confirmed by the scatterplots, where the order of variance along u_1 remains comparable to that of horizontal components, as d grows. The convergence to the vertices is noticeable even for $d = 30$, which becomes stronger for larger d, while the randomness along u_1 remains for large d. Also observe that the distance from each X_i to the space spanned by u_1 becomes deterministic for large d, supporting the first part of (3.11).


α = 1/2, d = 3

α = 1/2, d = 30

α = 1/2, d = 3,000

α = 1, d = 3

α = 1, d = 30

α = 1, d = 3,000

α = 2, d = 3

α = 2, d = 30

α = 2, d = 3,000

Figure 3.4: Gaussian toy example, showing the geometric representations of HDLSS data, with $n = 3$, for three different choices of $\alpha = 1/2, 1, 2$ of the spiked model and increasing dimensions $d = 3, 30, 3000$. For $\alpha \neq 1$, data converge to vertices of a regular 3-simplex (case $\alpha < 1$) or to the first PC direction (case $\alpha > 1$). When $\alpha = 1$, data are decomposed into the deterministic part on the horizontal axes and the random part along the vertical axis.

We conclude the discussion by pointing a connection to the phase transition phenomenon investigated in the random matrix theories. The phase transition phenomenon refers that, simply put, the eigenvalue and eigenvector of the sample covariance matrix S lose their connection of the population eigen-structure of $\Sigma = \mathrm{E}S$ if the signal variance is not much larger than the noise variance as $d, n \to \infty$ together. See Baik et al. (2005), Nadler (2008) and Rao et al. (2008) and references therein. The nature of the difference of the geometric representations in HDLSS asymptotics is similar to the phase transition in that the true eigenvector is drowned by large amount of noises when $\alpha \leqslant 1$. Although these two different asymptotic studies share a same message, they do not overlap to each other. For example, Nadler's result for the one spike model is that if $\sqrt{\frac{n}{d}} < \frac{\tau^2}{\lambda_{1,d}}$ then Angle$(\hat{u}_1, u_1)$ degenerates to zero in the limit. However, putting $\lambda_{1,d} = d^\alpha$ the condition turns to $\alpha < 1/2$ which does not coincide with the observation $\alpha < 1$ in the n fixed results. An inspection of the difference may be an interesting investigation.

3.5 Derivation of the density functions

The probability density functions of the limiting distributions in (3.7) and (3.8) will be derived for the case $m = 2$, under the normal assumption. The argument is readily generalized to all m, but the normal assumption is essential.

We first recall some necessary notions for treating the Wishart matrix $\mathbf{W}'\mathbf{W}$ and eigen-decompositions. Most of the results are adopted from Muirhead (1982). Let $A \sim W_m(n, \Sigma_m)$ and denote its eigen-decomposition as $A = HLH'$ with $L = \mathrm{diag}(l_1, \ldots, l_m)$. Assume Σ_m is positive definite and $n \geqslant m$ so that $l_1 > l_2 > \cdots > l_m > 0$ with probability 1. Denote $O(m) = \{H_{m\times m} : H'H = I_m\}$ for the set of orthonormal $m \times m$ matrices and (dH) for $H \in O(m)$ as the differential form representing the uniform probability measure on $O(m)$. The multivariate gamma function is defined as

$$\Gamma_m(a) = \pi^{m(m-1)/4} \prod_{i=1}^{m} \Gamma\left(a - \frac{1}{2}(i-1)\right),$$

where $\Gamma(\cdot)$ is the usual gamma function. For $H \equiv [\mathbf{h}_1, \ldots, \mathbf{h}_m] \in O(m)$,

$$(dH) \equiv \frac{1}{\mathrm{Vol}(O(m))}(H'dH) = \frac{\Gamma_m(\frac{m}{2})}{2^m \pi^{m^2/2}}(H'dH), \tag{3.12}$$

where $(H'dH) \equiv \prod_{i>j}^{m} \mathbf{h}_i' d\mathbf{h}_j$.

We are now ready to state the density function of $\varphi(\mathbf{W}'\mathbf{W})$ for $m = 2$. Note that under the Gaussian assumption $\mathbf{W}'\mathbf{W}$ is the 2×2 Wishart matrix with degree of freedom n and covariance matrix $\Sigma_W = \mathrm{diag}(\sigma_1^2, \sigma_2^2)$. For simplicity, write $(L_1, L_2) = \varphi(\mathbf{W}'\mathbf{W})$, and $L = \mathrm{diag}(L_1, L_2)$. Then the joint density function of L_1 and L_2 is given by e.g. Theorem 3.2.18 of

Muirhead (1982) with $m = 2$, and

$$f_L(l_1, l_2) = \frac{\pi 2^{-n}(\sigma_1^2\sigma_2^2)^{-\frac{n}{2}}}{\Gamma_2(\frac{n}{2})}(l_1 l_2)^{\frac{n-3}{2}}(l_1 - l_2)\int_{O(2)} \exp\left(\mathrm{tr}\left(-\frac{1}{2}\Sigma_W^{-1}HLH'\right)\right)(dH). \quad (3.13)$$

The integral can not be solved analytically but can be simplified by using the special orthogonal group $SO(2) = \{H \in O(2) : \mathrm{det}H = 1\}$. We can parameterize $H \in SO(2)$ as

$$H = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = [\mathbf{h}_1, \mathbf{h}_2] \quad (0 < \theta \leqslant 2\pi).$$

Then $(H'dH) = \mathbf{h}_2' d\mathbf{h}_1 = d\theta$. Moreover the integral in (3.13) over $O(2)$ is twice as large as the integral over $SO(2)$. This fact and the definition (3.12) together with the parametrization above give

$$\begin{aligned}
&\int_{O(2)} \exp\left(\mathrm{tr}\left(-\frac{1}{2}\Sigma_W^{-1}HLH'\right)\right)(dH) \\
&\quad = \frac{1}{2\pi}\int_{SO(2)} \exp\left(\mathrm{tr}\left(-\frac{1}{2}\Sigma_W^{-1}HLH'\right)\right)(H'dH) \\
&\quad = \frac{1}{2\pi}\int_0^{2\pi} \exp\left(-\frac{1}{2}\left[A\cos^2\theta + B\sin^2\theta\right]\right)d\theta \\
&\quad = \frac{1}{2\pi}e^{-\frac{A+B}{4}}\int_0^{2\pi} \exp\left(\frac{1}{4}(B-A)\cos t\right)dt, \\
&\quad = e^{-\frac{A+B}{4}} I_0\left(\frac{1}{4}(B-A)\right)
\end{aligned}$$

where $A = \frac{l_1}{\sigma_1^2} + \frac{l_2}{\sigma_2^2}$, $B = \frac{l_2}{\sigma_1^2} + \frac{l_1}{\sigma_2^2}$ and

$$I_0(x) = \frac{1}{2\pi}\int_0^{2\pi} \exp\left(x\cos t\right)dt$$

is the modified Bessel function of the first kind. Note that the integral can also be represented by the hypergeometric function of matrix arguments (see Section 7.3 of Muirhead (1982)). We chose to use $I_0(x)$ since it is numerically more stable than the hypergeometric function.

Then (3.13) becomes

$$f_L(l_1, l_2) = \frac{\pi 2^{-n}(\sigma_1^2\sigma_2^2)^{-\frac{n}{2}}}{\Gamma_2(\frac{n}{2})}(l_1 l_2)^{\frac{n-3}{2}}(l_1 - l_2)e^{-\frac{A+B}{4}} I_0\left(\frac{1}{4}(B - A)\right). \tag{3.14}$$

Now the distribution of the canonical angles (in (3.7) and Theorem 3.4) is obtained by applying the change of variable on the density (3.14). Let Y_1, Y_2 be the two canonical angles, in the reverse order. Then from

$$\begin{aligned}(Y_1, Y_2) &= \left(\cos^{-1}\{(1 + \tau^2/L_1)^{-1/2}\}, \cos^{-1}\{(1 + \tau^2/L_2)^{-1/2}\}\right) \\ &= \left(tan^{-1}(\sqrt{\tau^2/L_1}), tan^{-1}(\sqrt{\tau^2/L_2})\right),\end{aligned}$$

the joint density function of Y_1, Y_2 becomes

$$f_{Y_1,Y_2}(y_1, y_2) = f_L(\tau^2 \cot^2 y_1, \tau^2 \cot^2 y_2) \cdot (2\tau^2)^2 \frac{\cos y_1}{\sin^3 y_1} \frac{\cos y_2}{\sin^3 y_2} \tag{3.15}$$

on $0 < y_1 < y_2 < \frac{\pi}{2}$.

The limiting distribution of the distances between the empirical and population principal subspace, measured by the Euclidean sine metric, $\rho_s(\hat{\mathcal{L}}_1^m, \mathcal{L}_1^m)$ in (3.8) is obtained as follows. Let

$$Z_1 = \sqrt{\frac{\tau^2}{\tau^2 + L_1} + \frac{\tau^2}{\tau^2 + L_2}}, Z_2 = \frac{\tau^2}{\tau^2 + L_2}$$

so that

$$L_1 = \frac{\tau^2}{Z_1^2 - Z_2} - \tau^2, L_2 = \frac{\tau^2}{Z_2} - \tau^2.$$

The distribution of Z_1 is the limiting distribution of interest. Note that the eigenvalues $(L_1, L_2) \sim f_L$ must satisfy $0 < L_2 < L_1 < \infty$. This leads to the support for the joint distribution of (Z_1, Z_2):

$$D = \{Z_1, Z_2 \in \mathbb{R} : Z_2 < Z_1^2, Z_1^2 < 2Z_2, Z_2 < 1\}.$$

By the change of variable on f_L (3.14), we get

$$f_{Z_1,Z_2}(z_1,z_2) = f_L(\frac{\tau^2}{z_1^2 - z_2} - \tau^2, \frac{\tau^2}{z_2} - \tau^2) \cdot 2z_1 \frac{1}{\tau^4} \left(\frac{\tau^2}{z_2} \frac{\tau^2}{z_1^2 - z_2} \right)^2 1_{(z_1,z_2 \in D)}. \tag{3.16}$$

The marginal density of Z_1 can be obtained by numerical integration of f_{Z_1,Z_2}. The support of the density is then $(0, \sqrt{2})$.

Chapter 4

Discussions on High Dimensional Asymptotic Studies

4.1 Different regimes in high-dimensional asymptotics

The high dimension, low sample size data situation or the so called "large p, small n" situation becomes a common subject in statistics. In recent years substantial work has been done on the statistical inference in high dimensional sample spaces (or the space of parameters). A useful framework to organize these works is, using notations d for the dimension and n for the sample size:

1. Classical: $d/n \to 0$ as $n \to \infty$.
2. Random matrices: $d/n \to c \in (0, \infty)$, as $n \to \infty$.
3. HDLSS: $d/n \to \infty$ as $d \to \infty$.

The classical regime can be sub-divided into *1-i)* d fixed, $n \to \infty$ (the classical large sample theory), and *1-ii)* $d, n \to \infty$, $d/n \to 0$. (e.g. $d(n) \asymp \log n$.) Likewise, the HDLSS situations can also be further categorized into *3-i)* $d, n \to \infty$, $d/n \to \infty$ (e.g. $n(d) \asymp \log d$), and *3-ii)* $d \to \infty$ while n is fixed (the HDLSS asymptotics).

We view all of these as informative. Which is most informative will depend on the particular data analytic setting, in the same way that either the Normal or Poisson approximation can be most informative about the Binomial distribution. The theoretical findings and messages from different categories sometimes overlap with each other, but many times the results are completely different. Our focus is on the HDLSS asymptotics, which gives mathematical convenience and also is relatively easy to interpret the message.

The term "HDLSS" was first coined by the seminal paper of Hall et al. (2005), which introduced the geometric representation of the HDLSS data (cf. Section 3.4). The asymptotic direction was first studied a couple of decades back in a different context; Casella and Hwang (1982) studied the Stein estimator of a mean vector with d growing and n fixed. Continued from the Hall et al. (2005), Ahn et al. (2007) explored general conditions which give the geometric representation of HDLSS data, as well as strong inconsistency of eigenvectors. Discussions in Chapters 2–3 can be seen as an improvement of the theory glimpsed in Ahn et al. (2007). Similar sets of problems concerning estimation of PCA and covariance matrices are also studied by Yata and Aoshima (2009, 2010a), whose work can be categorized into the HDLSS asymptotics. The classification or discriminant analysis is also often under investigation along the HDLSS asymptotic directions (Hall et al. (2005), Lee (2007), Huang et al. (2010) and Qiao et al. (2010)). In Pesarin and Salmaso (2010a) and Chapter 4.5 of Pesarin and Salmaso (2010b) interesting HDLSS works on nonparametric permutation tests are discussed.

Some researchers have considered the situations that d, n both grow but d grows much faster than n (i.e. the category *3-i*). For example, Bickel and Levina (2004) showed that Fisher linear discriminant analysis performs poorly when $d/n \to \infty$. Variable selection problems in high dimensional space are widely investigated and we refer to Fan and Lv (2010) who reviewed recent advances in the area.

In the random matrix theory, substantial work has been done on the asymptotic behavior of eigenvalues of the sample covariance matrix in the limit $n, d \to \infty$ together, see Baik et al. (2005), Johnstone (2001) and Paul (2007) for Gaussian assumptions and Baik and Silverstein (2006) for non-Gaussian results when $\frac{n}{d} \to c > 0$. See also Nadler (2008), Lee et al. (2010) and references therein for more recent advances. Majority of results are well described in a *spiked covariance model*, proposed by Johnstone (2001). An exception we point out is a work by El Karoui (2008), where he proposed to estimate the spectral distribution of eigenvalues without assuming a spike model.

4.2 Open problems

The theoretical work in the previous chapters is a foundation of further study. Some future research directions are following:

1. Extension of the results for generalized assumptions on covariance structures: For example, exponential or polynomial decays of eigenvalues may be more realistic. The basic covariance model in Chapters 2 and 3 is not so much different than a simple spike covariance model where the first eigenvalue is of order d^{α} and all the other eigenvalues are 1. If we assume that the sequence of the *tail* eigenvalues $\{\lambda_2, \dots, \lambda_d\}$ decrease exponentially, then the ϵ-condition does not hold anymore. However, it is natural to conjecture that a large first eigenvalue will make the estimation of the first eigenvector more accurate. Measuring how accurate the estimate is an important question, but is not yet investigated .

2. The study of eigen-analysis for covariance estimates which are not the standard sample covariance matrix. There are many regularized covariance estimates proposed in literature to avoid singularity of the matrix, but there is no unified framework for analyzing eigen-decomposition of such estimates. We suggest to use theories of matrix perturbation (Stewart and Sun (1990)), by which we achieve bounds on the discrepancy between eigenvalues (and eigenvectors) of the estimated and population covariance matrices.

3. *Time-varying PCA.* High dimensional data set is frequently observed along different time points (or on the space) longitudinally or cross-sectionally. Examples of such situations include time course gene expression data (see for example Storey et al. (2005), who analyzed in a variable by variable fashion), and brain MRI data taken longitudinally (Thambisetty et al. (2010)). While a nonparametric estimation of the mean function of time can be done by a standard technique, changes in the second moments (i.e. the covariance matrix function) are not investigated. Instead of estimating all elements of covariance matrix, one can pursue to estimate PCs smoothly that vary along time.

4. Estimation of the number of important components, or the intrinsic dimensionality. When reducing the dimensionality in practice, determining the number of components

is an important issue. Moreover, in HDLSS situation, this can be used to determine an appropriate sample size. This problem is heavily investigated in the machine learning community. We refer to Levina and Bickel (2005) who tackled this in the statistical side. Yata and Aoshima (2010b) investigated this issue in the view of the HDLSS context.

Part II

Statistics on Manifold

Chapter 5

Data on manifolds

In this chapter we introduce special forms of data, which lie in curved manifolds. Direct application of traditional statistics based on Euclidean space often fails to have legitimate answers. As a simple illustration, the arithmetic mean of the angles $2^\circ, 4^\circ, 356^\circ$, and 358°, as numbers, is 180° which is the opposite point of the intuitive notion of center which is 0°. The latter is easily justified by considering the angles as points on the unit circle, which is a 1-dimensional curved manifold in $\mathbb{R}^2$.

We begin with a list of interesting manifold data types (Section 5.1). Some preliminary mathematical background that is essential in dealing with the manifold data is provided in Section 5.2. We then discuss some exploratory statistics on manifolds in Section 5.3 as well as a general idea in extensions of principal component analysis to manifolds.

5.1 Manifolds of interest

A growing number of data types are non-Euclidean. In terms of *object oriented data analysis*, proposed by Wang and Marron (2007), there is a useful distinction between mildly and strongly non-Euclidean data, depending on the existence of approximating Euclidean spaces for the data space. The term *strongly non-Euclidean* refers mostly to tree-structured data (see e.g. Billera et al. (2001) for phylogenetic trees, Bullitt and Aylward (2002) for blood vessel trees, and Tschirren et al. (2002) for lung airway trees), where development of appropriate statistical methods is more challenging (see e.g. Aydin et al. (2009)). We focus on the *mildly non-Euclidean* data, which are also referred to as manifold data, as in that context, the data objects are on the surface of a curved manifold forming a feature space.

Although there are infinitely many manifolds with different topological features, we are only interested in certain types of manifolds that form sample spaces of real objects. This section contains a brief (not complete) list of those manifolds of interest, first labeled by their application area then by their geometric features.

Data on curved manifolds have long been investigated. Among these the following are best studied;

Directional data 2-D or 3-D directions lie in the unit circle or the unit sphere (or a hemisphere), which include wind or ocean current directions, orientation of cracks on rocks, and directions from the earth to celestial objects. A substantial amount of literature can be found in the area of circular, angular or directional statistics, see Fisher (1993), Fisher et al. (1993), and Mardia and Jupp (2000) for a good introduction and further discussion of this topic.

Landmark-based shape spaces A landmark based shape analysis analyzes data lying on special manifolds. A shape is defined as an equivalence class under translation and rotation, scaling in many cases and sometimes reflection. Thus, shape spaces are constructed by removing the translation, scale, and rotation from the set of landmarks, as proposed and investigated by both Kendall (1984; 1999) and Bookstein (1986; 1991). A clear and detailed introduction can be found in Dryden and Mardia (1998).

Due to advances in technology, a demand to analyze different types of manifold data is growing. These modern data are mostly from medical imaging and include

Medial shape representations Shapes of 2-D or 3-D objects are represented in a parametric model, called *m-reps* in short, including directions and angles as parameters. The data space here is a manifold that is a direct product of Euclidean space and unit spheres. This type of data intrigued the research presented in Chapter 6, and thus we illustrate the m-reps in more detail, later in Section 6.5. A complete discussion on the subject can be found in Siddiqi and Pizer (2008).

DTI Diffusion Tensor Magnetic Resonance Imaging (Basser et al. (1994), Pennec et al. (2006)) is a recently developed and widely studied MRI technique that measures the diffusion of water molecules in a biological object. Random motion of water molecules in 3-D for each voxel of an image is represented by a 3-D tensor, i.e. a symmetric positive definite 3×3 matrix. Similar to a variance-covariance matrix, each tensor lies in a lower dimensional sub-manifold of $\mathbb{R}^9$ since it has to be symmetric positive definite. DTI data, consisting of multiple tensors, thus naturally lie in a manifold.

Diffeomorphisms A common methodology for comparing shapes in image analysis is to use diffeomorphisms (Joshi and Miller (2000), Joshi et al. (2004)), i.e. smooth space warping functions. This method delivers a new approach to shape analysis. A shape is considered as a distortion (i.e. diffeomorphism) of some template. Thus a set of shapes is represented as a set of diffeomorphisms and the variation in the population of diffeomorphisms can be studied to understand variation in shapes. The set of diffeomorphisms forms a very high dimensional manifold.

Shapes of curves A modern approach to analyze space curves, outlines of objects, and surfaces is to consider shapes of those, invariant to similarity transformations and reparametrization of the curves. The space of these invariant shapes form an infinite dimensional manifold that can be thought of as an extension of the landmark-based shape spaces. See e.g. Srivastava et al. (2010).

The list provided here contains most common manifolds of interest. To develop statistical methods on these different manifolds, it is helpful to classify the manifolds of interest by their topological features, as follows.

1. *Hyperspheres* are related to many types of manifolds yet easy to work on. For example, the shape space of landmarks or curves is a quotient space of hyperspheres under some actions of groups. A widely used technique in shape analysis is to lift (i.e. approximate) samples in shape space to its corresponding hypersphere and to work on the much simpler manifold, the sphere.

2. *Direct product manifolds* are identified with direct (Cartesian) products of other simple manifolds. For example, a torus can be understood as a direct product of two circles ($S^1 \times S^1$). Interesting data types are found in forms of generalized tori and direct

products of spheres and Euclidean spaces. The space of m-reps and the size-and-shape space (see Ch. 8 of Dryden and Mardia (1998)) belong to the class of direct product manifolds.

3. Space of *Symmetric Positive Definite matrices* is topologically a cone, and thus has different characteristics than spheres. Loosely speaking, the space is a generalization of the positive real line. The sample spaces of the diffusion tensors and the covariance matrices (as observed samples) belong to this category.

4. *Implicitly defined manifolds.* Some manifolds are only defined implicitly (i.e. through tangent spaces and metrics), an example of which is the space of diffeomorphisms. Distances and geodesics (an analog of lines) between two objects are found only through heavy optimization processes.

Throughout the discussion in the dissertation, we focus on direct product manifolds and spheres (and quotient spaces of those), in Chapters 6 and 7, respectively. Implementation of the proposed methods to other types of manifolds would be interesting and useful future work.

As an aside, Chikuse (2003) contains a detailed introduction to the statistics on special manifolds, including the Stiefel and Grassmann manifolds.

5.2 Mathematical background

5.2.1 Riemannian manifold

We begin with introducing well-known definitions and facts from differential geometry. Precise definitions and geometric discussions on the manifold can be found in Boothby (1986); Lee (1997); Helgason (2001).

A d-dimensional manifold can be thought of as a curved surface embedded in a Euclidean space of higher dimension d' ($\geqslant d$). The manifold is required to be smooth, i.e. infinitely differentiable, so that a sufficiently small neighborhood of any point on the manifold can be well approximated by a linear space. The *tangent space* at a point p of a manifold M, T_pM, is defined as a linear space of dimension d which is tangent to M at p. The notion of distance on manifolds is handled by a Riemmanian metric, which is a metric of tangent spaces.

Definition 5.1. A *Riemmanian metric* on a manifold M is a function that smoothly assigns to each point $p \in M$ an inner product $\langle \cdot, \cdot \rangle$ defined on the tangent space T_pM. A *Riemmanian manifold* is a smooth manifold equipped with such a Riemmanian metric.

Given a smooth curve segment $\gamma : [a, b] \longrightarrow M$, the length of γ can be defined by the Riemmanian metric as

$$L(\gamma) = \int_a^b \|\gamma'(t)\| \, dt,$$

where $\gamma'(t)$ lies in the tangent space at $\gamma(t)$ and $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$. In contrast to the Euclidean space, the shortest path between two points is in general not a line segment but a curve. The shortest path is closely related to the concept of *geodesics*, whose trajectory is (locally) the shortest path between two points. Geodesics are formally defined as follows.

Definition 5.2. A *geodesic* is a constant-velocity curve γ that locally minimizes $L(\gamma)$ among curves whose endpoints are fixed.

Note that a geodesic between two points may not be unique. Furthermore a geodesic may not be the same as the global shortest path. A natural metric on manifolds, the *geodesic distance* function $\rho(p, q)$, is the length of shortest geodesic segment between two points. In case there exists no such geodesic, ρ is defined by the infimum of lengths of all paths from p to q. The geodesic distance is formally defined as follows.

Definition 5.3. The *geodesic distance* function $\rho(p, q)$ is defined as the infimum of the lengths of all smooth curves γ between p and q, i.e. $\rho(p, q) = \inf_\gamma L(\gamma)$.

A geodesic γ is said to be a *minimal geodesic* if $L(\gamma) = \rho(p, q)$. Minimal geodesics are guaranteed to exists under certain conditions as stated in the following (Theorem 7.7 and Lemma 7.8 of Boothby (1986))).

Theorem 5.1 (Hopf-Rinow). *For a manifold M, The following are equivalent;*

(i) Any geodesic $\gamma : [a, b] \longrightarrow M$ can be extended to a geodesic from $\mathbb{R}$ to M.

(ii) With the metric $\rho(p, q)$, (M, ρ) is a complete metric space.

Furthermore, if (i) or (ii) is true, any point p of M can be joined to $q \in M$ by a minimal geodesic (whose length is necessarily $\rho(p, q)$).

In what follows we assume that our manifold M is a complete Riemmanian manifold with a metric ρ. Since the formulation of the Riemmanian manifold (and its metric) is closely related to its tangent spaces, mappings between the manifold and a tangent space are useful.

For any $p \in M$ and $v \in T_pM$, there exists a unique geodesic γ whose initial location is p with initial direction v, i.e. $\gamma(0) = p$ and $\gamma'(0) = v$ (Thoerem 5.8 of Boothby (1986)). If (M, ρ) is complete, then the geodesic γ is defined for all of $\mathbb{R}$. Now we can define the *exponential map* Exp_p, a mapping from T_pM to M, as

$$\mathrm{Exp}_p(v) = \gamma(1),$$

where γ is the unique geodesic defined by v. The exponential map is carefully chosen so that a geodesic in M through p is an image of a straight line in T_pM through the origin of T_pM. The exponential map Exp_p is a diffeomorphism (a smooth bijective function whose inverse is also smooth) in some neighborhood $U \in T_pM$ (Theorem 6.9 of Boothby (1986)). Thus Exp_p has an inverse at least in some neighborhood of $Exp_p(U)$. Denote the inverse exponential map as $\mathrm{Log}_p : \mathrm{Exp}_p(U) \longrightarrow T_pM$, sometimes called the *log map*.

The geodesic distance ρ of M can be formulated as the norm of a tangent vector by use of the exponential map.

Theorem 5.2. *For any two points p, q of a complete Riemmanian manifold M, there exists a tangent vector v of T_pM such that $Exp_p(v) = q$ and*

$$\rho(p, q) = \|v\| .$$

Proof. Since M is complete, there exists a minimal geodesic γ between p and q which can be parametrized as $\gamma(0) = p$ and $\gamma(1) = q$. Let $v = \gamma'(0) \in T_pM$. By the definition of the exponential map, $\mathrm{Exp}_p(v) = q$. Noting that $\|\gamma'(t)\| = \|\gamma'(0)\|$ for all $t \in \mathbb{R}$ because of its constant-velocity property, we have

$$\rho(p,q) = L(\gamma) = \int_0^1 \left\|\gamma'(t)\right\| dt = \int_0^1 \|v\| \, dt = \|v\| \, . \qquad \square$$

Remark 5.1. For the tangent vector v in Theorem 5.2, the unique geodesic γ defined by v is minimal.

Note that by the theorem, if q is in the domain of Log_p, then $\mathrm{Log}_p q = v$ so that $\rho(p,q) = \left\|\mathrm{Log}_p q\right\|$.

Remark 5.2. In many practical applications, a manifold can be represented as a Lie group or a quotient space of a Lie group. In this case, the exponential map is easily found by Lie algebra. We do not restate general Lie theory, but mention earlier work: Boothby (1986) and Fletcher (2004).

5.2.2 Direct Product manifold

Simple manifolds include some low dimensional manifolds such as S^1, S^2, $\mathbb{R}_+$, and $\mathbb{R}$. A direct product manifold, consisting of the direct product of simple manifolds, can be thought of as a metric space. In this section, a geodesic distance, an appropriate metric for the direct product manifold, is established.

In this section a direct product manifold is defined as a complete metric space with the geodesic distance as its metric. We begin by illustrating the main idea using simple manifolds, since they form a basis for direct product manifolds.

[$\mathbf{S^1}$] The unit circle S^1 is a complete 1-dimensional manifold embedded in $\mathbb{R}^2$. The geodesic distance is defined by the length of the shortest arc. Let $\theta \in \mathbb{R}$ denote an element of T_pS^1 where p is set to be $(1,0) \in S^1$. Then the exponential map is defined as

$$\mathrm{Exp}_p(\theta) = (\cos\theta, \sin\theta).$$

The corresponding log map of $x = (x_1, x_2)$, $\mathrm{Log}_p(x) = \mathrm{sign}(x_2) \cdot \arccos(x_1)$, is defined on $S^1/\{-p\}$.

Note that S^1 can be thought of as a manifold embedded in a complex plane. Then the exponential map at $p = 1 + 0i$ is just a complex exponential, i.e. $\text{Exp}_p(\theta) = e^{i\theta}$.

[$\mathbf{S^2}$] The unit sphere S^2 is a complete 2-dimensional manifold embedded in $\mathbb{R}^3$. The geodesics at the north pole $p = (0, 0, 1)$ are the great circles passing through p. The geodesic distance between p and $q \in S^2$, $\rho(p, q)$ is the length of the shortest great circle segment and is computed by

$$\rho(p, q) = \arccos(< p, q >),$$

where $<, >$ is the Euclidean inner product. Let $v = (v_1, v_2)$ denote a tangent vector in T_pS^2. Then the exponential map $\text{Exp}_p : T_pS^2 \longrightarrow S^2$ is defined by

$$\text{Exp}_p(v) = \left(\frac{v_1}{\|v\|} \sin \|v\|, \frac{v_2}{\|v\|} \sin \|v\|, \cos \|v\| \right).$$

This equation can be understood as a rotation of the base point p to the direction of v with angle $\|v\|$. The corresponding log map for a point $x = (x_1, x_2, x_3) \in S^2$ is given by

$$\text{Log}_p(x) = \left(x_1 \frac{\theta}{\sin \theta}, x_2 \frac{\theta}{\sin \theta} \right),$$

where $\theta = \arccos(x_3)$ is the geodesic distance from p to x. Note that the antipodal point of p is not in the domain of the log map, i.e. the domain of Log_p is $S^2/\{-p\}$.

[$\mathbb{R}_+$] The set of positive real numbers $\mathbb{R}_+$ needs special treatment. In many practical applications, $\mathbb{R}_+$ represents a space of scale parameters. A desirable property for a metric on scale parameters is the *scale invariance*, $\rho(rx, ry) = \rho(x, y)$ for any $x, y, r \in \mathbb{R}_+$. This can be achieved by differencing the logs, i.e.

$$\rho(x, y) = \left| \log \frac{x}{y} \right|, \quad \text{for } x, y \in \mathbb{R}_+. \tag{5.1}$$

It can be checked that the metric ρ is a complete geodesic distance function. Let geodesics of $\mathbb{R}_+$ be of the form $\gamma(t) = ae^{bt}$, for $a \in \mathbb{R}_+$, $b \in \mathbb{R}$ so that it can be extended for all $t \in \mathbb{R}$. The geodesic distance (5.1) is given by a Riemmanian metric at $p \in \mathbb{R}_+$, $\langle v, w \rangle_p = \frac{v \cdot w}{p^2}$. One may verify that a minimal geodesic between $x, y \in \mathbb{R}_+$ is $\gamma(t) = xe^{t \log \frac{y}{x}}$. Then $L(\gamma) = \int_0^1 \|\gamma'(t)\|_{\gamma(t)} \, dt = \int_0^1 \left|\log \frac{y}{x}\right| dt$ implies (5.1). The geodesic distance ρ is a complete metric by Theorem 5.1.

Note that another metric $d(x, y) = |x - y|$ on $\mathbb{R}_+$ is possible but is neither scale invariant nor complete. The exponential map of $\mathbb{R}_+$ is defined by the standard real exponential function. The domain of the inverse exponential map, the log map, is $\mathbb{R}_+$ itself.

[$\mathbb{R}$] The real line is a linear space. The exponential map on $\mathbb{R}$ is the identity map. The geodesic distance on $\mathbb{R}$ is the same as the Euclidean distance.

Note that the result of Theorem 5.2 can be applied to each of four simple manifolds. However, its major use for our development is on direct product manifolds, which will be stated shortly.

The *direct product manifold* is defined by the direct product of simple manifolds. Let $M = \otimes_{i=1}^{m} M_i$ be a d ($\geqslant m$) dimensional manifold, where $\otimes$ denotes the direct product of spaces. Let each M_i be one of the simple manifolds, that is, $M_i = S^1$, S^2, $\mathbb{R}_+$, or $\mathbb{R}$. The tangent space at $p = (p_1, \ldots, p_m)$ of M, T_pM is the d-dimensional vector space. Let $v = (v_1, \ldots, v_m)$, $w = (w_1, \ldots, w_m)$ be tangent vectors of T_pM. Define the Riemmanian metric at p by the inner product

$$\langle v, w\rangle_p = \langle v_1, w_1\rangle_{p_1} + \cdots + \langle v_m, w_m\rangle_{p_m}, \tag{5.2}$$

where each $\langle\cdot,\cdot\rangle_{p_i}$ is the Riemmanian metric at p_i of M_i.

Then geodesics between $x, y \in M$ are denoted as $\gamma(t) = (\gamma_1(t), \cdots, \gamma_m(t))$ where each $\gamma_i(t)$ is the geodesic between x_i and y_i. Since each $\gamma_i(t)$ is defined for all $\mathbb{R}$, $\gamma(t)$ is also defined for all $\mathbb{R}$, and thus by Theorem 5.1, M with the geodesic distance ρ is complete. The exponential map at $p \in M$ is defined by the exponential maps of each component, i.e.

$$\mathrm{Exp}_p(v) = (\mathrm{Exp}_{p_1}(v_1), \ldots, \mathrm{Exp}_{p_m}(v_m)), \quad v \in T_pM.$$

Likewise, the log map is defined as $\mathrm{Log}_p(x) = (\mathrm{Log}_{p_1}(x_1), \ldots, \mathrm{Log}_{p_m}(x_m))$, for x in the domain of Log_p, $D = D_1 \times \cdots \times D_d$, where D_i is the domain of the log map of M_i, Log_{p_i}.

By Theorem 5.2 and (5.2), for any $x, y \in M$, there exists a tangent vector $v \in T_xM$ such that $\mathrm{Exp}_x(v) = y$ and

$$\rho(x, y) = \|v\|_x = \left(\sum_{i=1}^{m} \|v_i\|_{x_i}^2\right)^{1/2}, \tag{5.3}$$

where the norm $\|\cdot\|_{x_i}$ on M_i is defined by the Riemmanian metric $\langle\cdot,\cdot\rangle_{x_i}$. Moreover, there exists a minimal geodesic $\gamma = (\gamma_i, \ldots, \gamma_m)$ associated with v. By using the constant-velocity property of geodesics, one may prove that each γ_i is also minimal. By a similar argument to the proof of Theorem 5.2, the geodesic distance between x_i, y_i of M_i is

$$\rho_i(x_i, y_i) = L(\gamma_i) = \int_0^1 \left\|\gamma_i'(t)\right\| dt = \int_0^1 \|v_i\|_{x_i} dt = \|v_i\|_{x_i}. \tag{5.4}$$

Using (5.4), the geodesic distance on M, defined by (5.3), can be written as

$$\rho(x, y) = \left(\sum_{i=1}^{m} \rho_i^2(x_i, y_i)\right)^{1/2}, \quad x, y \in M, \tag{5.5}$$

where each ρ_i is the geodesic distance function of M_i. The direct product manifold M with the metric ρ is a complete metric space.

5.3 Exploratory Statistics on manifolds

5.3.1 Extrinsic and Intrinsic means

Finding a center point that best represents the data is a natural starting point of a statistical analysis. In general there are two different notions of mean on manifolds. One of these is popular in directional statistics, which treats S^{d-1} as a subset of $\mathbb{R}^d$ (See e.g. Fisher et al. (1993), Mardia and Jupp (2000)). As an illustration, a set of angles $\{\theta_1, \ldots, \theta_n\}$ is represented as a set of vectors of length 1, i.e. $\vec{\theta} = (\cos\theta, \sin\theta)$. The mean of these angles is defined by the direction of the addition of the vectors, i.e.

$$\tilde{\theta} = \frac{\sum_{i=1}^{n} \vec{\theta}_i}{\left\|\sum_{i=1}^{n} \vec{\theta}_i\right\|}. \tag{5.6}$$

Note that the arithmetic mean of these vectors, $\frac{1}{n}\sum_{i=1}^{n} \vec{\theta}_i$, may not lie on the unit circle. So it is divided by its length to project back to the unit circle. This notion of mean can be generalized for general manifolds as follows. Given a d-dimensional manifold M, let $\mathbb{R}^{d_0}$ be the d_0-dimensional Euclidean space in which M is embedded. Define a distance function ρ_0 on M as $\rho_0(x, y) := \|x - y\|$, where $\|\cdot\|$ is the usual norm of $\mathbb{R}^{d_0}$. This definition of distance

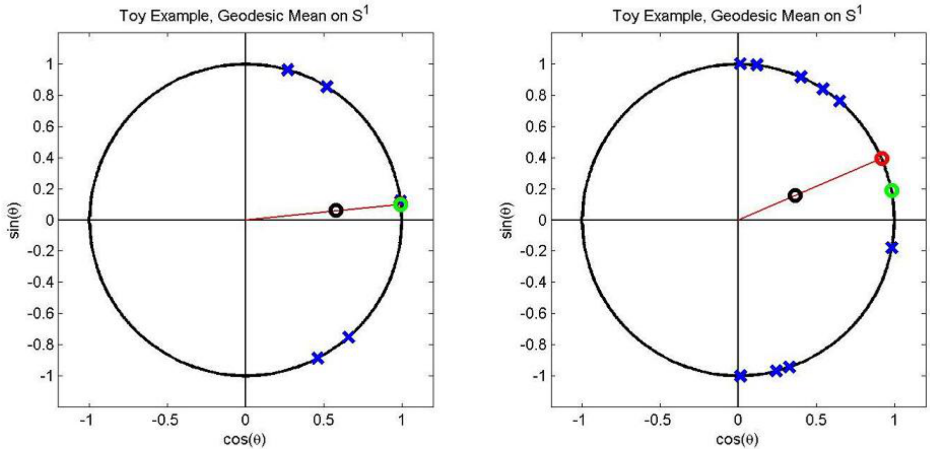


Figure 5.1: Illustration of the extrinsic and intrinsic means of the angular data (depicted as blue xs). The means may be the same as shown in the left panel, but in general different as shown in the right panel. It also shows that the simple average (black circle) leads to a point that is not on the circle, but the projection of the average onto the manifold is the extrinsic mean (red circle). The intrinsic mean is the green circle.

is extrinsic to M, that is, it depends on the choice of embedding space. Then, define the *extrinsic mean* of a set of points $x_1, \ldots, x_n \in M$ as

$$\tilde{x} = \operatorname*{argmin}_{x \in M} \sum_{i=1}^{n} \rho_0^2(x, x_i). \tag{5.7}$$

The extrinsic mean can also be obtained by finding the arithmetic mean of the embedded points (in $\mathbb{R}^{d_0}$) and then projecting this mean onto the manifold M (Srivastava and Klassen (2002)). Specifically, when the manifold is taken to be S^1, this definition is the same as (5.6).

A more natural choice of mean is obtained by using the geodesic distance ρ which is intrinsic to M. Define the *geodesic mean* of $x_1, \ldots, x_n \in M$ as the minimizer in M of the sum of squared geodesic distances to the data. Thus the geodesic mean is defined as

$$\bar{x} = \operatorname*{argmin}_{x \in M} \sum_{i=1}^{n} \rho^2(x, x_i), \tag{5.8}$$

which is often called the *intrinsic mean.*

The idea of intrinsic mean comes from that of the Fréchet mean (Fréchet (1944), Fréchet (1948), Karcher (1977), Le and Kume (2000), Bhattacharya and Patrangenaru (2003)), named after the famous mathematician Maurice Fréchet. The Fréchet mean generalizes the notion of the arithmetic mean of a vector space, preserving the property of minimizing the sum of squared distances to data. In a vector space, the Fréchet mean with the usual Euclidean distance is the same as the arithmetic mean. In a manifold, the Fréchet mean with its distance as the geodesic distance ρ is the geodesic mean. The existence and uniqueness of the geodesic mean is not in general guaranteed. However, Karcher (1977) showed that the geodesic mean exists and is unique when the given data are contained in a sufficiently small region of M. In case there are multiple geodesic means, the set of those will be called a *geodesic mean set*.

Note that those definitions of extrinsic and intrinsic means apply only to finite samples. We can also define extrinsic and intrinsic means of a distribution by substituting the sum by integration from (5.7) and (5.8), i.e. the extrinsic mean

$$\mu_E = \operatorname*{argmin}_{p \in M} \int_M \rho_0^2(p, x) dF(x),$$

where F denotes a distribution function on M, and the intrinsic mean

$$\mu_I = \operatorname*{argmin}_{p \in M} \int_M \rho^2(p, x) dF(x).$$

The extrinsic mean and the intrinsic mean are not in general the same, an example of which can be found in Srivastava and Klassen (2002). Bhattacharya and Patrangenaru (2003, 2005) studied the asymptotic properties of the sample means (5.7) and (5.8) when the sample size tends to infinity, including consistency and limiting distributions.

The extrinsic mean is easier and faster to compute and sometimes is used as an approximation of the intrinsic mean in practice (Bhattacharya and Patrangenaru (2003); Srivastava and Klassen (2002)). On the other hand, the intrinsic mean is widely used in the field of image analysis since any embedding is thought of as unnatural choice.

Figure 5.1 exemplifies the extrinsic and intrinsic means of angular data. A detailed investigation on the intrinsic mean is presented next.

5.3.2 Examples: Geodesic Means on Various Manifolds

Evaluating the geodesic mean on a manifold is an optimization problem, and usually there is no closed form solution. In this section aspects of evaluating geodesic means are discussed in the case of the direct product manifold. Evaluating geodesic means of data in S^1 or S^2, i.e.

directional data, are illustrated at first.

Geodesic mean of directional data

We begin with investigating geodesic means of directional data, particularly in S^1 or S^2. These manifolds are considered as spaces of angles, 2-D or 3-D directions.

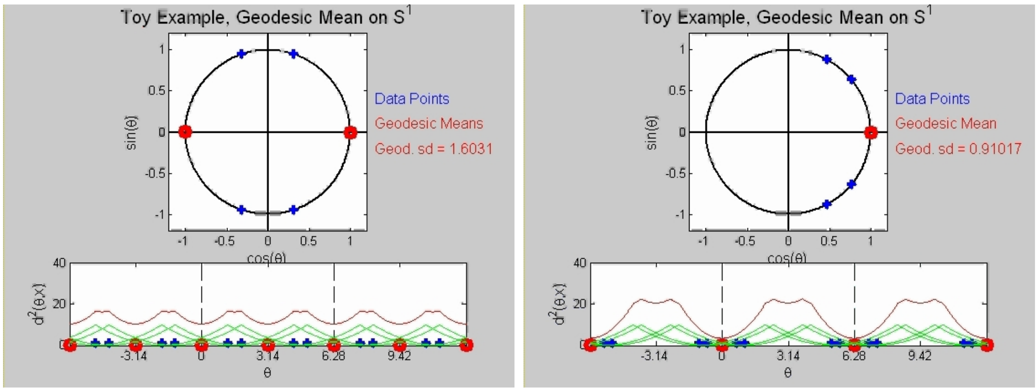
The unit circle S^1 can be considered as a wrapping of the real line such that θ and $\theta+2\pi \in \mathbb{R}$ are wrapped to the same location of the unit circle. The wrapping of the real line coincides with the exponential map from the tangent space as defined in section 5.2.2. Conversely for a given set of data points on S^1, it is possible to unwrap the unit circle to the real line so that there are infinitely many periodic copies of the data points. Figure 5.2 illustrates the procedure of finding the geodesic mean set via the wrapping approach. The top panel illustrates the unit circle, or S^1 embedded in $\mathbb{R}^2$, together with data points $\{x_1, \ldots, x_4\} \subset S^1$ as blue crosses. The unwrapped circle and data points (with replications) are described in the bottom panel. The bottom panel also shows the squared geodesic distances to each of the data points, $\rho^2(x, x_i)$, as green curves and the sum of squared distances $\sum_{i=1}^{4} \rho^2(x, x_i)$ as a red curve. By (5.8) the geodesic mean set (red dots) is found where the red curve is minimized. Figure 5.2a shows a particular example with multiple geodesic means. Figure 5.2b shows a particular example with a local minima of the object function, i.e. the red curve, that we want to minimize. It is clear that iterative algorithms to find geodesic means may fail because of these two cases.

A direct calculation for a geodesic mean set of angles, first developed by Moakher (2002) as an example of the geodesic mean of rotation matrices, can be used as a remedy for these problems. He showed that for a set of angles $\{\theta_1, \ldots, \theta_n\}$ the geodesic mean set is contained in the candidate set $\{\bar{\theta}_j\}_{j=0}^{n-1}$ where

$$\bar{\theta}_j = \frac{\sum_{i=1}^{n} \theta_i + 2j\pi}{n}, \quad j = 0, \ldots, n-1. \tag{5.9}$$

A simple proof of this can be found at Lu and Kulkarni (2008). Moreover, all local minima of the object function are found at the candidate set.

The movie `Geodmean2d.avi` at Jung (2008) describes aspects of the geodesic mean set with four data points moving around the circle.



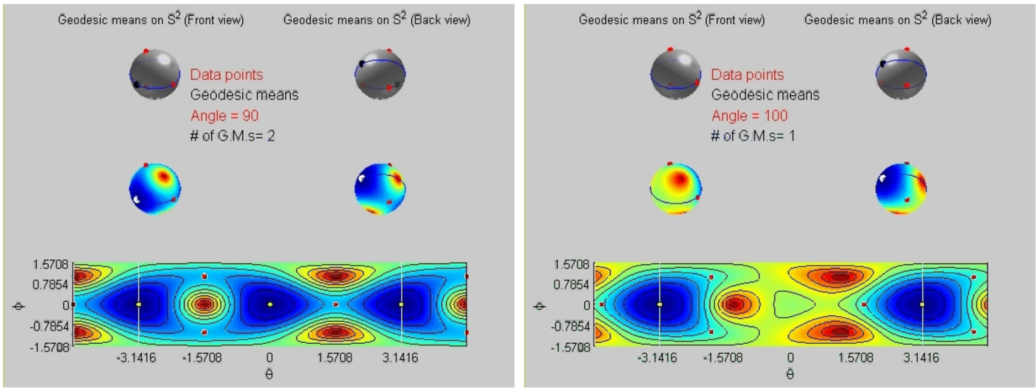
(a) Not unique geodesic mean

(b) local minima

Figure 5.2: Geodesic mean set on S^1. Four blue crosses (+) represent data points. The red curve in the bottom panel shows the sum of squared distances to the data. A geodesic mean set (shown as red dots) is found where the red curve is minimized. Examples are chosen to show (a) multiple geodesic means and (b) local minima of the object function.

Let us move on to geodesic means on S^2. One way to parameterize S^2 is by a spherical coordinate system using the longitude θ and the latitude ϕ. That is, a point on S^2 is represented by $(\theta, \phi) \in [0, 2\pi) \times [\frac{\pi}{2}, -\frac{\pi}{2}]$. Note that in this parametrization, the poles are each represented by infinitely many points, e.g. the north pole is represented by $(\theta, \frac{\pi}{2})$ for any $\theta \in [0, 2\pi)$. Similar to the wrapping approach in the S^1 case, consider unwrapping S^2 transversely (i.e., along the equator) so that (θ, ϕ) and $(\theta + 2\pi, \phi)$ are located at the same point on the sphere. These ideas are used to illustrate the procedure of finding a geodesic mean set on S^2 as in Figure 5.3. The top left and right show a sphere with its equator and three data points as red dots, viewed from the front and back sides. The bottom panel illustrates a part of the transversely unwrapped sphere with the longitude-latitude parameters. Note that a geodesic of S^2 is still a curve (not a line) in the parameter space. Thus a geodesic distance is not the same as the Euclidean distance in the parameter space. This parameter space does not agree with the tangent space defined in Section 5.2.2. Nevertheless, the sum of squared distances to the data, $\sum \rho^2(x_i, x)$, can be evaluated and illustrated based on the parameters. Contours of the object function (i.e. $\sum \rho^2(x_i, x)$) together with a heat map (high as red, low as blue) are plotted in the bottom panel. The geodesic mean set is found where the object function is minimized. The middle panel shows the sphere colored by the heat map of the object function. For more examples, movies at Jung (2008) illustrate various aspects of geodesic means of S^2.

Finding a geodesic mean set is an optimization problem. Similar to the S^1 case, a geodesic



(a) Not unique geodesic mean (b) local minima

Figure 5.3: Geodesic mean set on S^2. Three red points represent the location of data on the unit sphere. The heat map and contours of the object function evaluated on the latitude-longitude parameter space are illustrated in the bottom panel. The geodesic mean set (shown as yellow dots) is found where the object function is minimized. Examples are chosen to show (a) multiple geodesic means and (b) local minima of the object function.

mean of S^2 may not be unique (Figure 5.3a). Moreover, it is possible to have infinitely many geodesic means. For example, the geodesic mean set of two points in the north and south poles is the whole equator. Also, the object function that we want to minimize might have local minima (Figure 5.3b), which suggests that iterative algorithms may fail to find geodesic means.

Remark 5.3 (Uniqueness of geodesic mean in S^1 and S^2). As mentioned earlier, the uniqueness of a geodesic mean on manifolds is guaranteed when the support of data is sufficiently small. Specifically, Theorem 7.3 of Kendall (1990), also found in Karcher (1977) and Le (2001), applied to the case S^1 and S^2 gives the following. When the data in S^1 are contained in an arc which is a strict subset of the half circle, a geodesic mean is unique and found in the arc. Similarly, if the data in S^2 are contained in a geodesic ball $B(p, \delta) = \{x \in S^2 : \rho(x, p) \leqslant \delta\}$, for some $\delta < \frac{\pi}{2}$ so that $B(p, \delta)$ is strictly contained in a hemisphere, then there exists a unique geodesic mean of the data, contained in $B(p, \delta)$.

Extension to direct product manifolds

In the previous section we have focused on rather simple cases, S^1 and S^2. These special cases in fact form a fundamental basis for direct product manifolds. Finding a geodesic mean set on the direct product manifolds heavily relies on the geodesic mean set of each of the simple manifolds.

We begin by illustrating this idea on a 2-dimensional manifold $S^1 \times S^1$. For $x \in S^1 \times S^1$, denote $x = (x^1, x^2)$, where $x^1, x^2 \in S^1$. Let ρ_1 be the geodesic distance function defined on S^1. Then from the result of Section 5.2.2 the geodesic distance function on $S^1 \times S^1$ is defined as

$$\rho(x, y) = \sqrt{\rho_1^2(x^1, y^1) + \rho_1^2(x^2, y^2)}, \quad x, y \in S^1 \times S^1.$$

Now consider finding a geodesic mean set of $x_1, \ldots, x_n \in S^1 \times S^1$, where $x_i = (x_i^1, x_i^2)$, $i = 1, \ldots, n$. Let A_1, A_2 be the *component-wise geodesic mean sets*, i.e. the geodesic mean sets of $\{x_1^1, \ldots, x_n^1\}$ and $\{x_1^2, \ldots, x_n^2\}$ respectively. Then the geodesic mean set of $x_1, \ldots, x_n \in S^1 \times S^1$, say A, will be the direct product of the component-wise mean sets, i.e.

$$A = \{\underset{x \in (S^1)^2}{\operatorname{argmin}} \sum_{i=1}^{n} \rho^2(x, x_i)\} = A_1 \times A_2.$$

Figure 5.4 depicts a particular example of a geodesic mean set on $S^1 \times S^1$ with two data points

$$x_1 = \begin{pmatrix} 0 \\ \\ \pi \end{pmatrix}, \quad x_2 = \begin{pmatrix} \pi \\ \\ \frac{\pi}{2} \end{pmatrix} \in S^1 \times S^1,$$

where each component of S^1 is parametrized by the angle from the positive x-axis. The geodesic mean sets of the first and second components are $A_1 = \{\frac{\pi}{2}, \frac{3}{2}\pi\}$, $A_2 = \{\frac{3}{4}\pi\}$ respectively. The geodesic mean set $A = \{(\frac{\pi}{2}, \frac{3}{4}\pi), (\frac{3}{2}\pi, \frac{3}{4}\pi)\} \subset S^1 \times S^1$ can be found via minimizing the sum of squared distances, which is in fact the same as the set $A_1 \times A_2$.

Noting that S^2 is also a 2-dimensional manifold, we may hope that we could break further down the geodesic mean set of S^2 as done in $S^1 \times S^1$. However, the component-wise mean of points on S^2 does not in general give the geodesic mean. As an illustration, consider data points along the equator as in Figure 5.5. If we parameterize S^2 with the longitude-latitude parameters as done in previous section, then the component-wise mean of latitude angles is near zero. However, the geodesic mean is located near the north pole, which implies that the latitude angle of the geodesic mean is near $\frac{\pi}{2}$ (far from zero).

The component-wise mean idea illustrated, in Figure 5.4, on $S^1 \times S^1$ is in fact generalizable for more complicated direct product spaces. The following proposition shows that the geodesic

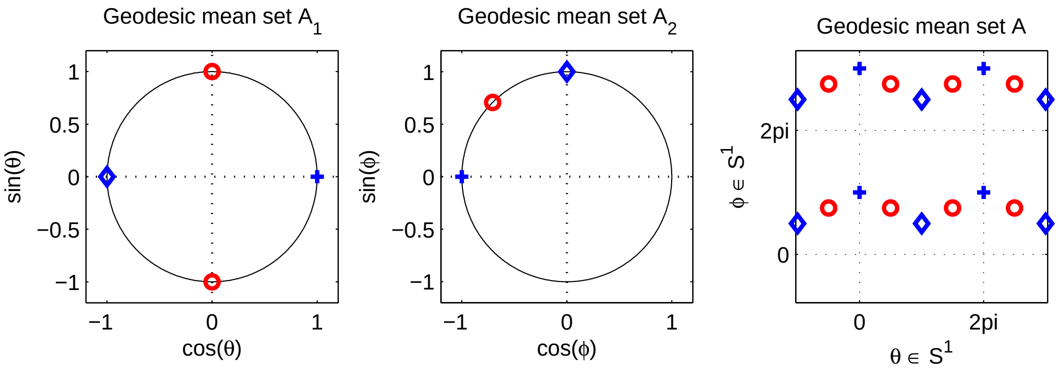


Figure 5.4: Geodesic mean set on $S^1 \times S^1$. Two data points are depicted as blue cross(+) and blue diamond (⋄) signs, respectively. The left and middle panels show the component-wise geodesic mean sets A_1 and A_2 (red circles) as well as the data for each component. The right panel is a part of the parametrization of $S^1 \times S^1$ via unwrapping and shows replicated data points and two geodesic means. It is easily seen that $A = A_1 \times A_2$.

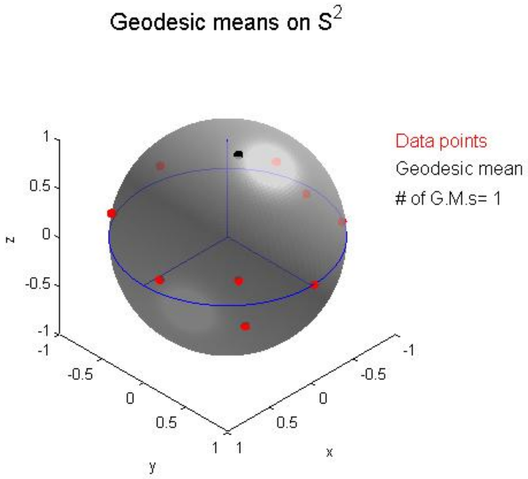


Figure 5.5: Plot of 10 points (red dots) along the equator with random perturbation and the geodesic mean (black dot) near the north pole illustrates that geodesic means on S^2 can not be calculated in the component-wise fashion, and that the geodesic mean may be located far from all data points.

mean set of data in a direct product manifold is the same as the direct product of component-wise geodesic means.

Proposition 5.3. *Let* $M = \otimes_{i=1}^{m} M_i$ *be a* $d(\geqslant m)$ *dimensional manifold. Let each* M_i *be a simple manifold as defined in Section 5.2.2. Let* $\{x_1, \dots x_n\}$ *be* n *points in* M, *where* $x_j = (x_j^1, \dots, x_j^m)$ *for each* $j = 1, \dots, n$. *Suppose* A_i *is the component-wise geodesic mean set of* $x_1^i, \dots, x_n^i \in M_i$ *for each* i. *Then the geodesic mean set of* $x_1, \dots x_n \in M$, *say* A, *is the same as the direct product of component-wise geodesic means, i.e.*

$$A = \otimes_{i=1}^{m} A_i.$$

Proof. Let ρ_i be the geodesic distance function defined on M_i respectively. Then from the result of Section 5.2.2 the geodesic distance function ρ_M on M is defined as

$$\rho_M(x_1, x_2) = \left(\sum_{i=1}^{m} \rho_i^2(x_1^i, x_2^i) \right)^{1/2}, \quad x_1, x_2 \in M. \tag{5.10}$$

For any $\bar{x} = (\bar{x}^1, \dots, \bar{x}^m) \in A$, we have by definition of the geodesic mean,

$$\sum_j \rho_M^2(\bar{x}, x_j) \leqslant \sum_j \rho_M^2(y, x_j), \quad \forall y \in M.$$

Plugging in (5.10) on both sides, we get for all $y^1 \in M_1$,

$$\sum_j \rho_1^2(\bar{x}^1, x_j) + \sum_{i>1} \sum_j \rho_i^2(\bar{x}^i, x_j) \leqslant \sum_j \rho_1^2(y^1, x_j) + \sum_{i>1} \sum_j \rho_i^2(\bar{x}^i, x_j),$$

which shows that $\bar{x}^1 \in A_1$. Since this is true for all $\bar{x}^i$, we get $A \subset \otimes_{i=1}^{m} A_i$. It can be proved similarly that $A \supset \otimes_{i=1}^{m} A_i$. □

Remark 5.4. The uniqueness of a geodesic mean on M is guaranteed by the uniqueness of every component-wise geodesic mean of M_i. Thus sufficiently small support of a data set in each M_i will give a unique geodesic mean. Note that a geodesic mean for $\mathbb{R}$ or $\mathbb{R}_+$ is always unique. For S^1 and S^2, see Remark 5.3. If there exist multiple geodesic means for each simple manifolds M_i, then the cardinality of the geodesic mean set of M grows exponentially.

5.4 Backward Generalization of PCA on manifolds

The rest of the dissertation is devoted to generalize Principal Component Analysis (PCA) to some types of manifold data. A main idea in generalization is to take a reverse viewpoint in understanding Euclidean PCA.

In Euclidean space, or simply a vector space of dimension d, let $X_1, \ldots, X_n$ be column vectors that are inputs for Classical (Euclidean) PCA. The data matrix is formed by aggregating the data vectors: $\mathbf{X} = [X_1, , \ldots, X_n]$. Euclidean PCA can be understood as an operation of finding affine subspaces AS^i, where $i = 1, \ldots, d$ represents the dimension of AS^i.

A traditional *forward stepwise* view to Euclidean PCA is understood by increasing the dimension i of AS^i, starting from the empirical mean $\bar{X} \equiv AS^0$. In particular, given AS^i, the direction $\vec{u}_{i+1}$ of great variance is added to AS^i, resulting in AS^{i+1}. Therefore, we have

$$AS^0 \subset AS^1 \subset AS^2 \subset \cdots \subset AS^d,$$

where each AS^i is the best fit containing AS^{i-1} in the whole space AS^d. A simple example of the forward operation in depicted in Figure 5.6. In 3-space, $\bar{X}$ is plotted as a black dot with the AS^1 drawn as a line segment. AS^2 is found by adding an orthogonal direction to AS^1, resulting in an affine plane AS^2 plotted in the right panel.

The viewpoint that seems more useful for generalization of PCA to manifold data is the *backward stepwise* view. In backward PCA, principal components are found in reverse order, i.e., AS^is are fitted from the largest dimension, which leads to

$$\mathbb{R}^d = AS^d \supset AS^{d-1} \supset \cdots \supset AS^1 \supset AS^0.$$

In particular, AS^{d-1} is found from AS^d by removing the direction $\vec{u}_d$ of least variance from all of the data points. Successively, AS^i is the best fit in AS^{i+1} (not in AS^d). In the toy example in Figure 5.6, the backward operation can be understood by viewing the plots from right to left. From $\mathbb{R}^3$, AS^2 is fitted by removing a direction $\vec{u}_3$, the direction of least variance. Then a line (AS^1) is found within AS^2 in the same fashion, and so on.

In Euclidean space the forward and backward approaches are equivalent. In practice, the

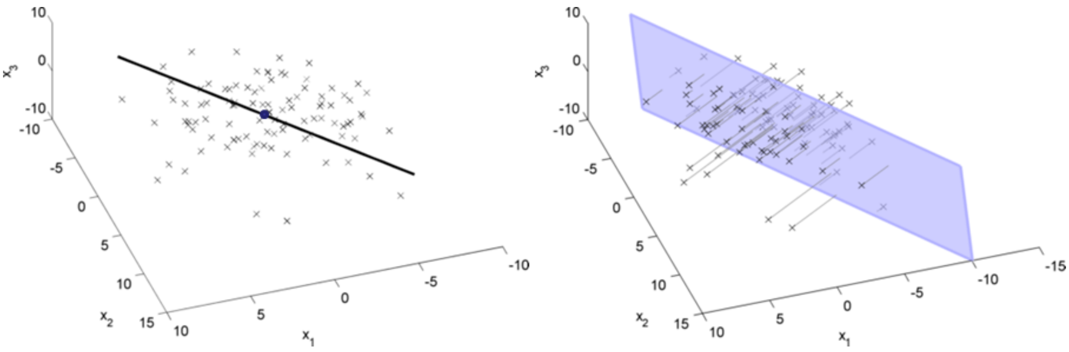


Figure 5.6: Sample points in 3-space with mean and PC1 direction, and the affine subspace formed by PC1–2 directions.

basis of AS^i is formed by the eigenvectors $\vec{u}_j$, $j = 1, \dots, i$, of the sample covariance matrix $S = \frac{1}{n-1}(\mathbf{X} - \bar{X})(\mathbf{X} - \bar{X})^T$ or the left singular vectors of the centered data matrix $(\mathbf{X} - \bar{X})$.

However, in non-Euclidean spaces the choice of viewpoint affects the generalizations of PCA, discussed next.

In curved manifolds we need to generalize important notions such as the sample mean and straight lines (or directions) as they are not defined in general manifolds. A useful notion for generalization of mean is the Fréchet mean or the geodesic mean (5.8). The Fréchet mean is widely applicable, since it only requires a metric on the manifold. In Euclidean space, the sample mean is the Fréchet mean with the usual metric $\rho(x, y) = \|x - y\|$.

A widely used approach to manifold PCA, called Principal Geodesic Analysis (PGA, Fletcher et al. (2004)), generalizes PCA in a forward stepwise manner. The first step in PGA is to find a center point for the manifold data. Having the geodesic mean as the center point in PGA, the second step is to find a geodesic (instead of a line) that best represents the data, among all geodesics that pass through the geodesic mean. The higher order components are again geodesics that are orthogonal to the lower order geodesics. In practice, these geodesic components are computed through a projection of the data onto the tangent space at the geodesic mean. PGA and similarly defined forward approaches are developed for various types of data; see e.g. Fletcher et al. (2004) for m-reps data, Fletcher and Joshi (2007) for DTI data, and Dryden and Mardia (1998) for landmark shape data.

However, there has been a concern that the geodesic mean and tangent space approximation can be very poor. As a simple example, consider the usual unit sphere S^2 and the data distributed uniformly along the equator of the sphere as illustrated in Figure 5.5. In this case, the equator itself is the geodesic that best represents the data. However, the geodesic mean is located near the north or the south pole, far from any data. PGA, as a forward method, finds principal geodesics through this geodesic mean, which fail to effectively describe the variation in the data.

This observation motivated Huckemann et al. (2010) to propose Geodesic PCA (GPCA). In GPCA, the geodesic mean or any pre-determined mean is no longer used; instead it finds the best approximating geodesic among all possible candidates. A center point of the data is then found in the first geodesic component, and all other components must be geodesics through the center point. In the equator example above, GPCA finds the equator as the first component. GPCA can be viewed as a backward approach, particularly when applied to S^2, since the center point is found last. In higher dimensional manifolds, for example in hyperspheres S^p with $p > 2$ and Kendall's shape spaces Dryden and Mardia (1998), GPCA is not fully backward, since the method is built by considering lower dimensional components first, only with an exception for center point. Nevertheless, the advantage of the method indeed comes from the backward viewpoint, i.e., from reversing the order of the first two steps.

Another method that can be viewed as the backward stepwise approach is Principal Arc Analysis (PAA), presented in Chapter 6, which is a non-geodesic generalization of PCA. PAA is motivated by data distributed along a *small circle* on S^2. Since the major variation is no longer along a geodesic, no geodesic based methods including PGA and GPCA capture the variation effectively. PAA begins with the full sphere S^2 and finds the small circle as the best fitting 1-d approximation of the data, followed by a center point contained in the small circle. PAA was shown to provide just this type of effective data approximation in S^2 and also in m-reps data.

In generalizations of PCA for higher dimensional manifolds, including hyperspheres S^p and Kendall's shape spaces, the backward stepwise principle led to a fully backward generalization of PCA: Principal Nested Spheres (PNS, see Chapter 7). In taking the backward approach, it inherits the advantages of GPCA. Moreover, this allows the successive submanifolds to be non-geodesic. PNS has been shown to provide more representative description of the data (compared to other forward stepwise approaches) in a number of standard examples.

Many kernel PCA (Schlkopf et al. (1998)) methods can also be used for dimension reduction and feature extraction, but the nature of these methodologies is quite different from PCA extensions for manifold data. Kernel PCA maps the data implicitly into a Hilbert space, then PCA is applied to the mapped data. While this approach gives a wide possibility of non-linear feature extraction, the interpretation of the result is difficult and reconstruction of components in the original space is impossible.

Chapter 6

Principal Arc Analysis on direct product manifold

The work presented in this chapter is based on and contained in Jung et al. (2011b).

6.1 Introduction

Principal Component Analysis (PCA) has been frequently used as a method of dimension reduction and data visualization for high dimensional data. For data that naturally lie in a curved manifold, application of PCA is not straightforward since the sample space is not linear. Nevertheless, the need for PCA-like methods is growing as more manifold data sets are encountered and as the dimensions of the manifolds increase.

In this chapter, we introduce a new approach for an extension of PCA on a special class of manifold data. We focus on direct products of simple manifolds, in particular, of the unit circle S^1, the unit sphere S^2, $\mathbb{R}_+$ and $\mathbb{R}^p$, as defined in the previous chapter. Many types of statistical sample spaces are special cases of the direct product manifold. A widely known example is the sample space for directional data (Fisher (1993), Fisher et al. (1993) and Mardia and Jupp (2000)), and their direct products. Applications include analysis of wind directions, orientations of cracks, magnetic field directions and directions from the earth to celestial objects. For example, when we consider multiple 3-D directions simultaneously, the sample space is $S^2 \otimes \cdots \otimes S^2$, which is a direct product manifold. Another example is the medial representation of shapes (m-reps, Siddiqi and Pizer (2008)) that provides a powerful parametrization of 3-D shapes of human organs and has been extensively studied in the image analysis field. The space of m-reps is usually a high-dimensional direct product manifold; see Section 6.5.

Our approach to a manifold version of PCA builds upon earlier work, especially the principal geodesic analysis proposed by Fletcher (2004) and the geodesic PCA proposed by Huckemann and Ziezold (2006) and Huckemann et al. (2010). A detailed catalogue of current methodologies can be found in Huckemann et al. (2010). An important approach among these is to approximate the manifold by a linear space. Fletcher et al. (2004) take the tangent space of the manifold at the geodesic mean as the linear space, and work with appropriate mappings between the manifold and the tangent space. This results in finding the best fitting geodesics among those passing through the geodesic mean. This was improved in an important way by Huckemann, who found the best fit over the set of all geodesics. Huckemann went on to propose a new notion of center point, the *PCmean*, which is an intersection of the first two principal geodesics. This approach gives significant advantages especially when the curvature of the manifold makes the geodesic mean inadequate, an example of which is depicted in Figure 6.2b.

Our method inherits advantages of these methods and improves further by effectively capturing more complex non-geodesic modes of variation. Note that the curvature of direct product manifolds is mainly due to the spherical part, which motivates careful investigation of S^2-valued variables. We point out that (small) circles in S^2, including geodesics, can be used to capture the non-geodesic variation. We introduce the *principal circles* and *principal circle mean*, analogous to, yet more flexible than, the geodesic principal component and PCmean of Huckemann *et al.*. These become *principal arcs* when the manifold is indeed S^2. For more complex direct product manifolds, we suggest transforming the data points in S^2 into a linear space by a special mapping utilizing the principal circles. For the other components of the manifold, the tangent space mappings can be used to map the data into a linear space as done in Fletcher et al. (2004). Once manifold-valued data are mapped onto the linear space, then the classical linear PCA can be applied to find principal components in the transformed linear space. The estimated principal components in the linear space can be back-transformed to the manifold, which leads to principal arcs.

We illustrate the potential of our method by an example of m-rep data in Figure 6.1. Here,

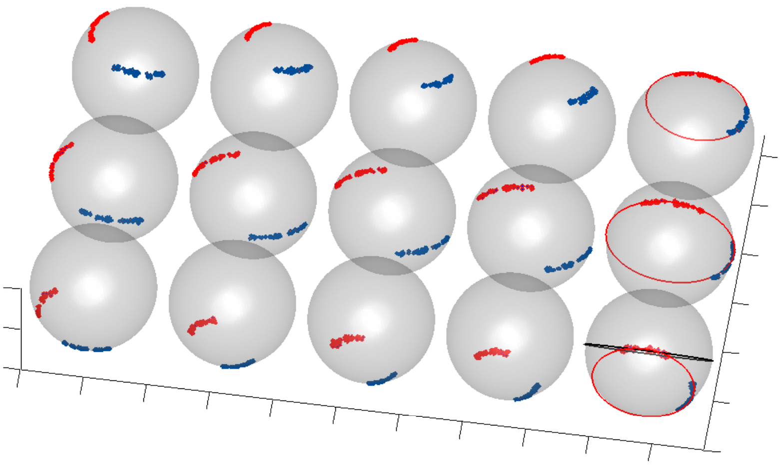


Figure 6.1: S^2-valued samples ($n = 60$) of the prostate m-reps with 15 medial atoms. One sample in this figure is represented by a 30-tuple of 3-D directions (two directions at each atom), which lies in the manifold $\otimes_{i=1}^{15}(S^2 \otimes S^2)$. Small and great circles are fitted and plotted in the rightmost atoms to emphasize the sample variation along small circles.

m-reps with 15 sample points called atoms model the prostate gland (an organ in the male reproductive system) and come from the simulator developed and analyzed in Jeong et al. (2008). Figure 6.1 shows that the S^2 components of the data tend to be distributed along small circles, which frequently are not geodesics. We emphasize the curvature of variation along each sphere by fitting a great circle and a small circle (by the method discussed in Section 6.2). Our method is adapted to capture this nonlinear (non-geodesic) variation of the data. A potential application of our method is to improve accuracy of segmentation of objects from CT images. Detailed description of the data and results of our analysis can be found in Section 6.5.

Note that the previous approaches (Fletcher et al. (2004), Huckemann and Ziezold (2006)) are defined for general manifolds, while our method focuses on these particular direct product manifolds. Although the method is not applicable for general manifolds, it is useful for this common class of manifolds that is often found in applications. Our results inform our belief that focusing on specific types of manifolds allow more precise and informative statistical modeling than methods that attempt to be fully universal. This happens through using special properties (e.g. presence of small circles) that are not available for all other manifolds.

The rest of the chapter is organized as follows. We begin by introducing a *circle class* on S^2 as an alternative to the set of geodesics. Section 6.2 discusses principal circles in S^2, which will be the basis of the special transformation. The first principal circle is defined by the least-squares circle, minimizing the sum of squared residuals. In Section 6.3, we introduce a data-driven method to decide whether the least-squares circle is appropriate. A recipe for principal arc analysis on direct product manifolds is proposed in Section 6.4 with discussion on the transformations. A detailed introduction of the space of m-reps, and the results from applying the proposed method follow. A novel computational algorithm for the least-squares circles is presented in Section 6.6.

6.2 Circle class for non-geodesic variation on S^2

Consider a set of points in $\mathbb{R}^2$. Numerous methods for understanding population properties of dataset in linear space have been proposed and successfully applied, which include rigid methods such as linear regression and principal components and very flexible methods such as scatterplot smoothing and principal curves (Hastie and Stuetzle (1989)). We make use of a parametric class of circles, including small and great circles, which allows much more flexibility than either methods of Fletcher (2004) or Huckemann et al. (2010), but less flexibility than a principal curve approach. Although this idea was motivated by examples such as those in Figure 6.1, there are more advantages gained from using the class of circles;

- (i) The circle class includes the simple geodesic case.
- (ii) Each circle can be parameterized, which leads to an easy interpretation.
- (iii) There is an orthogonal complement of each circle, which gives two important advantages:
  - (a) Two orthogonal circles can be used as a basis of a further extension to principal arc analysis.
  - (b) Building a sensible notion of principal components on S^2 alone is easily done by utilizing the circles.

The idea (iii-b) will be discussed in detail after introducing a method of circle fitting. A circle on S^2 is conveniently parameterized by center $\mathbf{c} \in S^2$ and geodesic radius r, and denoted by $\delta(\mathbf{c}, r) = \{\mathbf{x} \in S^2 | \rho(\mathbf{c}, \mathbf{x}) = r\}$. It is a geodesic when $r = \pi/2$. Otherwise it is a small circle.

A circle that best fits the points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in S^2$ is found by minimizing the sum of squared residuals. The residual of $\mathbf{x}_i$ is defined as the signed geodesic distance from $\mathbf{x}_i$ to the circle $\delta(\mathbf{c}, r)$. Then the least-squares circle is obtained by

$$\begin{aligned} &\min_{\mathbf{c},r} \sum_{i=1}^{n} (\rho(\mathbf{x}_i, \mathbf{c}) - r)^2 \qquad (6.1) \\ &\text{subject to } \ \mathbf{c} \in S^2, \ r \in (0, \pi), \end{aligned}$$

Note that there are always multiple solutions of (6.1). In particular, whenever $(\mathbf{c}, r)$ is a solution, $(-\mathbf{c}, \pi - r)$ also solves the problem as $\delta(\mathbf{c}, r) = \delta(-\mathbf{c}, \pi - r)$. This ambiguity does not affect any essential result. Our convention is to use the circle with smaller geodesic radius.

The optimization task (6.1) is a constrained nonlinear least squares problem. We propose an algorithm to solve the problem that features a simplified optimization task and approximation of S^2 by tangent planes. The algorithm works in a doubly iterative fashion, which has been shown by experience to be stable and fast. Section 6.6 contains a detailed illustration of the algorithm.

Analogous to principal geodesics in S^2, we can define *principal circles* in S^2 by utilizing the least-squares circle. The principal circles are two orthogonal circles in S^2 that best fit the data. We require the first principal circle to minimize the variance of the residuals, so it is the least-squares circle (6.1). The second principal circle is a geodesic which passes through the center of the first circle and thus is orthogonal at the points of intersection. Moreover, the second principal circle is chosen so that one intersection point is the intrinsic mean (defined in (6.2) later) of the projections of the data onto the first principal circle.

Based on a belief that the intrinsic (or extrinsic) mean defined on a curved manifold may not be a useful notion of center point of the data (see e.g. Huckemann et al. (2010). and Figure 6.2b), the principal circles do not use the pre-determined means. To develop a better notion of center point, we locate the best 0-dimensional representation of the data in a data-driven manner. Inspired by the PCmean idea of Huckemann *et al.*, given the first principal circle δ_1, the *principal circle mean* $\mathbf{u} \in \delta_1$ is defined (in intrinsic way) as

$$\mathbf{u} = \operatorname*{argmin}_{\mathbf{u} \in \delta_1} \sum_{i=1}^{n} \rho^2(\mathbf{u}, P_{\delta_1}\mathbf{x}_i), \tag{6.2}$$

where $P_{\delta_1}\mathbf{x}$ is the projection of $\mathbf{x}$ onto δ_1, that is the point on δ_1 of the shortest geodesic distance to $\mathbf{x}$. Then

$$P_{\delta_1(\mathbf{c},r)}\mathbf{x} = \frac{\mathbf{x}\sin(r) + \mathbf{c}\sin(\rho(\mathbf{x},\mathbf{c}) - r)}{\sin(\rho(\mathbf{x},\mathbf{c}))}, \tag{6.3}$$

as in Eq. (3.3) of Mardia and Gadsden (1977). We assume that $\mathbf{c}$ is the north pole $\mathbf{e}_3$, without losing generality since otherwise the sphere can be rotated. Then

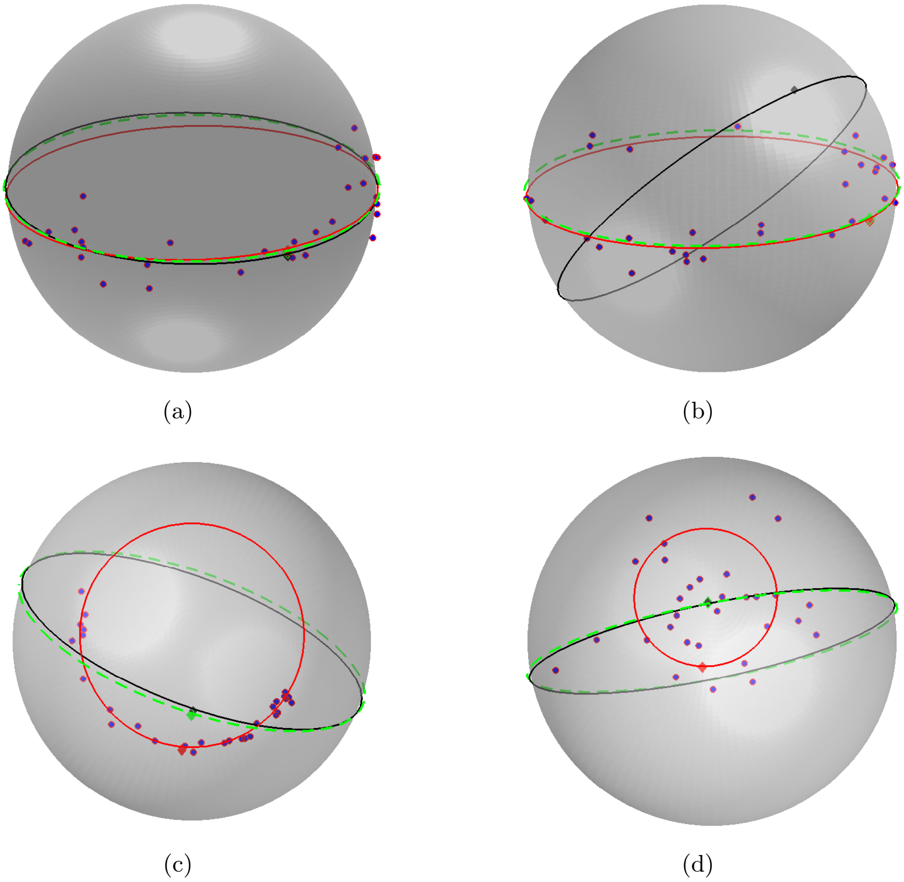
$$\rho(\mathbf{u}, P_{\delta_1}\mathbf{x}) = \sin(r)\rho_{S^1}\left(\frac{(u_1,u_2)}{\sqrt{1-u_3^2}}, \frac{(x_1,x_2)}{\sqrt{1-x_3^2}}\right), \tag{6.4}$$

where $\mathbf{u} = (u_1, u_2, u_3)'$, $\mathbf{x} = (x_1, x_2, x_3)'$ and ρ_{S^1} is the geodesic (angular) distance function on S^1. The optimization problem (6.2) is equivalent to finding the geodesic mean in S^1. See Eq. (5.9) for computation of the geodesic mean in S^1.

The second principal circle δ_2 is then the geodesic passing through the principal circle mean $\mathbf{u}$ and the center $\mathbf{c}$ of δ_1. Denote $\bar{\delta} \equiv \bar{\delta}(\mathbf{x}_1, \dots, \mathbf{x}_n)$ as a combined representation of $(\delta_1, \mathbf{u})$ or equivalently (δ_1, δ_2).

As a special case, we can force the principal circles to be great circles. The best fitting geodesic is obtained as a solution of the problem (6.1) with $r = \pi/2$ and becomes the first principal circle. The optimization algorithm for this case is slightly modified from the original algorithm for the varying r case, by simply setting $r = \pi/2$. The principal circle mean $\mathbf{u}$ and the δ_2 for this case are defined in the same way as in the small circle case. Note that the principal circles with $r = \pi/2$ are essentially the same as the method of Huckemann and Ziezold (2006).

Figure 6.2 illustrates the advantages of using the circle class to efficiently summarize variation. On four different sets of toy data, the first principal circle δ_1 is plotted with principal circle mean $\mathbf{u}$. The first principal geodesics from the methods of Fletcher and Huckemann are also plotted with their corresponding mean. Figure 6.2a illustrates the case where the data were indeed stretched along a geodesic. The solutions from the three methods



(a) (b)

(c) (d)

Figure 6.2: Toy examples on S^2 with $n = 30$ points showing the first principal circle (red) as a small circle and the first geodesic principal component (dotted green) by Huckemann, and the first principal geodesic (black) by Fletcher. Also plotted are the geodesic mean, PCmean and principal circle mean of the data as black, green and red diamonds, respectively. (a) The three methods give similar satisfactory answers when the data are stretched along a geodesic. (b) When the data are stretched along a great circle, covering almost all of it, the principal geodesic (black circle) and geodesic mean (black diamond) fail to find a reasonable representation of the data, while the principal circle and Huckemann's geodesic give sensible answers. (c) Only the principal circle fits well when the data are not along a geodesic. (d) For a small cluster without principal modes of variation, the principal circle gets too small. See Section 6.3 for discussion of this phenomenon.

are similar to one another. The advantage of Huckemann's method over Fletcher's can be found in Figure 6.2b. The geodesic mean is found far from the data, which leads to poor performance of the principal geodesic analysis, because it considers only great circles passing through the geodesic mean. Meanwhile, the principal circle and Huckemann's method, which do not utilize the geodesic mean, work well. The case where geodesic mean and any geodesic do not fit the data well is illustrated in Figure 6.2c, which is analogous to the Euclidean case, where a non-linear fitting may do a better job of capturing the variation than PCA. To this data set, the principal circle fits best, and our definition of mean is more sensible than the geodesic mean and the PCmean. The points in Figure 6.2d are generated from the von Mises–Fisher distribution with $\kappa = 10$, thus having no principal mode of variation. In this case the first principal circle δ_1 follows a contour of the apparent density of the points. We shall discuss this phenomenon in detail in the following section.

Fitting a (small) circle to data on a sphere has been investigated for some time, especially in statistical applications in geology. Those approaches can be distinguished in three different ways, where our choice fits into the first category.

1. Least-squares of intrinsic residuals: Gray et al. (1980) formulated the same problem as in (6.1), finding a circle that minimizes sum of squared residuals, where residuals are defined in a geodesic sense.

2. Least-squares of extrinsic residuals: A different measure of residual was chosen by Mardia and Gadsden (1977) and Rivest (1999), where the residual of $\mathbf{x}$ from $\delta(c, r)$ is defined by the shortest Euclidean distance between $\mathbf{x}$ and $\delta(\mathbf{c}, r)$. Their objective is to find

$$\operatorname*{argmin}_{\delta} \sum_{i=1}^{n} \|\mathbf{x}_i - P_\delta \mathbf{x}_i\|^2 = \operatorname*{argmin}_{\delta} \sum_{i=1}^{n} -\mathbf{x}_i' P_\delta \mathbf{x}_i = \operatorname*{argmin}_{\delta} \sum_{i=1}^{n} -\cos(\xi_i),$$

   where ξ_i denotes the intrinsic residual. This type of approach can be numerically close to the intrinsic method as $\cos(\xi_i) = 1 - \xi_i^2/2 + O(\xi_i^4)$.

3. Distributional approach: Mardia and Gadsden (1977) and Bingham and Mardia (1978) proposed appropriate distributions to model S^2-valued data that cluster near a small circle. These models essentially depend on the quantity $\cos(\xi)$, which is easily interpreted

in the extrinsic sense, but not in the intrinsic sense.

Remark 6.1. The principal circle and principal circle mean always exist. This is because the objective function (6.1) is a continuous function of $\mathbf{c}$, with the compact domain S^2. The minimizer r has a closed-form solution (see Section 6.6). A similar argument can be made for the existence of $\mathbf{u}$. On the other hand, the uniqueness of the solution is not guaranteed. We conjecture that if the manifold is approximately linear or equivalently the data set is well-approximated by a linear space, then the principal circle will be unique. However, this does not lead to the uniqueness of $\mathbf{u}$, whose sufficient condition is that the projected data on δ_1 is strictly contained in a half-circle (Karcher (1977)). Note that a sufficient condition for the uniqueness of the principal circle is not clear even in Euclidean case (Chernov (2010)).

6.3 Suppressing small least-squares circles

When the first principal circle δ_1 has a small radius, sometimes it is observed that δ_1 does not fit the data in a manner that gives useful decomposition, as shown in Figure 6.2d. This phenomenon has been also observed for the related Principal curve fitting method of Hastie and Stuetzle (1989). We view this as unwanted overfitting, which is indeed a side effect caused by using the full class of circles with free radius parameter instead a class of great circles. In this section, a data-driven method to flag this overfitting is discussed. In essence, the fitted small circle is replaced by the best fitting geodesics when the data do not cluster along the circle but instead tend to cluster near the center of the circle.

We first formulate the problem and solution in $\mathbb{R}^2$. This is for the sake of clear presentation and also because the result on $\mathbb{R}^2$ can be easily extended to S^2 using a tangent plane approximation.

Let $f_{\mathbf{X}}$ be a spherically symmetric density function of a continuous distribution defined on $\mathbb{R}^2$. Whether the density is high along some circle is of interest. By the symmetry assumption, density height along a circle can be found by inspecting a section of $f_{\mathbf{X}}$ along a ray from the origin (the point of symmetry). A section of $f_{\mathbf{X}}$ coincides with the conditional density $f_{X_1|X_2}(x_1|x_2 = 0) = \kappa^{-1} f_{\mathbf{X}}(x_1, 0)$. A random variable corresponding to the pdf $f_{X_1|X_2=0}$ is not directly observable. Instead, the radial distance $R = \|\mathbf{X}\|$ from the origin can be observed.

For the polar coordinates (R, Θ) such that $\mathbf{X} = (X_1, X_2) = (R\cos\Theta, R\sin\Theta)$, the marginal pdf of R is $f_R(r) = 2\pi r f_{\mathbf{X}}(r, 0)$ as $f_{\mathbf{X}}$ is spherically symmetric. A section of $f_{\mathbf{X}}$ is related to the observable density f_R as $f_R(r) \propto r f_{X_1|X_2=0}(r)$, for $r \geqslant 0$. This relation is called the length-biased sampling problem (Cox (1969)). The relation can be understood intuitively by observing that a value r of R can be observed at any point on a circle of radius r, circumference of which is proportional to r. Thus sampling of R from the density $f_{X_1|X_2=0}$ is proportional to its size.

The problem of suppressing a small circle can be paraphrased as "how to determine whether a nonzero point is a mode of the function $f_{X_1|X_2=0}$, when observing only a length-biased sample."

The spectrum from the circle-clustered case (mode at a nonzero point) to the center-clustered case (mode at origin), can be modeled as

$$\text{data} = \text{signal} + \text{error}, \tag{6.5}$$

where the signal is along a circle with radius μ, and the error accounts for the perpendicular deviation from the circle. (see Figure 6.3) Then, in polar coordinates (R, Θ), Θ is uniformly distributed on $(0, 2\pi]$ and R is a positive random variable with mean μ. First assume that R follows a truncated Normal distribution with standard deviation σ, with the marginal pdf proportional to

$$f_R(r) \propto \phi\left(\frac{r-\mu}{\sigma}\right), \text{ for } r \geqslant 0, \tag{6.6}$$

where ϕ is the standard Normal density function. The conditional density $f_{X_1|X_2=0}$ is then

$$f_{X_1|X_2=0}(r) \propto \frac{1}{r} f_R(r) \propto \frac{1}{r\sigma}\exp\left(-\frac{(r-\mu)^2}{\sigma^2}\right), \text{ for } r > 0.$$

Non-zero local extrema of $f_{X_1|X_2=0}$ can be characterized as a function of (μ, σ) in terms of $r_+, r_- = \{\mu \pm \sqrt{(\mu - 2\sigma)(\mu + 2\sigma)}\}/2$ as follows:

- When $\mu > 2\sigma$, $f_{X_1|X_2=0}$ has a local maximum at r_+, minimum at r_-.

- When $\mu = 2\sigma$, $r_+ = r_- = \frac{\mu}{2}$.

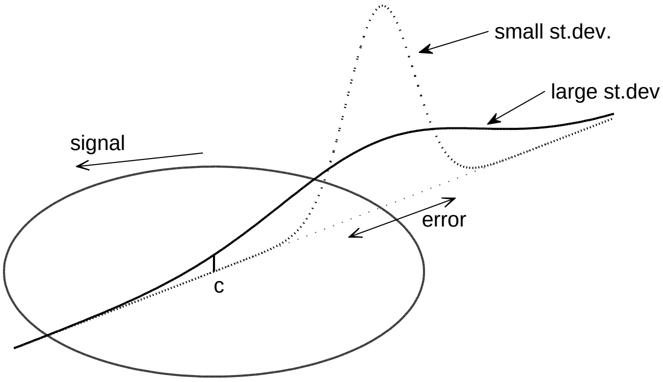


Figure 6.3: Illustration of the conceptual model (6.5) on $\mathbb{R}^2$, which can also be understood as a local approximation of S^2. The signal is along the circle centered at c and radius μ. The error is perpendicular to the signal. When the deviation σ is large, it is possible that the amount of error is even greater than the radius μ. This is incorporated in the wrapping approach (6.7).

- When $\mu < 2\sigma$, $f_{X_1|X_2=0}$ is strictly decreasing, for $r \geqslant 0$.

Therefore, whenever the ratio $\mu/\sigma > 2$, $f_{X_1|X_2=0}$ has a mode at r_+.

This idea can be applied for circles in S^2 with some modification, shown next. We point out that the model (6.5) is useful for understanding the small circle fitting: signal as a circle with radius μ, and error as the deviation along geodesics perpendicular to the circle. Moreover, a spherically symmetric distribution centered at $\mathbf{c}$ on S^2 can be mapped to a spherically symmetric distribution on the tangent space at $\mathbf{c}$, preserving the radial distances by the log map (defined in the Appendix). A modification need to be made on the truncated density f_R. It is more natural to let the error be so large that the deviation from the great circle is greater than μ. Then the observed value may be found near the opposite side of true signal, which is illustrated in Figure 6.3 as the large deviation case. To incorporate this case, we consider a wrapping approach. The distribution of errors (on the real line) is *wrapped* around the sphere along a great circle through $\mathbf{c}$, and the marginal pdf f_R in (6.6) is modified to

$$f_R^w(r) \propto \sum_{k=0}^{\infty} \left[\phi\left(\frac{r + 2\pi k - \mu}{\sigma} \right) + \phi\left(\frac{r - 2\pi k + \mu}{\sigma} \right) \right], \text{ for } r \in [0, \pi]. \tag{6.7}$$

The corresponding conditional pdf, $f^w_{X_1|X_2=0}$, is similar to $f_{X_1|X_2=0}$ and a numerical calculation shows that $f^w_{X_1|X_2=0}$ has a mode at some nonzero point whenever $\mu/\sigma > 2.0534$, for $\mu < \pi/2$.

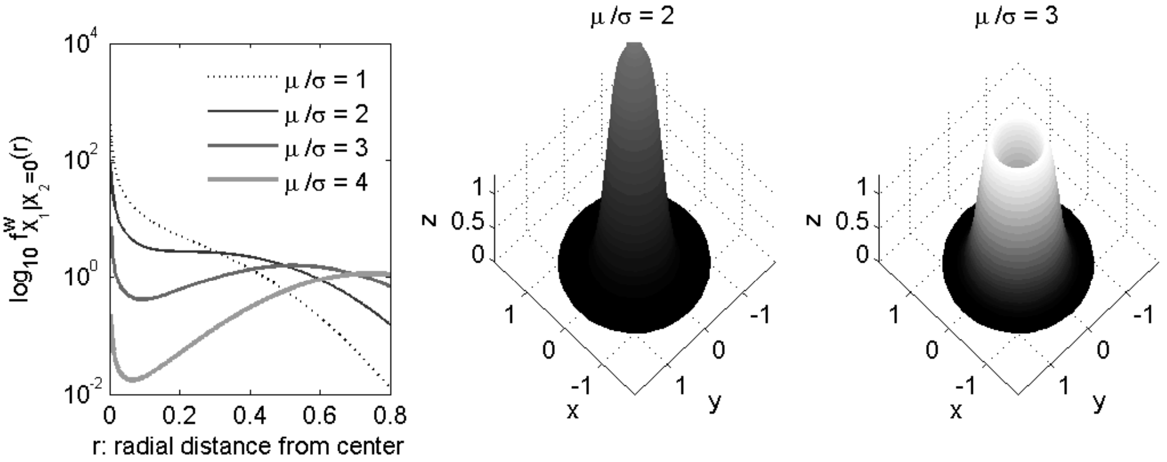


Figure 6.4: (left) Graph of $f^w_{X_1|X_2=0}(r)$ for $\mu/\sigma = 1, 2, 3, 4$. The density is high at a non-zero point when $\mu/\sigma > 2.0534$. (center, right) Spherically symmetric distributions f corresponding to $\mu/\sigma = 2, 3$. The ratio $\mu/\sigma > 2$ roughly leads to a high density along a circle.

In other words, we use the small circle when μ/σ is large. Note that in what follows, we only consider the first term ($k = 0$) of (6.7) since other terms are negligible in most situations. We have plotted $f^w_{X_1|X_2=0}$ for some selected values of μ and σ in Figure 6.4.

With a data set on S^2, we need to estimate μ and σ, or the ratio μ/σ. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in S^2$ and let $\hat{\mathbf{c}}$ be the samples and the center of the fitted circle, respectively. Denote ξ_i for the errors of the model (6.5) such that $\xi_i \sim N(0, \sigma^2)$. Then $r_i \equiv \rho(\mathbf{x}_i, \hat{\mathbf{c}}) = |\mu + \xi_i|$, which has the *folded normal* distribution (Leone et al. (1961)). Estimation of μ and σ based on unsigned r_i is not straightforward. We present two different approaches to this problem.

Robust approach The observations $r_1, \ldots, r_n$ can be thought of as a set of positive numbers contaminated by the folded negative numbers. Therefore the left half (near zero) of the data are more contaminated than the right half. We only use the right half of the data, which are less contaminated than the other half. We propose to estimate μ and σ by

$$\hat{\mu} = \mathrm{med}(r_1^n), \quad \hat{\sigma} = (\mathrm{Q}_3(r_1^n) - \mathrm{med}(r_1^n))/\mathrm{Q}_3(\Phi), \tag{6.8}$$

where $\mathrm{Q}_3(\Phi)$ is the third quantile of the standard normal distribution. The ratio can be estimated by $\hat{\mu}/\hat{\sigma}$.

Likelihood approach via EM algorithm The problem may also be solved by a likelihood approach. Early solutions can be found in Leone *et al.*, Elandt (1961) and Johnson

(1962), in which the MLEs were given by numerically solving nonlinear equations based on the sample moments. As those methods were very complicated, we present a simpler approach based on the EM algorithm. Consider unobserved binary variables s_i with values -1 and $+1$ so that $s_i r_i \sim N(\mu, \sigma^2)$. The idea of the EM algorithm is that if we have observed s_i, then the maximum likelihood estimator of $\vartheta = (\mu, \sigma^2)$ would be easily obtained. The EM algorithm is an iterative algorithm consisting of two steps. Suppose that the kth iteration produced an estimate $\hat{\vartheta}_k$ of ϑ. The E-step is to impute s_i based on r_i and $\hat{\vartheta}_k$ by forming a conditional expectation of log-likelihood for ϑ,

$$\begin{aligned}
Q(\vartheta) &= E\left[\log \prod_{i=1}^{n} f(r_i, s_i|\vartheta)\Big| r_i, \hat{\vartheta}_k\right] \\
&= \sum_{i=1}^{n} \left[\log f(r_i|s_i = +1, \vartheta) P(s_i = +1|r_i, \hat{\vartheta}_k) + \log f(r_i|s_i = -1, \vartheta) P(s_i = -1|r_i, \hat{\vartheta}_k)\right] \\
&= \sum_{i=1}^{n} \left[\log \phi(r_i|\vartheta) p_{i(k)} + \log \phi(-r_i|\vartheta)(1 - p_{i(k)})\right],
\end{aligned}$$

where f is understood as an appropriate density function, and $p_{i(k)}$ is easily computed as

$$p_{i(k)} = P(s_i = +1|r_i, \hat{\vartheta}_k) = \frac{\phi(r_i|\hat{\vartheta}_k)}{\phi(r_i|\hat{\vartheta}_k) + \phi(-r_i|\hat{\vartheta}_k)}.$$

The M-step is to maximize $Q(\vartheta)$ whose solution becomes the next estimator $\hat{\vartheta}_{k+1}$. Now the $k+1$th estimates are calculated by a simple differentiation and given by;

$$Q(\vartheta) = Q(\mu, \sigma^2) = \sum_{i=1}^{n} \left[\frac{1}{2}\log 2\pi\sigma^2 + \frac{(r_i - \mu)^2 p_{i(k)}}{2\sigma^2} \frac{(r_i + \mu)^2 (1 - p_{i(k)})}{2\sigma^2}\right]$$

$$\frac{\partial Q(\vartheta)}{\partial \mu} = 0 \Longleftrightarrow \hat{\mu}_{k+1} = \frac{1}{n}\sum_{i=1}^{n} (2p_{i(k)} - 1) r_i$$

$$\begin{aligned}
\frac{\partial Q(\vartheta)}{\partial \sigma^2} = 0 \Longleftrightarrow \hat{\sigma}^2_{k+1} &= \frac{1}{n}\sum_{i=1}^{n} (r_i - \hat{\mu}_{k+1})^2 p_{i(k)} + (r_i + \hat{\mu}_{k+1})^2 (1 - p_{i(k)}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \left(r_i^2 - \hat{\mu}^2_{k+1}\right).
\end{aligned}$$

With the sample mean and variance of $r_1, \ldots, r_n$ as an initial estimator $\hat{\vartheta}_0$, the algorithm iterates E-steps and M-steps until the iteration changes the estimates less than a pre-defined criteria (e.g. 10^{-10}). μ/σ is estimated by the ratio of the solutions.

Comparison Performance of these estimators are now examined by a simulation study. Normal random samples are generated with ratios μ/σ being 0, 1, 2, or 3, representing the transition from the center-clustered to circle-clustered case. For each ratio, $n = 50$ samples are generated, from which $\hat{\mu}/\hat{\sigma}$ is estimated. These steps are repeated 1000 times to obtain the sampling variation of the estimates. We also study $n = 1000$ case in order to investigate the consistency of the estimators. The results are summarized in Figure 6.5 and Table. 6.1.

The distribution of estimators are shown for $n = 50, 1000$ in Figure 6.5 and the proportion of estimators greater than 2 is summarized in Table. 6.1. When $n = 1000$, both estimators are good in terms of the proportion of correct answers. In the following, the proportions of correct answers are corresponding to $n = 50$ case. The top left panel in Figure 6.5 illustrates the circle-centered case with ratio 3. The estimated ratios from the robust approach give correct solutions (greater than 2) 95% of the time (98.5% for likelihood approach). For the borderline case (ratio 2, top right), the small circle will be used about half the time. The center-clustered case is demonstrated with the true ratio 1, that also gives a reasonable answer (proportion of correct answers 95.3% and 94.8% for the robust and likelihood answers respectively). It can be observed that when the true ratio is zero, the robust estimates are far from 0 (the bottom right in Figure 6.5). However, this is expected to occur because the proportion of uncontaminated data is low when the ratio is too small. However, those 'inaccurate' estimates are around 1 and less than 2 most of the time, which leads to 'correct' answers. The likelihood approach looks somewhat better with more hits near zero, but an asymptotic study (Johnson (1962)) showed that the variance of the maximum likelihood estimator converges to infinity when the ratio tends to zero, as glimpsed in the long right tail of the simulated distribution.

In summary, we recommend use of the robust estimators (6.8), which are computationally light, straightforward and stable for all cases.

In addition we point out that Gray et al. (1980) and Rivest (1999) proposed to use a goodness of fit statistic to test whether the small circle fit is better than a geodesic fit. Let r_g and r_c be the sums of squares of the residuals from great and small circle fits. They claimed

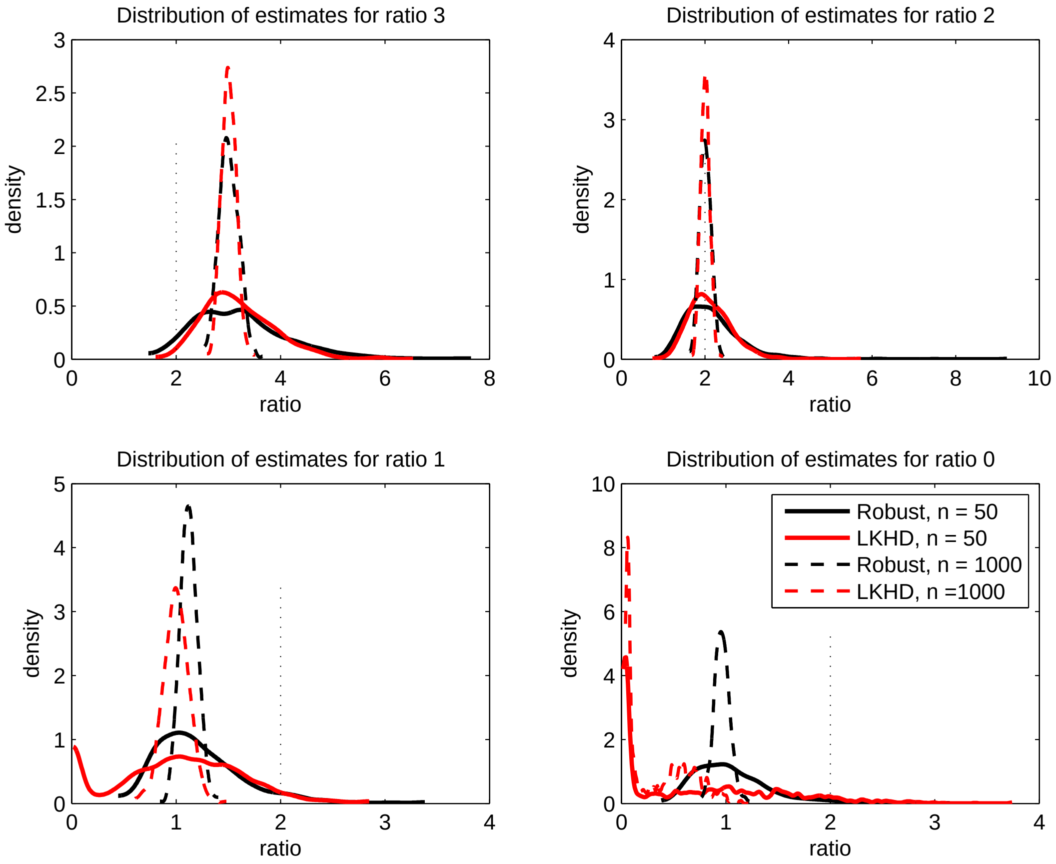


Figure 6.5: Simulation results of the proposed estimators for the ratio μ/σ. Different ratios represent different underlying distributions. For example, estimators in the top left are based on random samples from a folded Normal distribution with mean $\mu = 3$, standard deviation $\sigma = 1$. Curves are smooth histograms of estimates from 1000 repetitions. The thick black curve represents the distribution of the robust estimator from $n = 50$ samples. Likewise, the thick red curve is for the MLE with $n = 50$, the dotted black curve is for the robust estimator with $n = 1000$, and the dotted red curve is for the MLE with $n = 1000$. The smaller sample size represents a usual data analytic situation, while $n = 1000$ case shows an asymptotic situation.

Method	$\mu/\sigma = 3$	$\mu/\sigma = 2$	$\mu/\sigma = 1$	$\mu/\sigma = 0$
MLE, $n = 50$	98.5	55.2	5.2	6.8
Robust, $n = 50$	95.0	50.5	4.7	1.4
MLE, $n = 1000$	100	51.9	0	0
Robust, $n = 1000$	100	50.5	0	0

Table 6.1: Proportion of estimates greater than 2 from the data illustrated in Figure 6.5. For $\mu/\sigma = 3$, shown are proportions of correct answers from each estimator. For $\mu/\sigma = 1$ or 0, shown are proportions of *incorrect* answers.

that $V = (n-3)(r_g - r_c)/r_c$ is approximately distributed as $F_{1,n-3}$ for a large n if the great circle was true. However this test does not detect the case depicted in Figure 6.2d. The following numerical example shows the distinction between our approach and the goodness of fit approach.

Example 6.1. Consider the sets of data depicted in Figure 6.2. The goodness of fit test gives p-values of 0.51, 0.11359, 0, and 0.0008 for (a)-(d), respectively. The estimated ratios μ/σ are $14.92, 16.89$, 14.52 and 1.55. Note that for (d), when the least-squares circle is too small, our method suggests to use a geodesic fit over a small circle while the goodness of fit test gives significance of the small circle. The goodness of fit method is not adequate to suppress the overfitting small circle in a way we desire.

Remark 6.2. Note that the transition of the principal circle between great circle and small circle is not continuous. Specifically, when the data set is perturbed so that the principal circle becomes too small, then the principal circle and principal circle mean are abruptly replaced by a great circle and geodesic mean. As an example, we have generated a toy data set spread along a circle with some radial perturbation. The perturbation is continuously inflated, so that with large inflation, the data are no longer circle-clustered. In Figure 6.6, the $\widehat{\mu/\sigma}$ changes smoothly, but once the estimate hits 2 (our criterion), there is a sharp transition between small and great circles. Sharp transitions do naturally occur in the statistics of manifold data. For example, even the simple geodesic mean can exhibit a major discontinuous transition resulting from an arbitrarily small perturbation of the data. However, the discontinuity between small and great circles does seem more arbitrary and thus may be worth addressing. An interesting open problem is to develop a blended version of our two solutions, for values of $\widehat{\mu/\sigma}$ near 2, which could be done by fitting circles with radii that are smoothly blended between the small circle radius and $\pi/2$.

6.4 Principal Arc Analysis on direct product manifolds

The discussions of the principal circles in S^2 play an important role in defining the principal arcs for data in a direct product manifold $M = M_1 \otimes M_2 \otimes \cdots \otimes M_d$, where each M_i is one of the simple manifolds S^1, S^2, $\mathbb{R}_+$, and $\mathbb{R}$. We emphasize again that the curvature of the direct

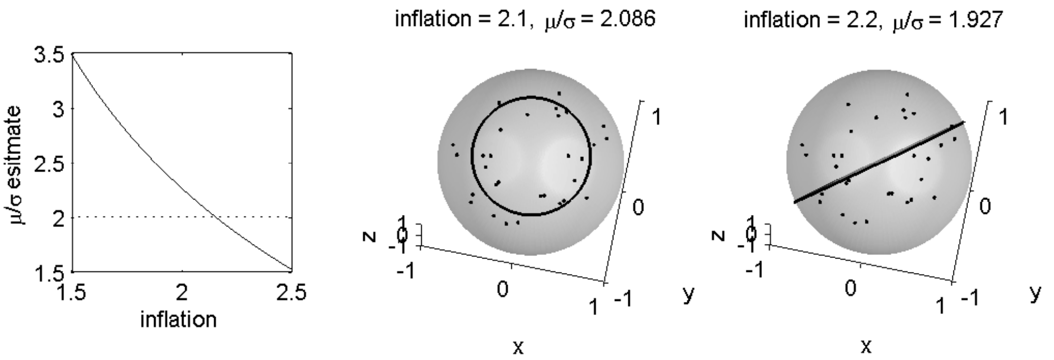


Figure 6.6: (left) The estimate $\widehat{\mu/\sigma}$ decreases smoothly as the perturbation is inflated. (center, right) Snapshots of the toy data on a sphere. A very small perturbation of the data set leads to a sharp transition between small circle (center) and great circle (right).

product manifold M is mainly due to the spherical components.

Consider a dataset $x_1, \ldots, x_n \in M$, where $x_i \equiv (x_i^1, \ldots, x_i^d)$ such that $x_i^j \in M_j$. Denote $d_0 \geqslant d$ for the intrinsic dimension of M. The geodesic mean $\bar{x}$ of the data is defined component-wise for each simple manifold M_j. Similarly, the tangent plane at $\bar{x}$, $T_{\bar{x}}M$, is also defined marginally, i.e. $T_{\bar{x}}M$ is a direct product of tangent spaces of the simple manifolds. This tangent space gives a way of applying Euclidean space-based statistical methods, by mapping the data onto $T_{\bar{x}}M$. We can manipulate this approximation of the data component-wise. In particular, the marginal data on the S^2 components can be represented in a linear space by a transformation $h_{\bar{\delta}}$, depending on the principal circles, that differs from the tangent space approximation.

Since the principal circles $\bar{\delta}$ capture the non-geodesic directions of variation, we use the principal circles as axes, which can be thought of as flattening the quadratic form of variation. In principle, we require a mapping $h_{\bar{\delta}} : S^2 \to \mathbb{R}^2$ to have the following properties: For $\bar{\delta} = (\delta_1, \delta_2) = ((\delta_1(\mathbf{c}, r), \mathbf{u})$,

- $\mathbf{u}$ is mapped to the origin,
- δ_1 is mapped onto the x-axis, and
- δ_2 is mapped onto the y-axis.

Two reasonable choices of the mapping $h_{\bar{\delta}}$ will be discussed in Section 6.4.1, in detail.

The mapping $h_{\bar{\delta}}$ and the tangent space projection together give a linear space representation of the data, where the Euclidean PCA is applicable. The line segments corresponding to the sample principal component direction of the transformed data can be mapped back to M, and become the principal arcs.

A procedure for principal arc analysis is as follows:

(1) For each j such that M_j is S^2, compute principal circles $\bar{\delta} = \bar{\delta}(x^j_1, \ldots, x^j_2)$ and the ratio $\widehat{\mu/\sigma}$. If the ratio is greater than the pre-determined value $\epsilon = 2$, then $\bar{\delta}$ is adjusted to be great circles as explained in Section 6.2.

(2) Let $\mathbf{h} : M \to \mathbb{R}^{d_0}$ be a transformation $\mathbf{h}(x) = (h_1(x^1), \ldots, h_d(x^d))$. Each component of $\mathbf{h}$ is defined as

$$h_j(x^j) = \begin{cases} h_{\bar{\delta}}(x^j) & \text{for } M_j = S^2, \\ \mathrm{Log}_{\bar{x}^j}(x^j) & \text{otherwise,} \end{cases}$$

where $\mathrm{Log}_{\bar{x}^j}$ and $h_{\bar{\delta}}$ are defined in the Appendix and Section 6.4.1, respectively.

(3) Observe that $\mathbf{h}(x_1), \ldots, \mathbf{h}(x_n) \in \mathbb{R}^{d_0}$ always have their mean at the origin. Thus, the singular value decomposition of the $d_0 \times n$ data matrix $\mathbf{X} \equiv [\mathbf{h}(x_1) \cdots \mathbf{h}(x_n)]$ can be used for computation of the PCA. Let $\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_m$ be the left singular vectors of $\mathbf{X}$ corresponding to the largest m singular values.

(4) The kth principal arc is obtained by mapping the direction vectors $\mathbf{v}_k$ onto M by the inverse of h, which can be computed component-wise.

The principal arcs on M are not, in general, geodesics. Nor are they necessarily circles, in the original marginal S^2. This is because $h_{\bar{\delta}}$ and its inverse $h_{\bar{\delta}}^{-1}$ are nonlinear transformations and thus a line on $\mathbb{R}^2$ may not be mapped to a circle in S^2. This is consistent with the fact that the principal components on a subset of variables are different from projections of the principal components from the whole variables.

Principal arc analysis for data on direct product manifolds often results in a concise summary of the data. When we observe a significant variation along a small circle of a marginal S^2, that is most likely not a random artifact but, instead, the result of a signal driving the circular variation. Non-geodesic variation of this type is well captured by our method.

Principal arcs can be used to reduce the intrinsic dimensionality of M. Suppose we want to reduce the dimension by k, where k can be chosen by inspection of the scree plot. Then each data point x is projected to a k-dimensional submanifold M_0 of M in such a way that

$$\mathbf{h}^{-1}\left(\sum_{i=1}^{k} \mathbf{v}_i \mathbf{v}_i' \mathbf{h}(x)\right) \in M_0,$$

where the $\mathbf{v}_i$'s are the principal direction vectors in $\mathbb{R}^{d_0}$, found by Step 3 above. Moreover, the manifold M_0 can be parameterized by the k principal components $z_1, \ldots, z_k$ such that $M_0(z_1, \ldots, z_k) = \mathbf{h}^{-1}(\sum_{i=1}^{k} z_i \mathbf{v}_i)$

6.4.1 Choice of the transformation $h_{\bar{\delta}}$

The transformation $h_{\bar{\delta}} : S^2 \to \mathbb{R}^2$ leads to an alternative representation of the data, which differs from the tangent space projection. The $h_{\bar{\delta}}$ transforms non-geodesic scatters along δ_1 to scatters along the x-axis, which makes a linear method like the PCA applicable. Among many choices of transformations that satisfy the three principles we stated, two methods are discussed here. Recall that $\bar{\delta} = (\delta_1, \delta_2) = (\delta_1(\mathbf{c}, r), \mathbf{u})$.

Projection The first approach is based on the projection of $\mathbf{x}$ onto δ_1, defined in (6.3), and a residual ξ. The signed distance from $\mathbf{u}$ to $P_{\delta_1}x$, whose unsigned version is defined in (6.4), becomes the x-coordinate, while the residual ξ becomes the y-coordinate. This approach has the same spirit as the model for the circle class (6.5), since the direction of the signal is mapped to the x-axis, with the perpendicular axis for errors.

The projection $h_{\bar{\delta}}(\mathbf{x})$ that we define here is closely related to the spherical coordinate system. Assume $\mathbf{c} = \mathbf{e}_3$, and $\mathbf{u}$ is at the Prime meridian (i.e. on the $x - z$ plane). For $\mathbf{x}$ and its spherical coordinates (ϕ, θ) such that $\mathbf{x} = (x_1, x_2, x_3) = (\cos\phi \sin\theta, \cos\phi \sin\theta, \cos\theta)$,

$$h_{\bar{\delta}}(\mathbf{x}) = (\sin(r)\phi, \theta - \theta_{\mathbf{u}}), \tag{6.9}$$

where $\theta_{\mathbf{u}} = \cos^{-1}(u_3)$ is the latitude of $\mathbf{u}$. The set of $h_{\bar{\delta}}(\mathbf{x}_i)$ has mean zero because the principal circle mean $\mathbf{u}$ has been subtracted.

Conformal map A conformal map is a function which preserves angles. We point out two conformal maps that can be combined to serve our purpose. See Chapter 9 of Churchill and Brown (1984) and Krantz (1999) for detailed discussions of conformal maps. A conformal map is usually defined in terms of complex numbers. Denote the extended complex plane $\mathbb{C} \cup \{\infty\}$ as $\mathbb{C}^*$. Let $\phi_{\mathbf{c}} : S^2 \to \mathbb{C}^*$ be the stereographic projection of the unit sphere when the point antipodal from $\mathbf{c}$ is the projection point. Then $\phi_{\mathbf{c}}$ is a bijective conformal mapping defined for all S^2 that maps δ_1 as a circle centered at the origin in $\mathbb{C}^*$. The linear fractional transformation, sometimes called Möbius transformation, is a rational function of complex numbers, that can be used to map a circle to a line in $\mathbb{C}^*$. In particular, we define a linear fractional transformation $f_{u^*} : \mathbb{C}^* \to \mathbb{C}^*$ as

$$f_{u^*}(z) = \begin{cases} \frac{\alpha i (z - u^*)}{-z - u^*} & \text{if } z \neq -u^* \\ \infty & \text{if } z = -u^*, \end{cases} \tag{6.10}$$

where $u^* = \phi_{\mathbf{c}}(\mathbf{u})$, and α is a constant scalar. Then the image of δ_1 under $f_{u^*} \circ \phi_c$ is the real axis, while the image of δ_2 is the imaginary axis. The mapping $h_{\bar{\delta}} : S^2 \to \mathbb{R}^2$ is defined by $f_{u^*} \circ \phi_c$ with the resulting complex numbers understood as members of $\mathbb{R}^2$. Note that orthogonality of any two curves in S^2 is preserved by the $h_{\bar{\delta}}$ but the distances are not. Thus we use the scale parameter α of the function f_{u^*} to match the resulting total variance of $h_{\bar{\delta}}(\mathbf{x}_i)$ to the geodesic variance of $\mathbf{x}_i$.

In many cases, both projection and conformal $h_{\bar{\delta}}$ give better representations than just using the tangent space. Figure 6.7 illustrates the image of $h_{\bar{\delta}}$ with the toy data set depicted in Figure 6.2c. The tangent space mapping is also plotted for comparison. The tangent space mapping leaves the curvy form of variation, while both $h_{\bar{\delta}}$s capture the variation and leads to an elliptical distribution of the transformed data.

The choice between the projection and conformal mappings is a matter of philosophy. The image of the projection $h_{\bar{\delta}}$ is not all of $\mathbb{R}^2$, while the image of the conformal $h_{\bar{\delta}}$ is all of $\mathbb{R}^2$. However, in order to cover $\mathbb{R}^2$ completely, the conformal $h_{\bar{\delta}}$ can grossly distorts the covariance structure of the data. In particular, the data points that are far from $\mathbf{u}$ are sometimes overly diffused when the conformal $h_{\bar{\delta}}$ is used, as can be seen in the left tail of the conformal mapped image in Figure 6.7. The projection $h_{\bar{\delta}}$ does not suffer from this problem. Moreover, the interpretation of projection $h_{\bar{\delta}}$ is closely related to the circle class model. Therefore we recommend the projection $h_{\bar{\delta}}$, which is used in the following data analysis.

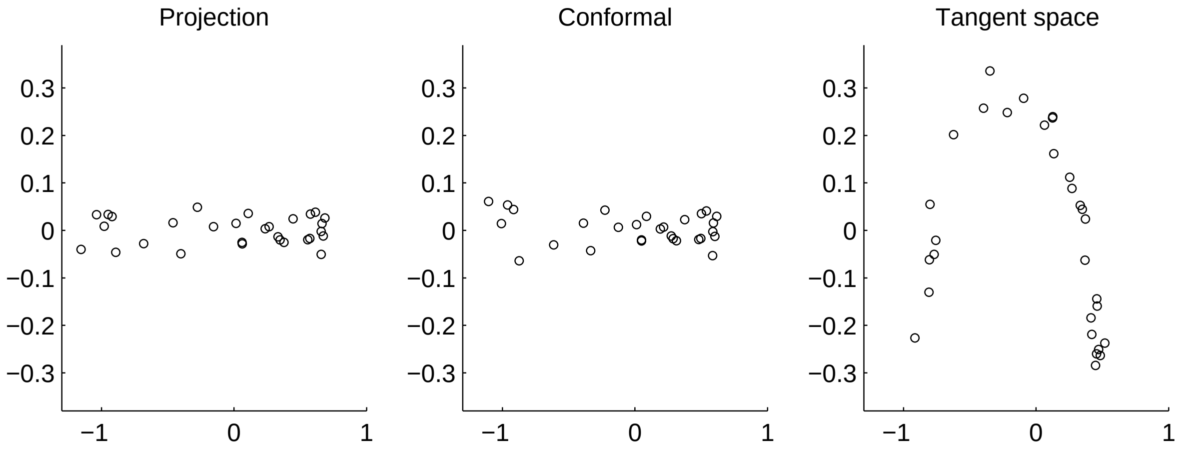


Figure 6.7: Illustration of projection $h_{\bar{\delta}}$ (left, Eq. (6.9)) and conformal $h_{\bar{\delta}}$ (center, Eq. (6.10)) compared to a tangent plane projection at the geodesic mean (right) of the data in Figure 6.2c. The $h_{\bar{\delta}}$ maps the variation along δ_1 to the variation along the x-axis, while the tangent plane mapping fails to do.

6.5 Application to m-rep data

In this section, an application of Principal Arc Analysis to the medial representation (m-rep) data is described.

6.5.1 The Medial Representation of prostate

The m-rep gives an efficient way of representing 2 or 3-dimensional objects. The m-rep consists of *medial atoms*. A medial atom consists of the location of the atom combined with two equal-length spokes, defined as a 4-tuple:

- location in $\mathbb{R}^3$
- spoke direction 1, in S^2
- spoke direction 2, in S^2
- common spoke length in $\mathbb{R}_+$

as shown in Figure 6.8. The size of the regular lattice is fixed for each object in practice. For example, the shape of a prostate is usually described by a 3 × 5 grid of medial atoms, across all samples. The collection of the medial atoms is called the medial representation (m-rep). An m-rep corresponds to a particular shape of prostate, and is a point in the m-rep space $\mathcal{M}$.

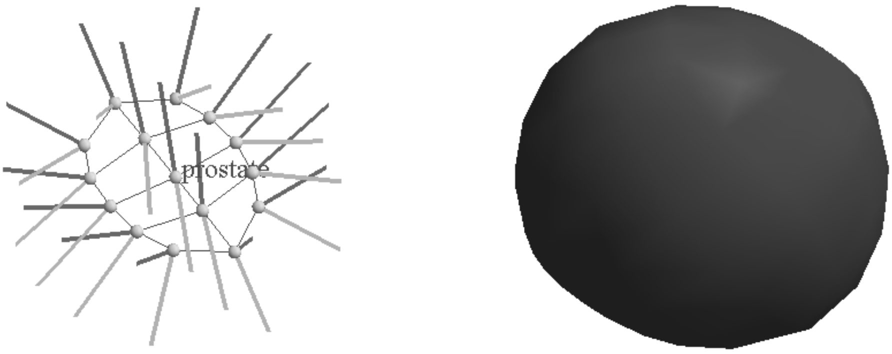


Figure 6.8: (left) An m-rep model with 3×5 grids of medial atoms. Each atom has its location ($\mathbb{R}^3$), and two equal-length spokes ($\mathbb{R}_+ \otimes S^2 \otimes S^2$) (right) The implied surface of the m-rep model, showing a prostate shape.

The space of prostate m-reps is then $\mathcal{M} = (\mathbb{R}^3 \otimes \mathbb{R}_+ \otimes S^2 \otimes S^2)^{15}$, which is a 120-dimensional direct product manifold with 60 components. The m-rep model provides a useful framework for describing shape variability in intuitive terms. See Siddiqi and Pizer (2008) and Pizer et al. (2003) for detailed introduction to and discussion of this subject.

An important topic in medical imaging is developing segmentation methods of 3D objects from CT images, see Cootes and Taylor (2001) and Pizer et al. (2007). A popular approach is similar to a Bayesian estimation scheme, where the knowledge of anatomic geometries is used (as a prior) together with a measure of how the segmentation matches the image (as a likelihood). A prior probability distribution is modeled using m-reps as a means of measuring geometric atypicality of a segmented object. PCA-like methods (including PAA) can be used to reduce the dimensionality of such a model. A detailed description can be found in Pizer et al. (2007).

6.5.2 Simulated m-rep object

The data set partly plotted in Figure 6.1 is from the generator discussed in Jeong et al. (2008). It generates random samples of objects whose shape changes and motions are physically modeled (with some randomness) by anatomical knowledge of the bladder, prostate and rectum in the male pelvis. Jeong et al. have proposed and used the generator to estimate the probability distribution model of shapes of human organs.

In the dataset of 60 samples of prostate m-reps we studied, the major motion of prostate is a rotation. In some S^2 components, the variation corresponding to the rotation is along a small circle. Therefore, PAA should fit better for this type of data than principal geodesics. To make this advantage more clear, we also show results from a dataset by removing the location and the spoke length information from the m-reps, the sample space of which is then $\{S^2\}^{30}$.

We have applied PAA as described in the previous section. The ratios μ/σ, estimated for the 30 S^2 components, are in general large (with minimum 21.2, median 44.1, and maximum 118), which suggests use of small circles to capture the variation.

Figure 6.9 shows the proportion of the cumulative variances, as a function of number of components, from the Principal Geodesic Analysis (PGA) of Fletcher et al. (2004) and PAA. In both cases, the first principal arc leaves smaller residuals than the first principal geodesic. What is more important is illustrated in the scatterplots of the data projected onto the first two principal components. The quadratic form of variation that requires two PGA components is captured by a single PAA component.

In Figure 6.10, the first principal arc is plotted in some marginal components of the m-rep space $\mathcal{M}$. Note that for the first and third S^2s (Figs 6.10a and 6.10c), the first principal arc fits the data very well, while the first principal geodesic does not. When the radius of the principal circle is close to $\pi/2$ (Figure 6.10b), two methods give similar results.

The probability distribution model estimated by principal geodesics is qualitatively different from the distribution estimated by PAA. Although the difference in the proportion of variance captured is small, the resulting distribution from PAA is no longer elliptical. In this sense, PAA gives a convenient way to describe a non-elliptical distribution by e.g. a Normal density.

6.5.3 Prostate m-reps from real patients

We also have applied PAA to a prostate m-rep data set from real CT images. Our data consist of five patients' image sets each of which is a series of CT scans containing prostate taken during a series of radiotherapy treatments. (Merck et al. (2008)) The prostate in each image is manually segmented by experts and an m-rep model is fitted. The patients, coded as 3106, 3107, 3109, 3112, and 3115, have different numbers of CT scans (17, 12, 18, 16, and 15 respectively). We have in total 78 m-reps.

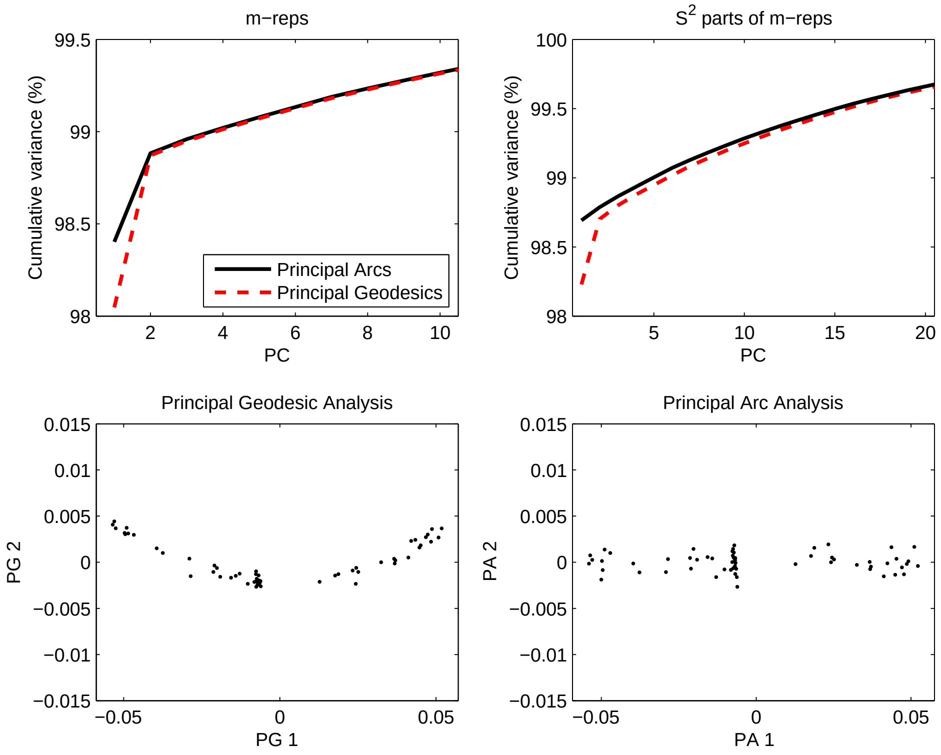


Figure 6.9: (Top) The proportion of variances captured by the first few components of PAA are compared to those from PGA for the simulated prostate m-reps. (Bottom) Scatter plots of the data on $\{S^2\}^{30}$ show that the major variation is explained more concisely by the first principal arc.

The proportion of variation captured in the first principal arc is 40.89%, slightly higher than the 40.53% of the first principal geodesic. Also note that the estimated probability distribution model from PAA is different from that of PGA. In particular, PAA gives a better separation of patients in the first two components, as depicted in the scatter plots (Figure 6.11).

6.6 Doubly iterative algorithm to find the least-squares small circle

We propose an algorithm to fit the least-squares small circle (6.1), which is a constrained nonlinear minimization problem. This algorithm is best understood in two iterative steps: The outer loop approximates the sphere by a tangent space; The inner loop solves an optimization problem in the linear space, which is much easier than solving (6.1) directly. In more detail,

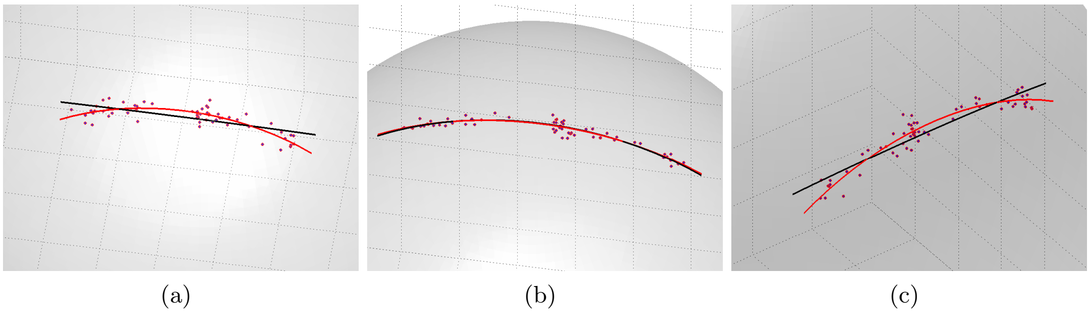

(a) (b) (c)

Figure 6.10: The first principal arc projected on the 5, 10, 15th S^2-components of the simulated prostate m-reps. Each sphere is magnified. The principal arc (more curved) is plotted from -2 to $+2$ standard deviation passing through the principal circle mean. The principal geodesic also covers ±2 standard deviations.

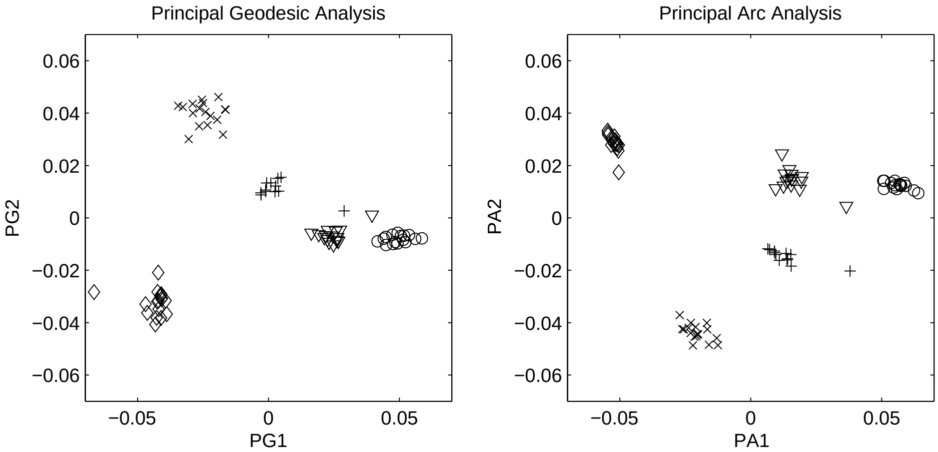


Figure 6.11: The scatter plots of the real prostate m-rep data. Different symbols represent different patients. PAA (right) gives a better separation of different patients in the first two components compared to PGA (left).

the $k+1$th iteration works as follows. The sphere is approximated by a tangent plane at $\mathbf{c}_k$, the kth solution of the center of the small circle. For the points on the tangent plane, any iterative algorithm to find a least-squares circle can be applied as an inner loop. The solution of the inner iteration is mapped back to the sphere and becomes the $k+1$th input of the outer loop operation. One advantage of this algorithm lies in the reduced difficulty of the optimization task. The inner loop problem is much simpler than (6.1) and the outer loop is calculated by a closed-form equation, which leads to a stable and fast algorithm. Another advantage can be obtained by using the *exponential map* and *log map* (6.12) for the tangent projection, since they preserve the distance from the point of tangency to the others, i.e. $\rho(\mathbf{x}, \mathbf{c}) = ||\mathrm{Log}_{\mathbf{c}}(\mathbf{x})||$ for any $\mathbf{x} \in S^2$. This is also true for radii of circles. The exponential map transforms a circle in $\mathbb{R}^2$ centered at the origin with radius r to $\delta(\mathbf{c}, r)$. Thus whenever (6.1) reaches its minimum, the algorithm does not alter the solution.

We first illustrate necessary building blocks of the algorithm. A tangent plane $T_{\mathbf{c}}$ at $\mathbf{c}$ can be defined for any $\mathbf{c}$ in S^2, and an appropriate coordinate system of $T_{\mathbf{c}}$ is obtained as follows. Basically, any two orthogonal complements of the direction $\mathbf{c}$ can be used as coordinates of $T_{\mathbf{c}}$. For example, when $\mathbf{c} = (0,0,1)' \equiv \mathbf{e}_3$, a coordinate system is given by $\mathbf{e}_1$ and $\mathbf{e}_2$. For a general $\mathbf{c}$, let $q_{\mathbf{c}}$ be a rotation operator on $\mathbb{R}^3$ that maps $\mathbf{c}$ to $\mathbf{e}_3$. Then a coordinate system for $T_{\mathbf{c}}$ is given by the inverse of $q_{\mathbf{c}}$ applied to $\mathbf{e}_1$ and $\mathbf{e}_2$, which is equivalent to applying $q_{\mathbf{c}}$ to each point of S^2 and using $\mathbf{e}_1$, $\mathbf{e}_2$ as coordinates.

The rotation operator $q_{\mathbf{c}}$ can be represented by a rotation matrix For $\mathbf{c} = (c_x, c_y, c_z)'$, the rotation $q_{\mathbf{c}}$ is equivalent to rotation through the angle $\theta = \cos^{-1}(c_z)$ about the axis $\mathbf{u} = (c_y, -c_x, 0)'/\sqrt{1-c_z^2}$, whenever $\mathbf{c} \neq \pm\mathbf{e}_3$. When $\mathbf{c} = \pm\mathbf{e}_3$, $\mathbf{u}$ is set to be $\mathbf{e}_1$. It is well known that a rotation matrix with axis $\mathbf{u} = (u_x, u_y, u_z)'$ and angle θ in radians is, for $c = \cos(\theta)$, $s = \sin(\theta)$ and $v = 1 - \cos(\theta)$,

$$R_{\mathbf{c}} = \begin{pmatrix} c + u_x^2 v & u_x u_y v - u_z s & u_x u_z v + u_y s \\ u_x u_y v + u_z s & c + u_y^2 v & u_y u_z v - u_x s \\ u_x u_z v - u_y s & u_y u_z v + u_x s & c + u_z^2 v \end{pmatrix}, \tag{6.11}$$

so that $q_{\mathbf{c}}(\mathbf{x}) = R_{\mathbf{c}}\mathbf{x}$, for $\mathbf{x} \in \mathbb{R}^3$.

With the coordinate system for $T_{\mathbf{c}}$, we shall define the exponential map $\mathrm{Exp}_{\mathbf{c}}$, a mapping from $T_{\mathbf{c}}$ to S^2, and the log map $\mathrm{Log}_{\mathbf{c}} = \mathrm{Exp}_{\mathbf{c}}^{-1}$. These are defined for $\mathbf{v} = (v_1, v_2) \in \mathbb{R}^2$ and $\mathbf{x} = (x_1, x_2, x_3)' \in S^2$, as

$$\mathrm{Exp}_{\mathbf{c}}(\mathbf{v}) = q_{\mathbf{c}} \circ \mathrm{Exp}_{\mathbf{e}_3}(\mathbf{v}), \mathrm{Log}_{\mathbf{c}}(\mathbf{x}) = \mathrm{Log}_{\mathbf{e}_3} \circ q_{\mathbf{c}}(\mathbf{x}), \tag{6.12}$$

for $\theta = \cos^{-1}(x_3)$. See Section 5.2.2 for $\mathrm{Exp}_{\mathbf{e}_3}$ and $\mathrm{Log}_{\mathbf{e}_3}$. Note that $\mathrm{Log}_{\mathbf{c}}(\mathbf{c}) = \mathbf{0}$ and $\mathrm{Log}_{\mathbf{c}}$ is not defined for the antipodal point of $\mathbf{c}$.

Once we have approximated each $\mathbf{x}_i$ by $\mathrm{Log}_{\mathbf{c}}(\mathbf{x}_i) \equiv \tilde{\mathbf{x}}_i$, the inner loop finds the minimizer $(\mathbf{v}, r)$ of

$$\min \sum_{i=1}^{n} \left(||\tilde{\mathbf{x}}_i - \mathbf{v}|| - r \right)^2, \tag{6.13}$$

which is to find the least-squares circle centered at $\mathbf{v}$ with radius r. The general circle fitting problem is discussed in e.g. Umbach and Jones (2003) and Chernov (2010). This problem is much simpler than (6.1) because it is an unconstrained problem and the number of parameters to optimize is decreased by 1. Moreover, optimal solution of r is easily found as

$$\hat{r} = \frac{1}{n} \sum_{i=1}^{n} ||\tilde{\mathbf{x}}_i - \mathbf{v}||, \tag{6.14}$$

when $\mathbf{v}$ is given. Note that for great circle fitting, we can simply put $\hat{r} = \pi/2$. Although the problem is still nonlinear, one can use any optimization method that solves non-linear least squares problems. We use the Levenberg-Marquardt algorithm, modified by Fletcher (1971) (see Chapter 4 of Scales (1985) and Chapter 3 of Bates and Watts (1988)), to minimize (6.13) with r replaced by $\hat{r}$. One can always use $\mathbf{v} = \mathbf{0}$ as an initial guess since $\mathbf{0} = \mathrm{Log}_{\mathbf{c}}(\mathbf{c})$ is the solution from the previous (outer) iteration.

The algorithm is now summarized as follows.

1. Given $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, $\mathbf{c}_0 = \mathbf{x}_1$.

2. Given $\mathbf{c}_k$, find a minimizer $\mathbf{v}$ of (6.13) with r replaced by (6.14), with inputs $\tilde{\mathbf{x}}_i = \mathrm{Log}_{\mathbf{c}_k}(\mathbf{x}_i)$.

3. If $\|\mathbf{v}\| < \epsilon$, then iteration stops with the solution $\hat{\mathbf{c}} = \mathbf{c}_k$, $r = \hat{r}$ as in (6.14). Otherwise, $\mathbf{c}_{k+1} = \mathrm{Exp}_{\mathbf{c}_k}(\mathbf{v})$ and go to step 2.

Note that the radius of the fitted circle in $T_{\mathbf{c}}$ is the same as the radius of the resulting small circle. There could be many variations of this algorithm: as an instance, one can elaborate the initial value selection by using the eigenvector of the sample covariance matrix of $\mathbf{x}_i$'s, corresponding to the smallest eigenvalue as done in Gray et al. (1980). Experience has shown that the proposed algorithm is stable and speedy enough. Gray *et al.* proposed to solve (6.1) directly, which seems to be unstable in some cases.

The idea of the doubly iterative algorithm can be applied to other optimization problem on manifolds. For example, the geodesic mean is also a solution of a nonlinear minimization, where the nonlinearity comes from the use of the geodesic distance. This can be easily solved by an iterative approximation of the manifold to a linear space (See Chapter 4 of Fletcher (2004)), which is the same as the gradient descent algorithms (Pennec (1999), Le (2001)). Note that the proposed algorithm, like other iterative algorithms, only finds one solution even if there are multiple solutions.

Chapter 7

Analysis of Principal Nested Spheres

The work presented in this chapter is based on and contained in Jung et al. (2011a).

7.1 Introduction

A general framework for a novel decomposition of hypersphere is introduced.The hypersphere is the sample space of directions (Fisher et al. (1993), Fisher (1993), Mardia and Jupp (2000)) and pre-shapes in Kendall's statistical theory of landmark shapes (Kendall (1984), Dryden and Mardia (1998)). The proposed decomposition method, Principal Nested Spheres (PNS), is a flexible extension of Principal Component Analysis (PCA) for curved manifolds. PCA provides an effective means of analyzing the main modes of variation of the dataset and also gives a basis for dimension reduction. There have been a number of extensions of PCA to manifold-valued data, most of which find principal geodesics (Fletcher (2004), Huckemann and Ziezold (2006), Huckemann et al. (2010), Kenobi et al. (2010)).

There has been a concern that when non-geodesic variation is major and apparent, the geodesic based PCA does not give a fully effective decomposition of the space. As an example, a dataset of shapes representing human movements, discussed later in Section 7.7 and introduced in Kume et al. (2007), is plotted in Figure 7.1, using the first two principal component directions. In this dataset and many other interesting real data sets, the major one dimensional variation of the data curves through at least two components, and thus at least two dimensions are needed to explain the major variation. PNS decomposes the data space in a way that the major one dimensional variation is linearly represented, as shown in the bottom of Figure 7.1.

For a unit d-sphere S^d, which is the set of unit vectors in $\mathbb{R}^{d+1}$, PNS gives a decomposition of S^d that captures the non-geodesic variation in a lower dimensional sub-manifold. The decomposition sequentially provides the best k-dimensional approximation $\mathcal{A}_k$ of the data for each $k = 0, 1, \ldots, d-1$. $\mathcal{A}_k$ is called the k-dimensional PNS, since it is essentially a sphere and is *nested* within (i.e. a sub-manifold of) the higher dimensional PNS. The sequence of PNS is then

$$\mathcal{A}_0 \subset \mathcal{A}_1 \subset \cdots \subset \mathcal{A}_{d-1} \subset S^d.$$

Since the preshape space of two dimensional landmark based shapes is also a hypersphere, the method can be readily applied to shape data, with some modifications (see Section 7.6). The analysis of PNS provides intuitive approximations of the directional or shape data for every dimension, captures the non-geodesic variation, and provides intuitive visualization of the major variability in terms of shape changes.

The procedure of fitting PNS involves iterative reduction of the dimensionality of the data. We first fit a $d-1$ dimensional *subsphere* $\mathcal{A}_{d-1}$ of S^d that best approximates the data. This subsphere is not necessarily a great sphere (i.e. a sphere with radius 1, analogous to the great circle for S^2), which makes the resulting decomposition non-geodesic. Nevertheless, $\mathcal{A}_{d-1}$ can be treated as if it was the unit $(d-1)$-sphere by some geometric facts discussed in Section 7.2.1 and in Appendix 7.8 in greater detail. Each data point has an associated residual, which is the geodesic distance to its projection on $\mathcal{A}_{d-1}$. Then for the data projected onto the subsphere, we continue to search for the best fitting $d-2$ dimensional subsphere. These steps are iterated to find lower dimensional PNS. A detailed discussion of the procedure is in Section 7.2. For visualization and further analysis, we obtain an Euclidean-type representation of the data, essentially consisting of the residuals of each level. The first two coordinates of this representation, related to the one and two dimensional PNS, applied to the human movement data are plotted in Figure 7.1.

In Figure 7.1, PNS (bottom panel) has less curving variation. The proportion of variance in the 1-d PNS is almost the proportion of the sum of the first two geodesic component variances. That is, the variation explained by two geodesic components is attained in only

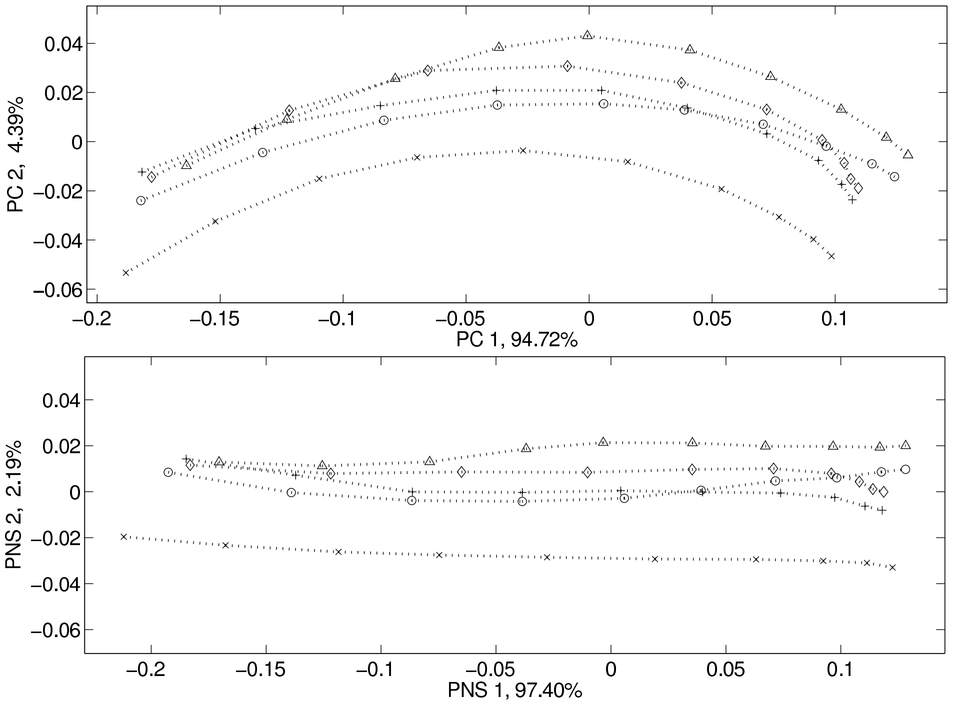


Figure 7.1: Human movement data: (top) Scatter plot of two major geodesic components, where the different symbols represent different tasks, and samples for each task are interpolated. (bottom) Scatter plot of the first two PNS. The number% is the percent variance explained. The first PNS captures more of the interesting variation.

one component of the PNS. Moreover, the graph in the top panel is indeed obtained by a special case of PNS analysis, which is similar to the geodesic-based PCA, as discussed in Section 7.2.6.

In Section 7.3, a visualization of a particular sphere is presented to give a way of understanding the decomposition. The procedure of PNS tends to find smaller spheres than the great sphere. Since this may cause an overfitting of the data, we developed a test procedure that can be applied to each layer to prevent the overfitting (see Section 7.4). A computational scheme for fitting PNS is proposed in Section 7.5. Necessary considerations and modifications for planar shape data are discussed in Section 7.6. In Section 7.7, we describe applications of the method to several interesting real datasets. We conclude with some geometric background to help define PNS, and proofs of the theorems.

7.2 Principal Nested Spheres

In this section, we provide a detailed procedure for fitting PNS to data on S^d, and discuss a Euclidean-type representation and principal arcs by PNS. Moreover, an important special case of PNS, Principal Nested Great spheres, is discussed.

7.2.1 Geometry of Nested Spheres

We begin with describing essential geometric background for nested spheres. More detailed discussion of the arguments in this section can be found in Section 7.8.

For a unit sphere S^d, a geodesic joining any two points is a great circle joining the two points. A natural distance function on S^d is the Riemannian (geodesic) distance function $\rho_d(\cdot, \cdot)$ defined as the length of the shortest great circle segment joining $\mathbf{x}, \mathbf{y} \in S^d$, $\rho_d(\mathbf{x}, \mathbf{y}) = \cos^{-1}(\mathbf{x}^T \mathbf{y})$.

A sequence of nested spheres of S^d is roughly a set of lower dimensional submanifolds that are essentially spheres. In the following, the precise form of nested spheres is introduced. We first define a subsphere of S^d, which induces the nested spheres.

Definition 7.1. A *subsphere* A_{d-1} of S^d is defined by an orthogonal axis $\mathbf{v} \in S^d$ and a distance $r \in (0, \pi/2]$, as follows:

$$A_{d-1}(\mathbf{v}, r) = \{\mathbf{x} \in S^d : \rho_d(\mathbf{v}, \mathbf{x}) = r\},$$

where $\rho_d(\cdot, \cdot)$ is the geodesic distance function on S^d, $d \geqslant 2$.

The subsphere A_{d-1} can be viewed as an intersection of $S^d \subset \mathbb{R}^{d+1}$ and an affine d dimensional hyperplane, $\{\mathbf{x} \in \mathbb{R}^{d+1} : \mathbf{v}^T \mathbf{x} - \cos(r) = 0\}$. In other words, A_{d-1} is identified with a “slicing” of S^d with the affine hyperplane, an example of which is illustrated as a shaded plane in Fig 7.2. A subsphere A_{d-1} is indeed a $d-1$ dimensional nested sphere $\mathcal{A}_{d-1}$ of S^d.

The subsphere A_{d-1} is isomorphic to S^{d-1}, as shown in Proposition 7.7, so we can treat the subsphere as a unit sphere S^{d-1}. This is done by an isomorphism $f_1 : A_{d-1} \to S^{d-1}$ and its inverse f_1^{-1}, defined in Eq. 7.1 below and also depicted in Figure 7.2. Now a subsphere

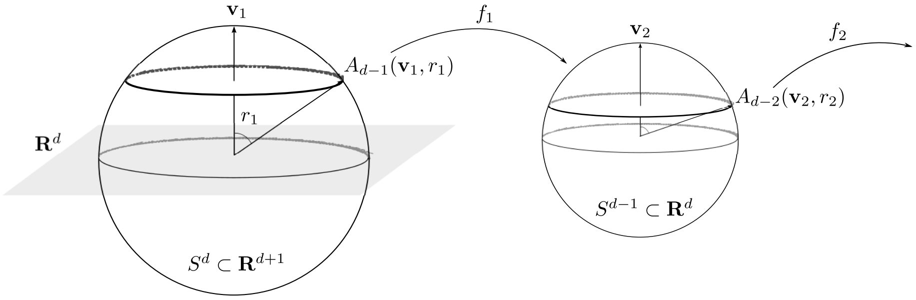


Figure 7.2: The subsphere $A_{d-1}(\mathbf{v}_1, r_1)$ in S^d and its relation to S^{d-1}, through the isomorphism f_1. Recursively, $A_{d-2}(\mathbf{v}_2, r_2)$ is found in S^{d-1}, and is isomorphic to S^{d-2} with the isomorphism f_2.

A_{d-2} of S^{d-1} can be obtained by applying Def. 7.1 with dimension d reduced by 1. For a general subsphere A_{d-k} of S^{d-k+1}, where $k = 1, \ldots, d-1$, we also use the isomorphic transformation $f_k : A_{d-k} \to S^{d-k}$ and its inverse f_k^{-1}. Let $m = d-k+1$, so that the subsphere $A_{d-k} \in S^m \subset \mathbb{R}^{m+1}$. The transformations are defined by $\mathbf{v}_k \in S^m$ and $r_k \in (0, \pi/2]$ as

$$\begin{aligned} f_k(\mathbf{x}) &= \frac{1}{\sin(r_k)} \mathbf{R}^-(\mathbf{v_k})\mathbf{x}, \ \mathbf{x} \in A_{d-k}, \\ f_k^{-1}(\mathbf{x}^\dagger) &= \mathbf{R}^T(\mathbf{v_k}) \begin{bmatrix} \sin(r_k) \cdot \mathbf{x}^\dagger \\ \cos(r_k) \end{bmatrix}, \ \mathbf{x}^\dagger \in S^{d-k}, \end{aligned} \tag{7.1}$$

where $\mathbf{R}(\mathbf{v_k})$ is the $(m+1) \times (m+1)$ rotation matrix that moves $\mathbf{v_k}$ to the north pole (see Appendix 7.8), and $\mathbf{R}^-(\mathbf{v_k})$ is the $m \times (m+1)$ matrix consisting of the first m rows of $\mathbf{R}(\mathbf{v_k})$.

The subspheres A_{d-k} are defined in different spaces (in S^{d-k+1} for each k). A nested sphere is defined by the subsphere located in the original space S^d.

Definition 7.2. A $d-k$ *dimensional nested sphere* $\mathcal{A}_{d-k}$ of S^d is defined as

$$\mathcal{A}_{d-k} = \begin{cases} f_1^{-1} \circ \cdots \circ f_{k-1}^{-1}(A_{d-k}) & \text{if } k = 2, \ldots, d-1 \\ A_{d-1} & \text{if } k = 1 \end{cases}$$

A $d-k$ dimensional nested sphere $\mathcal{A}_{d-k}$ is indeed identified with a slicing of S^d by a $d-k+1$ dimensional affine hyperplane. Note that, however, we work with each S^{d-k}, as it is logically simple in terms of dimensionality reduction as described in Section 7.2.3.

7.2.2 The Best Fitting Subsphere

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be samples in S^d, $d \geqslant 2$. We first define the residual ξ of $\mathbf{x}$ from a subsphere $A_{d-1}(\mathbf{v}_1, r_1)$ of S^d as the signed length of the minimal geodesic that joins $\mathbf{x}$ to A_{d-1}. Then $\xi = \rho_d(\mathbf{x}, \mathbf{v}_1) - r_1$. The sign of ξ is negative if $\mathbf{x}$ is in the interior of the geodesic ball corresponding to A_{d-1}, and is positive if $\mathbf{x}$ is in the exterior.

The *best fitting subsphere* $\widehat{A}_{d-1} \equiv A_{d-1}(\hat{\mathbf{v}}_1, \hat{r}_1)$ is found by minimizing the sum of squares of residuals of the data points to $\widehat{A}_{d-1}$. In other words, $\hat{\mathbf{v}}_1$ and $\hat{r}_1$ minimize

$$\sum_{i=1}^{n} \xi_i(\mathbf{v}_1, r_1)^2 = \sum_{i=1}^{n} \{\rho_d(\mathbf{x}_i, \mathbf{v}_1) - r_1\}^2, \tag{7.2}$$

among all $\mathbf{v}_1 \in S^d$, $r_1 \in (0, \pi/2]$.

Note that the method can be extended using other objective functions, e.g. sum of absolute deviations for more robust fitting.

Each $\mathbf{x}_i$ can be projected on $\widehat{A}_{d-1}$ along the minimal geodesic that joins $\mathbf{x}$ to $\widehat{A}_{d-1}$. The projection of $\mathbf{x}$ onto A_{d-1} is defined as

$$P(\mathbf{x}; A_{d-1}(\mathbf{v}, r)) = \frac{\sin(r)\mathbf{x} + \sin(\rho_d(\mathbf{x}, \mathbf{v}) - r)\mathbf{v}}{\sin\{\rho_d(\mathbf{x}, \mathbf{v})\}}. \tag{7.3}$$

Denote $\mathbf{x}^P = P(\mathbf{x}; \widehat{A}_{d-1}) \in \widehat{A}_{d-1}$ for the projected $\mathbf{x}$. We use the isomorphism $\hat{f}_1 \equiv f(\hat{\mathbf{v}}_1, \hat{r}_1)$, as defined in Eq. (7.1), to transform $\widehat{A}_{d-1}$ to S^{d-1} so that $\hat{f}_1(\mathbf{x}_i^P) \in S^{d-1}$.

7.2.3 The sequence of Principal Nested Spheres

The sequence of PNS are fully meaningful when they are in the same space. On the other hand, utilizing the isomorphic spaces of the nested spheres, i.e. the unit spheres, makes the process simpler. Therefore the procedure to find the sample PNS consists of iteratively finding the best fitting subsphere and mapping to the original space.

The $d-1$ dimensional sample PNS $\widehat{\mathcal{A}}_{d-1}$ is the same as the best fitting subsphere $A_{d-1}(\hat{\mathbf{v}}_1, \hat{r}_1)$ because both are in the original space S^d. The second layer, the $d-2$ dimensional sample PNS, is obtained from the subsphere that best fits $\hat{f}_1(\mathbf{x}_i^P) \in S^{d-1}$. The best fitting subsphere $A_{d-2}(\hat{\mathbf{v}}_2, \hat{r}_2)$ is then mapped to S^d by the relevant isomorphism f_1^{-1} and becomes $\widehat{\mathcal{A}}_{d-2}$.

In general, we recursively find the sequence of best fitting subspheres from the projected and transformed samples, i.e. $\mathbf{x} \mapsto \hat{f}_k(P(\mathbf{x}; \widehat{A}_{d-k}))$. In the '$k$th level', where we fit a subsphere from S^{d-k+1}, we denote the best fitting subsphere as $\widehat{A}_{d-k} \equiv A_{d-k}(\hat{\mathbf{v}}_k, \hat{r}_k)$ and keep residuals $\xi_i \equiv \xi_{i,d-k}$, $i = 1, \dots, n$, for later use as analogs of principal component scores.

The lowest level best fitting subsphere $\widehat{A}_1$ is then a small circle isomorphic to S^1. No further sphere or circle can be used to reduce the dimensionality. Instead, we find the Fréchet mean (Fréchet (1944, 1948) and Karcher (1977)) $\widehat{A}_0$ of $\mathbf{x}_1^\dagger, \dots, \mathbf{x}_n^\dagger$ (the projected and transformed samples in S^1) which can be thought of as a best 0-dimensional representation of the data in the framework of PNS. The Fréchet mean $\widehat{A}_0$ is defined as the minimizer of the squared distances to the $\mathbf{x}_i^\dagger$s, i.e.

$$\widehat{A}_0 = \operatorname*{argmin}_{\mathbf{x} \in S^1} \sum_{i=1}^{n} \rho_1(\mathbf{x}, \mathbf{x}_i^\dagger)^2.$$

The Fréchet mean is unique when the support of $\mathbf{x}_i^\dagger$ is a proper subset of a half circle in S^1, which is often satisfied in practice. If there are multiple Fréchet means, then careful inspection of the data must be followed. A typical case for having multiple means is that the data are uniformly distributed on the circle. If this is the case, then $\widehat{A}_0$ can be chosen to be any solution of the above criterion, and since it does not summarize the data well we may not lay much emphasis on $\widehat{A}_0$.

The sequence of best fitting subspheres including the $\widehat{A}_0$ can be located in the original space S^d, as follows.

Definition 7.3. The *sequence of sample Principal Nested Spheres* in S^d is then $\{\widehat{\mathcal{A}}_0, \widehat{\mathcal{A}}_1, \dots, \widehat{\mathcal{A}}_{d-1}\}$, where

$$\widehat{\mathcal{A}}_{d-k} = \begin{cases} \hat{f}_1^{-1} \circ \cdots \circ \hat{f}_{k-1}^{-1}(\widehat{A}_{d-k}) & \text{if } k = 2, \dots, d, \\ \widehat{A}_{d-1} & \text{if } k = 1. \end{cases}$$

We call $\widehat{\mathcal{A}}_0$ the *PNSmean.*

7.2.4 Euclidean-type representation

We wish to represent the data in an Euclidean space for visualization and further analysis.

Recall that in the kth level of the procedure, we have collected the signed residuals which we denote by $\xi_{i,d-k}$, $i = 1, \dots, n$. These were measured by the metric ρ_{d-k} in a space different from S^d. Therefore we scale these residuals by multiplying $\prod_{i=1}^{k-1} \sin(\hat{r}_i)$ which makes the magnitude of residuals commensurate (see Proposition 7.9). We put the scaled residuals in a row vector

$$\Xi(d-k)_{1\times n} \doteq \prod_{i=1}^{k-1} \sin(\hat{r}_i)[\xi_{1,d-k}, \ldots, \xi_{n,d-k}].$$

We further define $\xi_{i,0}$ as the ith sample's signed deviation from $\widehat{A}_0$ measured by ρ_1. Similar to before, rescale the deviations and let

$$\Xi(0)_{1\times n} \doteq \prod_{i=1}^{d-1} \sin(\hat{r}_i)[\xi_{1,0}, \ldots, \xi_{n,0}].$$

These commensurate residuals are combined into a $d \times n$ data matrix

$$\widehat{X}_{PNS} = \begin{bmatrix} \Xi(0) \\ \Xi(1) \\ \vdots \\ \Xi(d-1) \end{bmatrix},$$

where each column is the corresponding sample's coordinates in terms of the sample PNS. Each entry in row k works like the kth principal component score.

The data matrix $\widehat{X}_{PNS}$ can be used to visualize the structure of the data. For example, the graph in Figure 7.1 is a scatterplot of $\Xi(0)$ and $\Xi(1)$. The variance of each component is defined by the variance of the corresponding residuals. Moreover, conventional multivariate statistics based on Euclidean space can be applied to $\widehat{X}_{PNS}$ for further analysis (e.g. PCA and classification methods).

7.2.5 Principal Arcs

In analogy to the principal component directions in Euclidean space, or the manifold extension principal geodesics, the *principal arcs* that represent the *direction* of major variations are defined by PNS. These arcs are space curves lying in the manifold S^d, which frequently are not equivalent to any geodesic.

Given a sequence of PNS $\{\widehat{\mathcal{A}}_0, \widehat{\mathcal{A}}_1, \ldots, \widehat{\mathcal{A}}_d - 1\}$, the first principal arc coincides with the 1-d PNS $\widehat{\mathcal{A}}_1$. This arc may be parameterized by the signed distance from the PNSmean $\widehat{\mathcal{A}}_0$.

In the space of the Euclidean-type representation $\widehat{X}_{PNS}$, the first principal arc coincides with the direction $\mathbf{e}_1 = (1, 0, \dots, 0)^T$.

The second principal arc lies in $\widehat{\mathcal{A}}_2$ and is orthogonal to the first principal arc at all points in common. The first and second arcs cross at $\widehat{\mathcal{A}}_0$ and also at the farthest point from $\widehat{\mathcal{A}}_0$ on $\widehat{\mathcal{A}}_1$. The second arc is in general a small cirle in S^d but is identified with a great circle in S^2, the isomorphic space of $\widehat{\mathcal{A}}_2$. The second principal arc in S^2 must pass through the axis $\mathbf{v}_{d-1}$ in order to be orthogonal to the first. This arc may be parameterized by the signed distance from $\widehat{\mathcal{A}}_0$, and coincides with the direction $\mathbf{e}_2$ in the space of $\widehat{X}_{PNS}$.

The higher order principal arcs are defined in the same manner. The kth principal arc can be defined and identified with the direction $\mathbf{e}_k$ in the space of $\widehat{X}_{PNS}$. The kth arc is then orthogonal to the principal arcs of order $1, \dots, k-1$, and passes through the PNSmean.

In addition, a space curve in S^d may be parameterized by the coordinates of the Euclidean-type representation. Fitting a space curve that further smooths the data is a separate issue.

7.2.6 Principal Nested Spheres restricted to Great Spheres

An important special case of the PNS is obtained by setting $r = \pi/2$ for each subsphere fitting. This restriction leads to the nested spheres being great spheres, and the principal arcs become geodesics. In all data sets we tested, the resulting principal geodesics are similar to the previous geodesic-based PCA methods. In the following, we indicate this special case as *Principal Nested Great spheres* (PNG).

We conjecture that the principal geodesics, found by PNG, are more similar to the Geodesic Principal Component of Huckemann et al. (2010) than the usual tangent space projection methods. This is mainly because any pre-determined mean (either geodesic mean or Procrustes mean) is not used in PNG nor Huckemann's. The PNSmean in this special case is similar to the notion of mean of Huckemann, and is identical when the sphere has dimension $d = 2$. Although we have not yet found a significant difference of PNG than previous methods, we point out that the PCA extension approach of PNG (and PNS) is different from those methods.

7.3 Visualization of nested spheres of S^3

In this section, a visualization of nested spheres of S^3 is used to give insight into the meaning of fitting A_2 from S^3 in a non-geodesic way, i.e. $r < \pi/2$. This gives a low dimensional answer to a central question: what is the meaning of $A_{d-k}(\mathbf{v}, r)$ with $r < \pi/2$? Note that it is difficult to visualize a space with dimension greater than 3.

A stereographic projection can be used for visualization of the 3-sphere S^3. Specifically, to visualize the nested spheres of S^3, we use the stereographic projection with projection point $\mathbf{p}$ as the antipodal point from the PNSmean. With $\mathbf{p} = (0, 0, 0, -1) \in S^3$, the stereographic coordinates of $\mathbf{x} = (x_1, x_2, x_3, x_4) \in S^3$ are given by

$$\frac{1}{1 + x_4}(x_1, x_2, x_3).$$

Note that one can move the PNSmean (and all other points) to the north pole without losing any information, which is done by the rotation operator $\mathbf{R}(\mathbf{p})$ defined in Section 7.8. The stereographic projection defined above maps the PNSmean at the north pole to the origin of the new coordinates. Moreover, any geodesic (great circle) passing through the PNSmean is mapped to a straight line, and any great sphere containing the PNSmean is mapped to a flat plane through the origin. On the other hand, a small sphere or small circle passing through the PNSmean is mapped to a sphere or circle.

Decomposition of S^3 by PNS is a sequential reduction of dimensionality. We first find a two dimensional PNS $\mathcal{A}_2$. When the underlying variation of the data is along a surface with more curvature than great spheres, then PNS shall use a small sphere for $\mathcal{A}_2$. This corresponds to the projected images in Figure 7.3, cases c and d, where the meshed surface is the image of $\mathcal{A}_2$. Likewise, when fitting $\mathcal{A}_1$ from $\mathcal{A}_2$, whether to use a geodesic or small circle depends on the underlying variation. The one dimensional PNS $\mathcal{A}_1$ is depicted as a line or arc in Figure 7.3.

There are four different possible situations for the PNS decomposition of S^3. Recall that from S^3, r_1 determines the radius of the two dimensional nested sphere $\mathcal{A}_2$ (2-sphere), and r_2 together with r_1 determines the radius of the $\mathcal{A}_1$ (circle). The image of $\mathcal{A}_1$ ($\subset \mathcal{A}_2$) by f_1 is A_1, which is the fitted subsphere in S^2, the isomorphic space of $\mathcal{A}_2$. The four possible

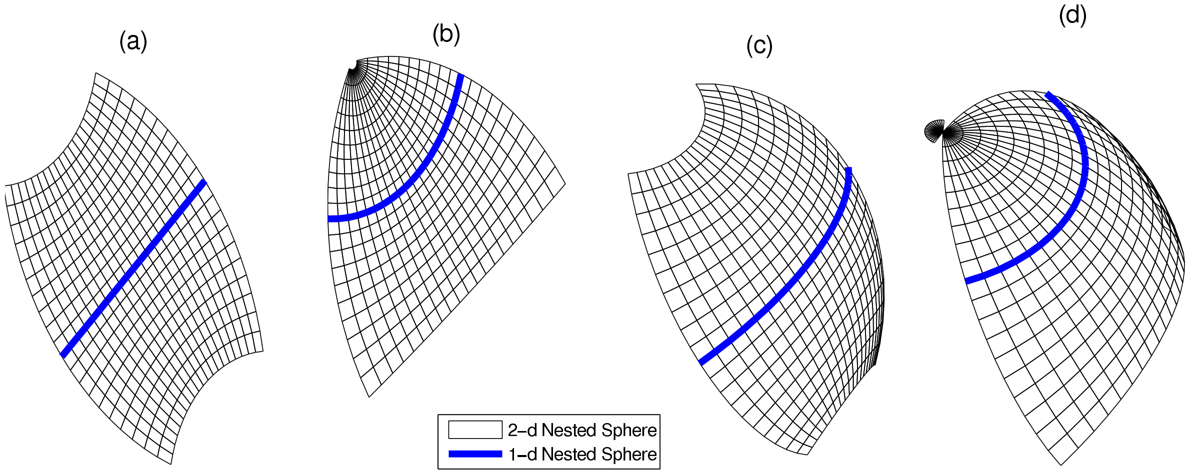


Figure 7.3: Stereographic projection of S^3, with four different decompositions by PNS. The two dimensional PNS $\mathcal{A}_2$ is depicted as the meshed surface. The one dimensional PNS $\mathcal{A}_1$ is nested in $\mathcal{A}_2$ and is illustrated as a line or arc. In the first level of PNS fitting, where we fit $\mathcal{A}_2$ from S^3, the cases a and b correspond to $\mathcal{A}_2$ with $r = \pi/2$ (great sphere), while cases c and d correspond to $\mathcal{A}_2$ being a small sphere. Furthermore, in the level of fitting $\mathcal{A}_1$, as part of the PNS fitting we test whether the major variation $\mathcal{A}_1$ is along a geodesic in $\mathcal{A}_2$ (cases a and c), or along a small circle (cases b and d).

decompositions are characterized as:

(a) $r_1 = \pi/2$ ($\mathcal{A}_2$ great sphere), $r_2 = \pi/2$ ($\mathcal{A}_1$ great circle).

(b) $r_1 = \pi/2$ ($\mathcal{A}_2$ great sphere), $r_2 < \pi/2$ ($\mathcal{A}_1$ small circle).

(c) $r_1 < \pi/2$ ($\mathcal{A}_2$ small sphere), $r_2 = \pi/2$ (A_1 great circle).

(d) $r_1 < \pi/2$ ($\mathcal{A}_2$ small sphere), $r_2 < \pi/2$ (A_1 small circle).

These four cases are illustrated in Figure 7.3 with $r_i = \pi/3$ for the small sphere/circle case. Case (a) corresponds to PNG, i.e. major variation in any dimension is along a geodesic or a great sphere. Case (b) is when the best two dimensional approximation is a great sphere, but the best one dimensional approximation is not a geodesic. $\mathcal{A}_2$ being small sphere (cases c and d) can be understood as the major two dimensional variation is along a small 2-sphere with more curvature than the great sphere. Once we reduce the dimensionality to two by $\mathcal{A}_2$, we find either a geodesic of $\mathcal{A}_2$ (case c) or a small circle of $\mathcal{A}_2$ (case d) as $\mathcal{A}_1$, according to the curvature of underlying variation.

Similar understanding follows for S^d with any d, when fitting A_{d-1}. In particular, when the underlying variation in S^d is best captured by a $(d-1)$ dimensional curved sub-manifold (with more curvature than a great sphere), we then use $\mathcal{A}_{d-1}$ with $r < \pi/2$.

7.4 Prevention of overfitting by sequential tests

In this section, the significance of small spheres against the great sphere is discussed. We propose a test procedure consisting of two different tests for each level of subsphere fitting. Similar to the backward regression procedure, sequentially testing small spheres at each layer may prevent overfitting.

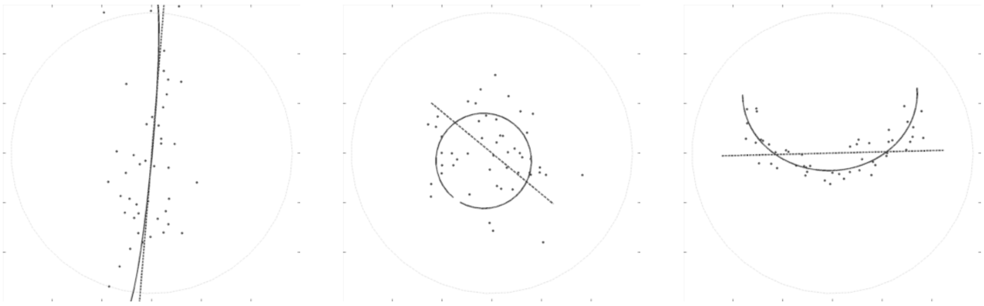
There are two cases where a great sphere provides more appropriate fit to the data, yet the sum of squared residuals is minimized by a small sphere. The first case is where a true major variation is along a great sphere, an example of which on S^2 is illustrated in Fig 7.4a. The second case is when the underlying distribution is isotropic with a single mode, so that there is no major variation along any direction. An example of such a distribution is $N(\mathbf{0}, \mathbf{I}_k)$ (in linear space), or the von Mises–Fisher distribution on S^d (Fisher (1953), Mardia and Jupp (2000)), as illustrated in Fig 7.4b. In this situation, small spheres centered at the point of isotropy are frequently obtained, which do not give a useful decomposition.

We have developed two different tests to handle these cases. The first is a likelihood ratio test (LRT) for the detection of the first case above (Fig 7.4a), which tests the significance of the reduction of residual variances. The second is a parametric bootstrap test aimed at the second case above(Fig 7.4b), which tests the isotropy of the underlying distribution. A detailed description of the tests is given in the following subsections. A procedure to apply these tests to PNS fitting is then discussed in Section 7.4.3.

7.4.1 Likelihood ratio test

We define a likelihood ratio statistic for each level to sequentially test the significance of small sphere fitting against the great sphere.

For the kth level of the procedure, where A_{d-k} is fitted to $\mathbf{x}_1, \ldots, \mathbf{x}_n \in S^{d-k+1}$, we assume that the deviations of the samples $\mathbf{x}_i$ from the subsphere $A_{d-k}(\mathbf{v}, r)$ are independent $N(0, \sigma^2)$. It makes more sense when a truncated Normal distribution on a range $[-\pi/2, \pi/2]$ is assumed. However unless the data spread too widely (e.g. Uniform on the sphere), the distribution will



(a) (b) (c)

Figure 7.4: Simulated data examples on S^2 projected by an equal area projection, and the fitted best small (solid) and great (dotted) spheres, which are arcs in this two dimensional case. (a) The LRT gives the p-value 0.338, while p-value of the bootstrap test is 0. The LRT detects the overfitting. (b) The LRT leads to p-value $\approx$ 0), and the bootstrap p-value is 0.3. The bootstrap test detects the overfitting. (c) When the fitted small sphere (circle) is not overfitted, both tests give very small p-values ($\approx$ 0). This assures that the small sphere is not overly fitted.

be approximately Normal. Thus we use the approximate likelihood function of $(\mathbf{v}, r, \sigma^2)$, given by

$$L(\mathbf{v}, r, \sigma^2 | \mathbf{x}_1^n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\rho(\mathbf{x}_i, \mathbf{v}) - r)^2\right),$$

where ρ is the geodesic distance function on S^{d-k+1}. The approximate maximum likelihood estimator (m.l.e.) of $(\mathbf{v}, r)$ coincides with $(\hat{\mathbf{v}}, \hat{r})$, the solution of Eq. (7.2), and the approximate m.l.e. of σ^2 is given by $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}\xi_i(\hat{\mathbf{v}}, \hat{r})^2$, which is obtained by differentiating the log-likelihood function and setting the derivative equal to zero.

We can test $H_{0a} : r = \pi/2$ (i.e. the great sphere), versus $H_{1a} : r < \pi/2$ (i.e. some small sphere), using a likelihood ratio test. The m.l.e. of $(\mathbf{v}, r, \sigma^2)$ under H_{0a} is given by $(\hat{\mathbf{v}}^0, \pi/2, \hat{\sigma}_0^2)$, where $\hat{\mathbf{v}}^0$ minimizes the sum of squared residuals of Eq. (7.2) with $r = \pi/2$, and $\hat{\sigma}_0^2 = n^{-1}\sum_{i=1}^{n}\xi_i(\hat{\mathbf{v}}^0, \pi/2)^2$. The log-likelihood ratio is $(\hat{\sigma}^2/\hat{\sigma}_0^2)^{-n/2}$. Then using Wilks' theorem, for large samples $n\log(\hat{\sigma}^2/\hat{\sigma}_0^2) \approx \chi_1^2$ under H_{0a}, and the test rejects H_{0a} in favor of H_{1a} for large values of $n\log(\hat{\sigma}^2/\hat{\sigma}_0^2)$.

7.4.2 Parametric bootstrap test

For each level of PNS fitting, suppose $\mathbf{X} \in S^m$ has a distribution function $F_{\mathbf{X}}$. We wish to test for the underlying distribution $F_{\mathbf{X}}$, $H_{0b} : F_X$ is an isotropic distribution with a single mode, versus $H_{1b} :$ not H_{0b} (i.e. anisotropic). We develop a parametric bootstrap test with an assumption of the von Mises-Fisher distribution. The von Mises-Fisher distribution is an analogue of Normal distribution on the unit sphere with concentration parameter κ and directional parameter μ, denoted as $\text{vMF}(\mu, \kappa)$.

We build a test statistic that is large when F_X is neither isotropic nor having a single mode. For this purpose, we derive the following test statistic. Given $\mathbf{x}_1, \ldots, \mathbf{x}_n \in S^m$, estimate the best fitting subsphere $A(\hat{\mathbf{v}}, \hat{r})$ as done in Eq. (7.2). Let $\zeta_i = \rho_d(\mathbf{x}_i, \hat{\mathbf{v}}) = \cos^{-1}(\mathbf{x}_i^T \hat{\mathbf{v}})$ be the radial distances from the axis of the subsphere. Then the test statistic to use is the coefficient of variation of ζ,

$$Z = Z(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \frac{\bar{\zeta}}{\text{std}(\zeta)} = \frac{\frac{1}{n}\sum_{i=1}^n \zeta_i}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (\zeta_i - \bar{\zeta})^2}}.$$

The next step is to estimate a null distribution of this statistic. We have assumed that under H_{0b}, $F_\mathbf{X}$ is $\text{vMF}(\mu, \kappa)$. The unknown parameters are estimated from the data. μ is estimated by a standard likelihood approach, see Mardia and Jupp (2000). For an estimate of κ, Banerjee et al. (2005) empirically derived an approximation of the m.l.e. of κ. The estimates are

$$\hat{\mu}^{MLE} = \frac{r}{\|r\|} = \frac{\sum_{i=1}^n x_i}{\|\sum_{i=1}^n x_i\|}, \ \hat{\kappa}^{MLE} \approx \frac{\bar{r}(d+1) - \bar{r}^3}{1 - \bar{r}^2},$$

where $\bar{r} = \frac{\|r\|}{n}$. Then we generate $B \geqslant 100$ random samples of size n from $\text{vMF}(\hat{\mu}^{MLE}, \hat{\kappa}^{MLE})$ and calculate $Z_1, \ldots, Z_B$. The test rejects H_{0b} with a significance level α if

$$\frac{1}{B}\sum_{i=1}^B 1_{\{Z_i > Z\}} < \alpha.$$

7.4.3 Application procedure

As discussed in section 7.2.3, a sequence of sample PNS is obtained by iterative fitting of subspheres. In each layer of subsphere fitting, both of the tests developed in this section will be used, due to the observation in Figure 7.4. We first illustrate how these tests are applied to the examples in Figure 7.4, then propose a procedure to apply the tests to the PNS fitting procedure.

Some typical data examples on the 2-sphere and the results of the two tests are illustrated in Figure 7.4. When the true major variation is along a great circle, as in Figure 7.4a, the LRT works well and accepts H_{0a} (great sphere) but the bootstrap test rejects H_{0b}. On the other hand, when the underlying distribution is von Mises–Fisher, the LRT rejects H_{0a} in favor of H_{1a}: small sphere. However, the best fitting small sphere is frequently inappropriate, as shown in Figure 7.4b. The bootstrap test accepts H_{0b} and thus can be used to detect such a case. Therefore, in order to prevent an overfitting, we proposed to sequentially apply both tests in each level of subsphere fitting. In a case where a true variation is along a small sphere, both tests reject the null hypotheses, and we assure that the small subsphere is not overfitting.

In each level of subsphere fitting, we use the following testing procedure to test the significance of “small” subsphere fitting.

1. Test H_{0a} versus H_{1a} by the likelihood ratio test. If H_{0a} is accepted, then fit a great sphere with $r = \pi/2$ and proceed to the next layer.

2. If H_{0a} is rejected, then test the isotropy of the distribution by the parametric bootstrap test. If H_{0b} is accepted, then use great spheres for ‘all’ further subsphere fittings.

3. If both tests do not reject the null hypotheses, then use the fitted small sphere for decomposition.

Note that in step 2, when H_{0b} is accepted, we use great sphere fitting not only for the level, but also for all further levels with smaller dimensions. This is because once H_{0b} is accepted, the underlying distribution at the level is assumed to be a von Mises-Fisher. An analogy in Euclidean space is $N(0, I_k)$ where a non-linear mode of variation is meaningless. Therefore, great spheres are used for all further nested spheres, without further application of tests.

Note that for S^d, we test at most $2(d-1)$ hypotheses. This brings us a multiple testing problem, i.e. using significance level $\alpha = 0.05$ for every test may result in a larger overall type I error. This phenomenon can be treated by, for example, using Bonferroni’s correction. Deeper discussion of more advanced treatments, such as False Discovery Rate, might be an interesting open problem.

7.5 Computational Algorithm

The computation of sample PNS involves iterative applications of minimization, projection and transformation. We have given explicit formulas for the projection (Eq. 7.3) and the transformation (Eq. 7.1). The least squares problem (Eq. 7.2) is a constrained non-linear minimization problem. It can be solved by the doubly iterative algorithm described in Section 6.6 with some modifications. The algorithm is best understood in two iterative steps: The outer loop finds the point of tangency to approximate S^d by a tangent space; the inner loop solves an optimization problem in the linear space.

We make use of the exponential map and its inverse for mappings between the manifold and tangent spaces (see Helgason (2001) and Buss and Fillmore (2001)). A tangent space at $\mathbf{p} \in S^m$, $T_{\mathbf{p}}S^m$, is an affine m-dimensional vector space and can be identified by $\mathbb{R}^m$. Without loss of generality set the point of tangency $\mathbf{p} = \mathbf{e}_{m+1} = (0, \ldots, 0, 1)$, because one can use the rotation operator $\mathbf{R}(\mathbf{p})$ to transform $\mathbf{p}$ to $\mathbf{e}_{m+1}$ while preserving all data structure. The exponential map $\mathrm{Exp}_{\mathbf{p}} : T_{\mathbf{p}}S^m \longrightarrow S^m$ is defined for $\mathbf{z} \in \mathbb{R}^m \cong T_{\mathbf{p}}S^m$,

$$\mathrm{Exp}_{\mathbf{p}}(\mathbf{z}) = \left(\frac{\sin(\|\mathbf{z}\|)}{\|\mathbf{z}\|}\mathbf{z}^T, \cos(\|\mathbf{z}\|) \right)^T \in S^m.$$

The inverse exponential map (log map) $\mathrm{Log}_{\mathbf{p}} : S^m \longrightarrow T_{\mathbf{p}}S^m$ is defined for $\mathbf{x} = (x_1, \ldots, x_{m+1})^T \in S^m$,

$$\mathrm{Log}_{\mathbf{p}}(\mathbf{x}) = \frac{\theta}{\sin(\theta)}(x_1, \ldots, x_m)^T \in \mathbb{R}^m,$$

where $\cos(\theta) = x_{m+1}$. These mappings preserve the distances to the point of tangency. By using the exponential mapping and its inverse, a hypersphere with radius r in the tangent space corresponds to a subsphere in S^m with distance r. In particular, $A_{m-1}(\mathbf{v}, r)$ is equivalent to the image of $\{\mathbf{x} \in \mathbb{R}^m : \|\mathbf{x}\| = r\}$ by $\mathrm{Exp}_{\mathbf{v}}$.

The algorithm finds a suitable point of tangency $\mathbf{v}$, which is also the center of the fitted subsphere. Given a candidate $\mathbf{v}_0$, the data are mapped to the tangent space $T_{\mathbf{v}_0}S^m$ by the log map. Write $\mathbf{x}_i^\dagger = \mathrm{Log}_{\mathbf{v}_0}(\mathbf{x}_i)$, then the inner loop finds the minimizer of

$$\min_{\mathbf{v}^\dagger, r} \sum_{i=1}^{n} (\|\mathbf{x}_i^\dagger - \mathbf{v}^\dagger\| - r)^2,$$

which is a non-linear least-squares problem and can be solved numerically by e.g. the Levenberg-Marquardt algorithm (see e.g. Ch.4 of Scales (1985)). The solution $\mathbf{v}^\dagger$ is then mapped to S^m by the exponential map, and becomes the $\mathbf{v}_1$. This procedure is repeated until $\mathbf{v}$ converges.

A main advantage of this approach is the reduced difficulty of the optimization task. The inner loop solves an unconstrained problem in a vector space, which is much simpler than the original constrained problem on manifolds. Experience has shown that with a carefully chosen initial value, the algorithm has worked well in a wide range of simulated and real applications.

It becomes increasingly common in modern applied problems that the sample size is less than the dimension of the manifold, i.e. $\mathbf{x}_1, \ldots, \mathbf{x}_n \in S^d$ with $n \leqslant d$, which is frequently referred to as the high dimension, low sample size situation (Hall et al. (2005), Dryden (2005)). In Euclidean space, the dimensionality of the data can be reduced to n without losing any information. Likewise, the intrinsic dimensionality of the data on the hypersphere can be reduced to $n-1$, where additional reduction of 1 occurs because there is no 'origin' in S^d. For the simplest yet intuitive example, let $n = 2$. Then there is a geodesic joining the two points, which is the sub-manifold containing all information. A generalization of this fact can be made for any $n > 2$, by the following theorem.

Theorem 7.1. *There exists an $n-1$ dimensional nested sphere $\mathcal{A}_{n-1}$ of S^d satisfying $\mathbf{x}_i \in \mathcal{A}_{n-1}$ for all $i = 1, \ldots, n$. Moreover, there exist $\mathcal{A}_{d-1} \supset \cdots \supset \mathcal{A}_{n-1}$, all of which are great spheres (i.e. with radius 1).*

As can be seen in the proof of the theorem in the Appendix, the singular value decomposition of the data matrix $[\mathbf{x}_1 \cdots \mathbf{x}_n]$ gives the appropriate $\mathcal{A}_{n-1}$. Let $\mathcal{L}_1^n$ be the vector space of dimension n that all data points span. Then, the intersection of $\mathcal{L}_1^n$ and S^d is the $n-1$ dimensional manifold $\mathcal{A}_{n-1}$.

For a faster computation (when $n < d$), we reduce the dimensionality to $\mathcal{A}_{n-1}$ by the singular value decomposition, and use the proposed algorithm to fit $\mathcal{A}_{n-2}$, and so on.

7.6 Application to Shape space

The shape of an object is what is left after removing location, scale, and rotation. The classical approach in shape analysis (see e.g. Dryden and Mardia (1998)) is to work with (biological) landmarks of the objects. Each shape determined by a set of landmarks can be represented by a point in Kendall's (1984) shape space. A useful approach to understanding the non-Euclidean shape space is through preshape space, which is a high dimensional sphere. In this section, we discuss necessary considerations to apply PNS to shape space through the preshape space.

7.6.1 Planar shape space

Consider a set of $k > 2$ landmarks in $\mathbb{R}^2$ and the corresponding configuration matrix $\mathbf{X}$, which is a $k \times 2$ matrix of Cartesian coordinates of landmarks. The preshape of the configuration $\mathbf{X}$ is invariant under translation and scale, which is given by $\mathbf{Z} = \mathbf{HX}/\left\|\mathbf{HX}\right\|$, where $\mathbf{H}$ is the $(k-1) \times k$ Helmert sub-matrix with elements

$$h_{ij} = \begin{cases} -1/\sqrt{i(i+1)}, & 1 \leqslant i \leqslant j \\ 1/\sqrt{i(i+1)}, & i = j+1 \\ 0, & i > j+1. \end{cases}$$

Provided that $\|\mathbf{HX}\| > 0$, $\mathbf{Z} \in S^{2(k-1)-1}$. The unit sphere $S^{2(k-1)-1}$ in $\mathbb{R}^{2(k-1)}$ is the space of all possible preshapes, and is called the preshape space. Conversely, a configuration matrix corresponding to a preshape $\mathbf{Z}$ is given by $\mathbf{H}^T\mathbf{Z}$ since $\mathbf{H}^T\mathbf{H} = \mathbf{I}_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k'$, often referred to as the centering matrix.

The shape of a configuration matrix $\mathbf{X}$ can be represented by the equivalence set under rotation, $[\mathbf{Z}] = \{\mathbf{Z}\Gamma : \Gamma \in SO(2)\}$, where $SO(2)$ is the set of all 2×2 rotation matrices. The space of all possible shapes is then a non-Euclidean space called the shape space and is denoted by Σ_2^k.

We also write the preshape $\mathbf{Z}$ as a vectorized version $\mathbf{z} = \mathrm{vec}(\mathbf{Z}^T)$, where $\mathrm{vec}(\mathbf{A})$ is obtain by stacking the columns of the matrix $\mathbf{A}$ on top of one another. Then the following facts are well-known (see e.g. Kume et al. (2007) and Dryden and Mardia (1998)).

Suppose that $\mathbf{v}, \mathbf{w} \in S^{2(k-1)-1}$ are preshapes satisfying $\mathbf{v}^T\mathbf{w} \geqslant 0$ and $\mathbf{v}^T\mathbf{M}\mathbf{w} = 0$, where $\mathbf{M}$ is the $2(k-1) \times 2(k-1)$ skew-symmetric matrix consisting of $k-1$ diagonal blocks

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

Then

$$M = \begin{pmatrix} 0 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & -1 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix}.$$

Then the geodesic that joins $\mathbf{v}$ to $\mathbf{w}$, $\mathbf{Q}(\mathbf{v} \to \mathbf{w}, \theta)\mathbf{v}$, $\theta \in [0, \pi/2]$ is said to be a horizontal geodesic, and the Riemannian distance between $\mathbf{v}, \mathbf{w}$ is the same as the Riemannian distance between the corresponding shapes $[\mathbf{v}]$ and $[\mathbf{w}]$ in Σ_2^k.

For the preshapes $\mathbf{w}, \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$, as long as the shapes of those are of interest, we assume without loss of generality that $\mathbf{w}^T\mathbf{v}_i \geqslant 0$ and $\mathbf{w}^T\mathbf{M}\mathbf{v}_i = 0$.

7.6.2 Principal Nested Spheres for planar shapes

The intrinsic dimension of the shape space is $2k-4$, since the degrees of freedom are reduced from $2k$ (of the set of landmarks) by 2 for translation, 1 for scale, and 1 for rotation. This is less than the dimension of the preshape space $d = 2k-3$. It is thus desired that the $d-1$ dimensional PNS of S^d leaves no residuals. This is achieved by the theory and practical modifications in this section. In short, Procrustes fit of configurations or preshapes to a common alignment base (e.g. the Procrustes mean) results in the desired decomposition of the shape space.

Theorem 7.2. *Suppose the preshapes* $\mathbf{w}, \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n \in S^d$ *satisfy* $\mathbf{w}^T\mathbf{z}_i \geqslant 0$ *and* $\mathbf{w}^T\mathbf{M}\mathbf{z}_i = 0$ *for all* $i = 1, \ldots, n$. *Let* $\mathbf{w}^* = \mathbf{M}\mathbf{w}$ *for* M *defined above. Then* $\mathbf{w}, \mathbf{z}_i \in A_{d-1}(\mathbf{w}^*, \pi/2)$.

Moreover, define $hA_{d-1} = \{\mathbf{z} \in A_{d-1} : \mathbf{z}^T\mathbf{w} \geqslant 0\}$ *as a hyper-hemisphere. Then* $\mathbf{w}, \mathbf{z}_i \in hA_{d-1}(\mathbf{w}^*, \pi/2)$*.*

We have the following comments:

- The dimension of preshape space can be reduced by 1 without loss of any shape information.
- For the nested hemisphere hA_{d-1},
    1. the intrinsic distance $\rho^*_{d-1}(\mathbf{w}, \mathbf{z})$ defined on hA_{d-1} (see Proposition 7.7(b) in Appendix 7.8) is the same as the Riemannian distance $\rho([\mathbf{w}], [\mathbf{z}])$ in Σ_2^k for any $\mathbf{z} \in hA_{d-1}$.
    2. the tangent space of hA_{d-1} at $\mathbf{w}$ is in fact identical to the horizontal subspace of the tangent space of S^d at $\mathbf{w}$.
- The hA_{d-1} is closely related to Σ_2^k, but is not identical.

When $k = 3$, the preshape space has dimension $d = 2(k-1) - 1 = 3$. The corresponding shape space of planar triangles Σ_2^3 is $S^2(\frac{1}{2})$. hA_2 obtained from some $\mathbf{w} \in S^3$ is isometric to a unit hemisphere in $\mathbb{R}^3$. A geodesic in Σ_2^3 may or not be identified with a geodesic in hA_2. A geodesic in Σ_2^3 through $[\mathbf{w}]$ is identified with a geodesic in hA_2 through $\mathbf{w}$. On the other hand, a set of points in distance $\pi/4$ from $[\mathbf{w}]$ in Σ_2^3 is a geodesic, but is identified with a small circle with center $\mathbf{w}$ and radius $\pi/4$ in hA_2.

The choice of the alignment base $\mathbf{w}$ is an important issue because the Riemannian distance in hA_{d-1} is the same as the Riemannian distance in the shape space Σ_2^k when compared to $\mathbf{w}$, i.e.

$$\rho_d(\mathbf{w}, \mathbf{z}) = \rho_\Sigma([\mathbf{w}], [\mathbf{z}]), \text{ for } \mathbf{z} \in hA_{d-1}.$$

Moreover, $\rho_d(\mathbf{z}_1, \mathbf{z}_2)$ for $\mathbf{z}_1, \mathbf{z}_2 \in hA_{d-1}$ is closer to $\rho_\Sigma([\mathbf{z}_1], [\mathbf{z}_2])$ when $\mathbf{z}_1, \mathbf{z}_2$ are close to $\mathbf{w}$.

In general, we wish to set the alignment base $\mathbf{w}$ as a center of the data. Among many reasonable options of $\mathbf{w}$, we recommend to use the preshape of the Procrustes mean of the data. Other reasonable candidates for $\mathbf{w}$ are the geodesic mean and the PNSmean. We have tested these options to a number of real and simulation datasets. Setting $\mathbf{w}$ as the PNSmean or the geodesic mean usually takes longer computation time than using the full Procrustes mean, and the resulting decompositions are virtually the same in most cases.

In the following, we describe all candidates of $\mathbf{w}$ in more detail, giving the advantages and disadvantages of each option.

We have first considered use of the PNSmean $\mathcal{A}_0$ as the alignment base. $\mathcal{A}_0$ is identified with the origin of the coordinate system for the Euclidean representation of data $\widehat{X}_{PNS}$. Since the PNSmean is estimated from the data, we begin with the preshape of the full Procrustes mean as an initial guess for $\mathcal{A}_0$ and recursively update $\widehat{\mathcal{A}}_0$ on which preshapes are aligned. The algorithm consists of the following steps.

1. Initialize $\mathbf{w}$ as the preshape of the Procrustes mean of $\mathbf{z}_i$.

2. Align $\mathbf{z}_1, \ldots, \mathbf{z}_n$ to $\mathbf{w}$ and compute the sample PNSmean $\widehat{\mathcal{A}}_0$ of aligned $\mathbf{z}_i$.

3. If $\rho_d(\mathbf{w}, \widehat{\mathcal{A}}_0) < \epsilon$, then set $\mathbf{w} = \widehat{\mathcal{A}}_0$ and stop. Otherwise update $\mathbf{w} = \widehat{\mathcal{A}}_0$ and go to Step 2.

Note that in practice, there is no guarantee that this algorithm should converge.

Other candidates of $\mathbf{w}$ are the full Procrustes mean preshape and the geodesic mean of the preshapes. These are relevant to the Fréchet mean, where the geodesic mean is the Fréchet mean with the intrinsic (Riemannian) distance and the Procrustes mean is using the full Procrustes distance which is extrinsic to S^d. Recently, it has been observed that the curvature of the manifold S^d sometimes makes the Fréchet mean inadequate, see e.g. Huckemann et al. (2010). When the Fréchet mean is indeed a useful representation of the data, the PNSmean is usually found at a point close to the Fréchet mean. Note that even if the Fréchet mean is far from the data, the PNSmean is nevertheless located at the appropriate center of the data.

We have also considered $\mathbf{v}_{d-1}^{\dagger}$, which is relevant to the axis for the 1-dimensional subsphere $A_1(\mathbf{v}_{d-1}, r_{d-1})$ and defined in Lemma 7.8. Since $\mathbf{v}_{d-1}^{\dagger}$ is in the same distance to any points in $\mathcal{A}_1$, where the most important variation will be captured, choosing $\mathbf{v}_{d-1}^{\dagger}$ for $\mathbf{w}$ makes the criteria for the PNS adequate in the sense that residual distance in S^d is the same as the distance in the shape space. However, there is a substantial drawback that if $\{\mathbf{z}_i\}$ were aligned to $\mathbf{v}_{d-1}^{\dagger}$, then the fitted $\mathcal{A}_1$ is not a great circle in S^d even though the shapes of $\{\mathbf{z}_i\}$ are along a geodesic in Σ_2^k.

We have tried these options to a number of real and simulation datasets. Setting $\mathbf{w}$ as the PNSmean or the geodesic mean usually takes longer computation time than using the full Procrustes mean, and the resulting decompositions are virtually the same in most cases. Therefore, we recommend to use the full Procrustes mean as the alignment base.

Finally, the tests for overfitting discussed in Section 7.4 can be applied for the planar shapes case too, as the residuals are all obtained after optimal procrustes rotation.

7.6.3 Principal Nested Spheres for spaces of $m > 2$ dimensional shapes

Shapes in three or more dimensional space can also be analyzed PNS, with a similar treatment as in the planar case.

Consider a set of k landmarks in $\mathbb{R}^m$, $m \geqslant 2$, and the corresponding configuration matrix $\mathbf{X}$, which is a $k \times m$ matrix of Cartesian coordinates of landmarks. The preshape of the configuration $\mathbf{X}$ is invariant under translation and scale. The preshape $\mathbf{Z}$ of $\mathbf{X}$ is obtained by centering $\mathbf{HX}$, where $\mathbf{H}$ is the $(k-1) \times k$ Helmert sub-matrix, and then scaling $\mathbf{Z} = \mathbf{HX}/\left\|\mathbf{HX}\right\|$. All such $\mathbf{Z}$ form the unit sphere $S^{m(k-1)-1}$ embedded in $\mathbb{R}^{m(k-1)}$, which we call the preshape space.

The shape of a configuration matrix $\mathbf{X}$ can be represented by the equivalence set under rotation, $[\mathbf{Z}] = \{\mathbf{Z}\Gamma : \Gamma \in SO(m)\}$, where $SO(m)$ is the set of all $m \times m$ rotation matrices. The space of all possible shapes is then a non-Euclidean space called the shape space and is denoted by Σ_m^k. Since the dimension of the set $SO(m)$ is $\frac{m(m-1)}{2}$, the intrinsic dimensionality of Σ_m^k is

$$d = m(k-1) - 1 - \frac{m(m-1)}{2}.$$

Therefore, we wish to find the d dimensional PNS $\mathcal{A}_d$ of $S^{m(k-1)-1}$ without losing any information. Similar to the $m = 2$ case, this is achieved by the Procrustes fit of configurations or preshapes to a common alignment base $\mathbf{W}$. In the following theorem, preshapes are represented by $(k-1) \times m$ matrices. These are understood as a member of $\mathbb{R}^{m(k-1)}$ without vectorizing the matrix.

Theorem 7.3. *Suppose the preshapes $\mathbf{W}, \mathbf{Z}_1, \ldots, \mathbf{Z}_n \in S^{m(k-1)-1}$ satisfy the following:*

(a) $tr(\mathbf{W}^T\mathbf{Z}_i) \geqslant 0$,

(b) $\mathbf{W}^T\mathbf{Z}_i = \mathbf{Z}_i^T\mathbf{W}$,

(c) $\mathbf{W}$ *has rank at least* $m-1$,

(d) $k \geqslant m+1$.

Then, there exists the d *dimensional PNS* $\mathcal{A}_d$ *of* $S^{m(k-1)-1}$ *such that* $\mathbf{W}, \mathbf{Z}_i \in \mathcal{A}_d$. *Moreover, define* $h\mathcal{A}_d = \{\mathbf{Z} \in \mathcal{A}_d : tr(\mathbf{Z}^T\mathbf{W}) \geqslant 0$ *as a hyper-hemisphere. Then* $\mathbf{W}, \mathbf{Z}_i \in h\mathcal{A}_d$.

The theorem tells that when preshapes $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are aligned to a common preshape $\mathbf{W}$ (i.e. satisfying (a-b)), then the dimension of preshape space can be reduced by $m(m-1)/2$. As long as the shape is of interest, the Procrustes fit of $\mathbf{Z}_i$ to $\mathbf{W}$ results in (a-b), without losing any shape information.

The assumptions (c) and (d) are very reasonable in practice. For example, when $m = 3$ and the rank of $\mathbf{W}$ is 1, the corresponding landmarks do not form a legitimate shape, since they are aligned in a line. Similarly, $k \leqslant 3$ landmarks do not span the 3-space.

The hyper-hemisphere $h\mathcal{A}_d$ is similar to the horizontal tangent space of $S^{m(k-1)-1}$ in the sense that the distance $\rho_d(\mathbf{W}, \mathbf{Z})$ is the same as the Riemannian distance $\rho([\mathbf{W}], [\mathbf{Z}])$ in Σ_m^k for any $\mathbf{Z} \in h\mathcal{A}_d$ and the tangent space of $h\mathcal{A}_d$ at $\mathbf{W}$ is identical to the horizontal tangent space of $S^{m(k-1)-1}$ at $\mathbf{W}$.

Similar to the $m = 2$ case, for general m, we recommend to set the alignment base $\mathbf{W}$ as the Procrustes mean of the data.

7.7 Real Data Analysis

Four datasets are analyzed by the proposed method. The first two are spherical data, i.e. on S^d, and the latter two are shape data. For those shape datasets, we have used Procrustes mean as the alignment base $\mathbf{w}$.

Migration path of an elephant seal: As a simplest example, consider a dataset on the usual sphere S^2. The dataset consists of $n = 73$ daily location measurements of a migrating female elephant seal, presented in Brillinger and Stewart (1998) and also discussed in Rivest (1999). The seal migrates from the southern California coast to the eastern mid-north Pacific Ocean. Of interest is to investigate whether the seal migrates along a great circle path, i.e. the shortest distance path. Note that Brillinger and Stewart (1998) and Rivest (1999) have

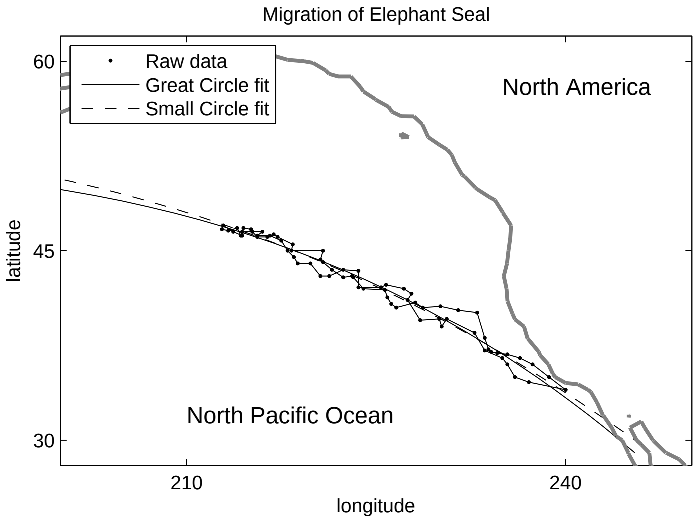


Figure 7.5: Daily observations of migration path of an elephant seal, in the latitude-longitude coordinates, and the great circle and small circle fit of the data.

analyzed this dataset in greater detail. We briefly re-analyze this data set with our hypothesis test.

Figure 7.5 shows the path of the migration, including both forward journey and return trip. Since the dataset in the latitude-longitude coordinates can be converted to points on the unit sphere, it is viewed as a set of points on S^2 and we fitted PNS, with only one nested sphere (circle in this case). We fit the best fitting great circle and small circle with fitted distance $\hat{r} = 75.45°$. The likelihood ratio test developed in Section 7.4.1 results in p-value 0.0851 (with $H_{0a} : r = 90°$, the great circle). Therefore, the migration is not significantly different at the level $\alpha = 0.05$ from a great circle path, which is consistent with the results from Brillinger and Stewart (1998) and Rivest (1999).

River and sea sand grains: We consider sand grain outlines that can be parameterized as a set of points in a hypersphere. The dataset was originally analyzed in Kent et al. (2000), and consists of outlines of sand grains in two dimensional view. There are $n_1 = 25$ river and $n_2 = 24$ sea sand grains. We illustrate an application of PNS, and use of the Euclidean-type representation to test for group mean difference.

The outline of each sand grain is represented in polar coordinates $(r_1, \ldots, r_k)$ at each equally spaced angle $(\theta_1, \ldots, \theta_k)$, with $k = 20$. The scale is removed so that $\sum_{i=1}^{k} r_i^2 = 1$. The origin for each sand grain is its center of gravity, and we keep the grains fixed in the orientation that they were recorded. With θ_i fixed throughout the samples (as $\theta_i = (i-1)2\pi/k$), $\mathbf{r} = (r_1, \ldots, r_k)$ on the unit $(k-1)$-sphere represents the shape of sand grain. Note that the size of river sand grains are typically larger than that of sea sand (see Kent et al. (2000)), but this analysis focuses on the variability in the scale invariant profiles of sand grains.

To the 49 $(= n_1 + n_2)$ data points on the 19-sphere, we have applied the procedure of sample PNS, with significance level $\alpha = 0.05$ for every test applied. The small sphere is significant for only three layers of the procedure, when fitting A_{18}, A_{17} and A_{11}, with both p-values less than 0.05. The smallest dimension nested sphere $\mathcal{A}_1$ has radius 0.8738, suggesting that the captured principal variation is not so much curved than geodesics. The PNS leads to the Euclidean-type representation X_{PNS} of the dataset, in a way that the curved principal arcs are flattened. The first three coordinates in X_{PNS} are used for visualization of major variation as in Figure 7.6.

To test the group mean difference between river and sea sand grains, we can use any Euclidean space based test procedure applied to X_{PNS}. Since we do not have any prior information on the underlying distribution, it makes sense to use a nonparametric permutation test. In particular, we use the *DiProPerm* test (*Di*rection-*Pro*jection-*Perm*utation), described in Wichers et al. (2007). The test finds a direction vector pointing from one group to the other, and computes a t-statistic of the projected values onto the direction. The null distribution of the t-statistic is found by permutation of group labels. We have used the DiProPerm test with the *Distance Weighted Discrimination* (DWD) direction (Marron et al. (2007)). The DWD is a classification tool that separates two groups with more generalizability than e.g. the popular SVM (Vapnik (1995)). The subspace found by the DWD direction and first three coordinates of X_{PNS} is illustrated as a scatterplot matrix in Figure 7.6. Although the first three coordinates of X_{PNS} do not give a visual separation between the groups, X_{PNS} turns out to be a useful Euclidean space for linear classification methods such as DWD.

DiProPerm tests the null hypothesis of equal group means. In our analysis, the test with 1000 permutations rejects the null hypothesis with p-value 0.0292. The difference of shapes in the overlay of the outlines of sand grains (Figure 7.7) is statistically significant.

Human movement: A human movement dataset, introduced in Kume et al. (2007), contains 50 samples of $k = 4$ landmarks (lower back, shoulder, wrist, and index finger). The

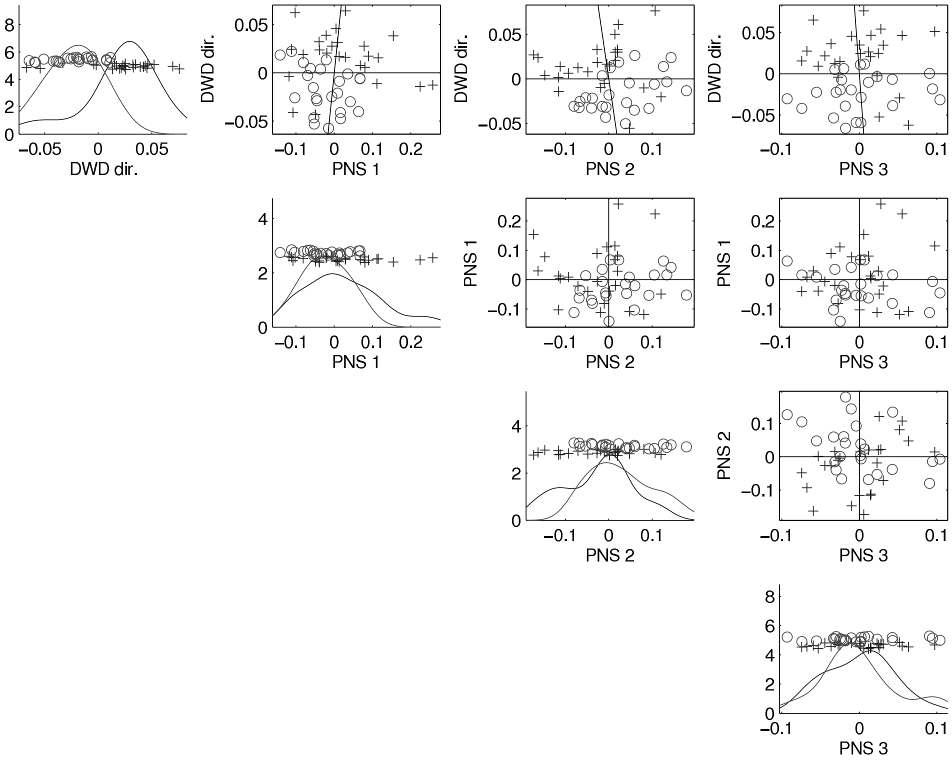


Figure 7.6: Scatterplot matrix of sand grain data set, by the DWD direction and the first three coordinates of X_{PNS}. (+: river sand grains, ∘: sea sand grains) Diagonal entries are jitter plots of one dimensional projections with kernel density estimates for each group. The DWD direction separating the two groups is found in the Euclidean space, X_{PNS}.

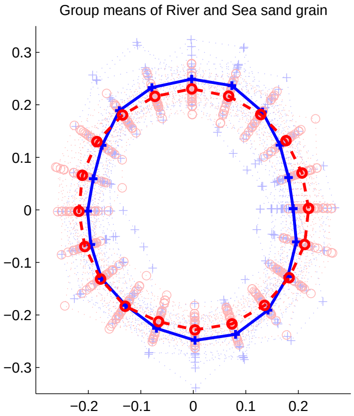


Figure 7.7: Overlaid outlines of 25 river sand grains (+) and 24 sea sand grains (∘) with the group means (thick outlines) identified with the geodesic mean of each group. The DiProPerm test rejects a null hypothesis of equal group means with p-value 0.0292

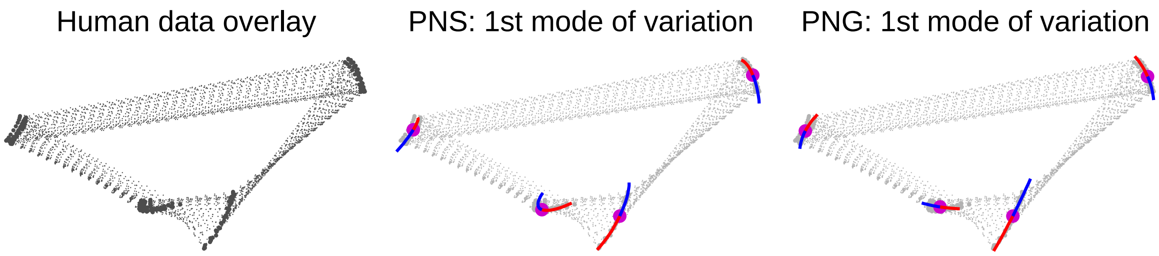


Figure 7.8: (left) Procrusted fitted human data, (center) The first principal mode of variation by PNS, showing ±2 standard deviation from PNSmean. (right) The first principal mode of variation by Principal Nested Great sphere (PNG). PNS allows to capture curving variation. In particular, the variation in the bottom left landmark is more precisely captured in PNS than PNG.

dataset consists of shape configurations in the plane of a table, of one person performing five different tasks, each observed at ten different time points. The raw data are plotted in Figure 7.8.

In the left panel of Figure 7.8, overlaid are 50 quadrilaterals, each of which is a shape configuration. Vertices of the quadrilateral are the locations of the landmarks. These 50 samples are Procrustes fitted to each other, i.e. translated, scaled, and rotated to each other so that they are as close as possible.

We have applied PNG and PNS. The fitted nested spheres of PNS have radii 1, 0.7019, 0.3967, and 0.2473 (from the 4-sphere to the 1-sphere, respectively). Note that since the dimension of the corresponding shape space is 4, the 4-d PNS is a great sphere and leaves no residuals, as expected. P-values of the sequential LRT are at most 0.0013, supporting the significance of the fitted PNS. The quadratic form of variation in the PNG coordinates is captured by the 1-d PNS, as illustrated in Figure 7.1. The principal mode of variation found by PNS is plotted in Figure 7.8, where the four bold dots together represent the shape of the PNSmean, and the curves through the PNSmean illustrate the shape change captured in the first PNS. The curvy form of variation apparent in the raw data are well captured.

Each task can be modeled as a 1-d arc, by applying PNS to the samples corresponding to each task. The results are plotted in Figure 7.9. Each task is curving through at least three geodesic components, and is well approximated by the separately fitted PNS.

Rat skull growth: The shape and size changes of rat skulls are described in Bookstein (1991) and studied by several other authors including Kenobi et al. (2010). The data are eight

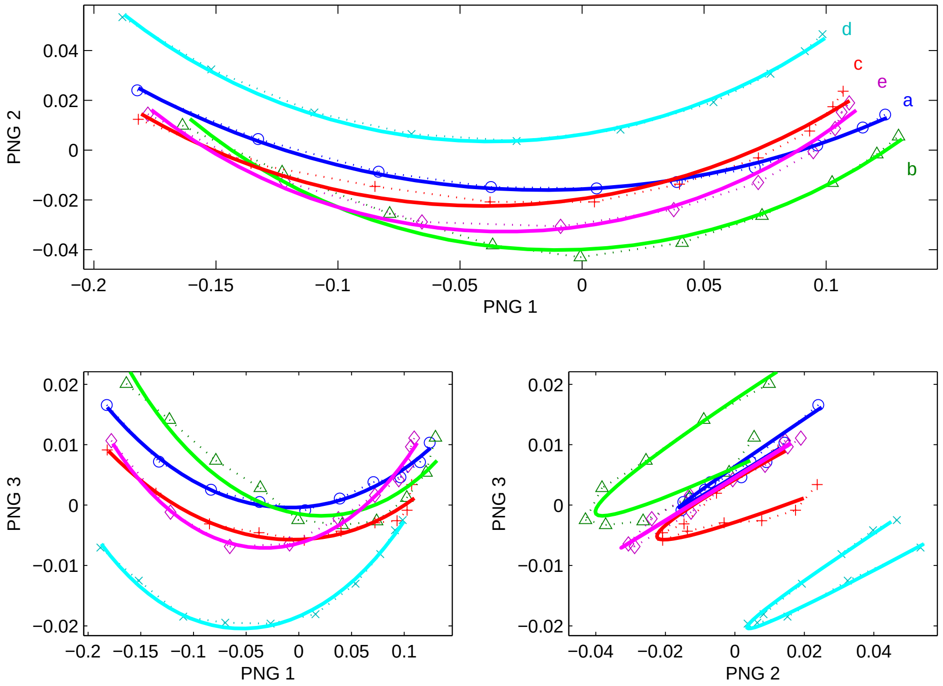


Figure 7.9: Human movement data—fitted curves (solid curves) for different tasks labeled a, b, c, d, and e, plotted in PNG coordinates.

landmark locations for skulls of 21 laboratory rats observed at eight ages (days 7, 14, 21, 30, 40, 60, 90, and 150). We discard 4 missing samples, and analyze the remaining 164 samples.

A non-geodesic variation curving through three geodesic components, in Figure 7.10(a-b), is captured in the 2-d PNS (Figure 7.10(c)). The first two principal arcs are plotted in PNG coordinates, showing the non-geodesic variation captured by PNS. The PNS coordinates capture more interesting variability in fewer components and give concise and useful representation of the data. In particular, the shape change due to the growth of the rat is well captured by the first PNS, which can be checked by inspecting the relation to the size of the rat skulls (Figure 7.10(d)). The skull size is naturally larger for older rats. The first PNS coordinates are strongly associated with the size with the sample correlation coefficient 0.9705.

The shape change due to the first PNS is illustrated in Figure 7.11. The two extreme shape configurations are also overlaid, which shows a typical effect of age.

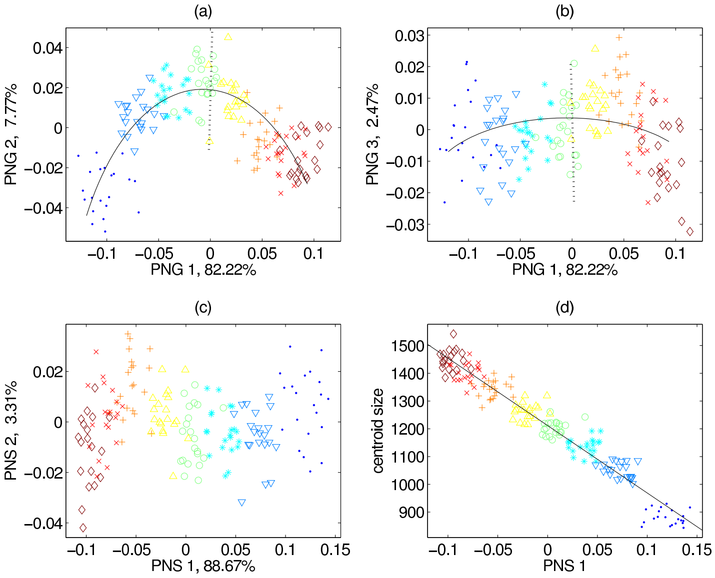


Figure 7.10: Rat skull growth: (a-b) data plotted by PNG coordinates. (—) represents the first principal arc, and (· · ·) represents the second principal arc. (c) data plotted by PNS coordinates. (d) scatterplot with centroid size and the regression line.

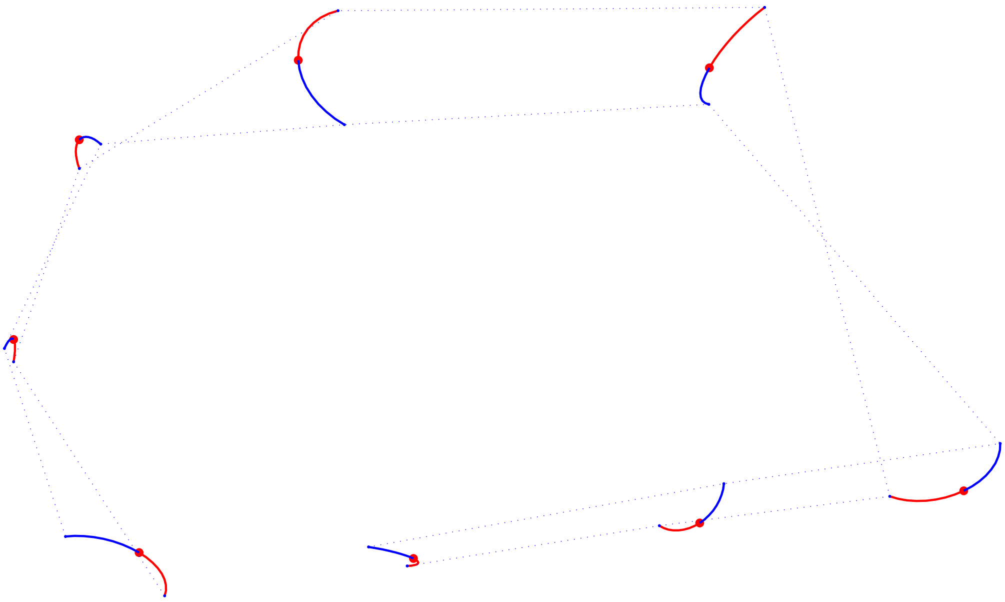


Figure 7.11: Rat skull growth: the first principal mode of variation by PNS, (· · ·) represents the shape of a typical young skull (at −2 s.d.), and (—) represents the shape of a typical old rat skull (at +2 s.d.).

7.8 Geometry of Nested Spheres

Geometric properties of nested spheres are discussed in this section. Initially it will be necessary to introduce a particular type of transformation on the sphere in order to help define PNS. Specifically, we describe a rotation matrix for moving a dataset on a sphere along a particular minimal geodesic which retains the interpoint geodesic distances after the transformation. We then describe the subsphere and the sequence of nested spheres as defined in section 7.2.1, and discuss the geometric properties of these.

7.8.1 Preliminary Transformations: Rotation matrices

Suppose that $\mathbf{a}$ and $\mathbf{b}$ are unit vectors in $\mathbb{R}^m$ and we wish to "move $\mathbf{b}$ to $\mathbf{a}$ along the geodesic path on the unit sphere in $\mathbb{R}^m$ which connects $\mathbf{b}$ to $\mathbf{a}$." Amaral et al. (2007) showed that a rotation matrix is determined in a natural way.

Define $\mathbf{c} = \{\mathbf{b} - \mathbf{a}(\mathbf{a}^T\mathbf{b})\}/\left\|\mathbf{b} - \mathbf{a}(\mathbf{a}^T\mathbf{b})\right\|$, where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^m$. Provided that $|\mathbf{a}^T\mathbf{b}| < 1$, $\mathbf{c}$ is well defined. Let $\mathbf{A} = \mathbf{a}\mathbf{c}^T - \mathbf{c}\mathbf{a}^T$. The following lemma is proved in Amaral et al. (2007).

Lemma 7.4. *Assume that* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$ *are unit vectors such that* $|\mathbf{a}^T\mathbf{b}| < 1$, *and let* $\mathbf{A}$ *and* $\mathbf{c}$ *be defined as earlier. Then for* $\theta \in (0, \pi]$, *the matrix*

$$\mathbf{Q}(\theta) = \exp(\theta\mathbf{A}) = \mathbf{I}_d + \sum_{j=1}^{\infty} \frac{\theta^j}{j!}\mathbf{A}^j$$

has the following properties:

(a) $\mathbf{Q}(\theta)$ *is an* $m \times m$ *rotation matrix,*

(b) $\mathbf{Q}(\theta)$ *can be written as*

$$\mathbf{Q}(\theta) = \mathbf{I}_d + \sin(\theta)\mathbf{A} + (\cos(\theta) - 1)(\mathbf{a}\mathbf{a}^T + \mathbf{c}\mathbf{c}^T),$$

(c) $\mathbf{Q}(\alpha)\mathbf{b} = \mathbf{a}$ *for* $\alpha = \cos^{-1}(\mathbf{a}^T\mathbf{b})$ *and*

(d) for any $\mathbf{z} \in \mathbb{R}^m$ *such that* $\mathbf{a}^T\mathbf{z} = 0$ *and* $\mathbf{b}^T\mathbf{z} = 0$, *we have* $\mathbf{Q}\mathbf{z} = \mathbf{z}$.

The path of minimum length on the surface of the unit sphere in $\mathbb{R}^m$ connecting $\mathbf{b}$ to $\mathbf{a}$ is given by $\{\mathbf{x}(\theta) = \mathbf{Q}(\theta)\mathbf{b} : \theta \in [0, \cos^{-1}(\mathbf{a}^T\mathbf{b})\}$. We write this $\mathbf{Q}(\theta)$ as $\mathbf{Q}(\mathbf{b} \to \mathbf{a}, \theta)$ and denote $\mathbf{Q}(\mathbf{b} \to \mathbf{a}) \doteq \mathbf{Q}(\mathbf{b} \to \mathbf{a}, \cos^{-1}(\mathbf{a}^T\mathbf{b}))$ for the rotation matrix that moves $\mathbf{b}$ to $\mathbf{a}$. The path defined here is indeed a minimal geodesic on the sphere. If $\mathbf{b}$ and $\mathbf{a}$ were orthogonal, then

$$\mathbf{Q}(\theta)\mathbf{b} = \cos(\theta)\mathbf{b} + \sin(\theta)\mathbf{a}, \ -\pi/2 < \theta \leqslant \pi/2, \tag{7.4}$$

which corresponds to a definition of the unit speed geodesic (Kendall et al. (1999)).

We also define $\mathbf{R}(\mathbf{v})$, for $\mathbf{v} \in \mathbb{R}^m$, as a rotation matrix that rotates $\mathbf{v}$ to the north pole $\mathbf{e}_m = (0, \ldots, 0, 1)^T$, i.e.

$$\mathbf{R}(\mathbf{v}) = \mathbf{Q}(\mathbf{v} \to \mathbf{e_m}).$$

Note that the last row of $\mathbf{R}(\mathbf{v})$ is $\mathbf{v}^T$. If $\mathbf{v} = \mathbf{e}_m$, then $\mathbf{R}(\mathbf{v}) = \mathbf{I}_m$.

Lemma 7.5. *Assume that* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$ *are unit vectors such that* $|\mathbf{a}^T\mathbf{b}| < 1$*, and let* $\theta \in (0, 2\pi]$*.*

(a) Let $\mathbf{R}$ *be an* $m \times m$ *rotation matrix. Then,* $\mathbf{Q}(\mathbf{R}\mathbf{b} \to \mathbf{R}\mathbf{a}, \theta) = \mathbf{R}\mathbf{Q}(\mathbf{b} \to \mathbf{a}, \theta)\mathbf{R}^T$*. Equivalently,* $\mathbf{Q}(\mathbf{R}^T\mathbf{b} \to \mathbf{R}^T\mathbf{a}, \theta) = \mathbf{R}^T\mathbf{Q}(\mathbf{b} \to \mathbf{a}, \theta)\mathbf{R}$*.*

(b) Let $\mathbf{a}' = (\mathbf{a}^T, 0)^T$*,* $\mathbf{b}' = (\mathbf{b}^T, 0)^T$*. Then the* $(m+1) \times (m+1)$ *rotation matrix that moves* $\mathbf{b}'$ *to* $\mathbf{a}'$ *is given by*

$$\mathbf{Q}(\mathbf{b}' \to \mathbf{a}', \theta) = \begin{bmatrix} \mathbf{Q}(\mathbf{b} \to \mathbf{a}, \theta) & \mathbf{0}_{m\times 1} \\ \mathbf{0}_{1\times m} & 1 \end{bmatrix},$$

where $\mathbf{0}_{m\times n}$ *is the* $m \times n$ *matrix of zeros.*

7.8.2 Geometry of Subsphere

The nested spheres of S^d are lower dimensional submanifolds of S^d, each of which is isomorphic to the unit spheres in different dimensions. We first define a subsphere of S^m, $m \geqslant 2$ and discuss the relevant geometry, which induces the nested spheres.

Definition (7.1)**.** A *subsphere* A_{m-1} of S^m is defined by an orthogonal axis $\mathbf{v} \in S^m$ and a distance $r \in (0, \pi/2]$, as follows:

$$A_{m-1}(\mathbf{v}, r) = \{\mathbf{x} \in S^m : \rho_m(\mathbf{v}, \mathbf{x}) = r\},$$

where $\rho_m(\cdot, \cdot)$ is the great circle distance function on S^m.

Note that $A_{m-1}(\mathbf{v}, r)$ is the boundary of the geodesic ball in S^m with center $\mathbf{v}$ and radius r. The $\mathbf{v}$ is said to be orthogonal to A_{m-1} in a sense of the following lemma.

Lemma 7.6. *(a) For any* $\mathbf{x}, \mathbf{y} \in A_{m-1}$, $(\mathbf{x} - \mathbf{y})^T\mathbf{v} = 0$.

(b) $\mathbf{x} \in A_{m-1}$ *if and only if* $\mathbf{v}^T(\mathbf{x} - \cos(r)\mathbf{v}) = 0$ *and* $\|\mathbf{x}\| = 1$.

A subsphere A_{m-1} of S^m is essentially an $(m-1)$ dimensional sphere. The following properties of subspheres give the mathematical background to treat A_{m-1} as S^{m-1}.

Proposition 7.7. *Let* $A_{m-1}(\mathbf{v}, r)$ *be a subsphere in* S^m. *Then*

(a) (A_{m-1}, ρ_m) *is isomorphic to* (S^{m-1}, ρ_{m-1}) *with an isomorphism* $f : A_{m-1} \longrightarrow S^{m-1}$ *defined by*

$$f(\mathbf{x}) = \frac{1}{\sin(r)}\mathbf{R}^-(\mathbf{v})\mathbf{x}, \ \mathbf{x} \in A_{m-1}$$

with inverse

$$f^{-1}(\mathbf{x}^\dagger) = \mathbf{R}^T(\mathbf{v}) \begin{bmatrix} \sin(r) \cdot \mathbf{x}^\dagger \\ \cos(r) \end{bmatrix}, \ \mathbf{x}^\dagger \in S^{m-1},$$

where $\mathbf{R}(\mathbf{v})$ *is the* $(m+1) \times (m+1)$ *rotation matrix that moves* $\mathbf{v}$ *to the north pole,* $\mathbf{R}^-(\mathbf{v})$ *is the* $m \times (m+1)$ *matrix consisting of the first* m *rows of* $\mathbf{R}(\mathbf{v})$.

(b) Let $\rho^*_{m-1}(\mathbf{x}, \mathbf{y}) = \sin(r)\rho_{m-1}(f(\mathbf{x}), f(\mathbf{y}))$. *Then* ρ^*_{m-1} *is a metric on* A_{m-1}.

(c) (A_{m-1}, ρ^*_{m-1}) *is isometric to* $(S^{m-1}, \sin(r)\rho_{m-1})$.

(d) The two metrics ρ_m *and* ρ^*_{m-1} *are equivalent, in a sense that the following inequalities*

$$\rho_m(\mathbf{x}, \mathbf{y}) \leqslant \rho^*_{m-1}(\mathbf{x}, \mathbf{y}) \leqslant \frac{\pi \sin(r)}{2r}\rho_m(\mathbf{x}, \mathbf{y})$$

hold for all $\mathbf{x}, \mathbf{y} \in A_{m-1}$ *and both equalities hold if and only if* $r = \pi/2$ *or* $\mathbf{x} = \mathbf{y}$.

(e) $\rho^*_{m-1}(\mathbf{x},\mathbf{y}) - \rho_m(\mathbf{x},\mathbf{y}) \leqslant \pi\sin(r) - 2r$ *for all* $\mathbf{x},\mathbf{y} \in A_{m-1}$.

The $\rho^*_{m-1}(\mathbf{x},\mathbf{y})$ can be interpreted as the length of a minimal arc in A_{m-1} that joins $\mathbf{x},\mathbf{y}$. Precisely, the minimal arc is the image by f^{-1} of the minimal geodesic segment joining $f(\mathbf{x})$ and $f(\mathbf{y})$. Let $\mathbf{x}^\dagger = f(\mathbf{x})$, $\mathbf{y}^\dagger = f(\mathbf{y})$. Then the geodesic segment is given by

$$\Gamma = \{\gamma(\theta) = \mathbf{Q}(\mathbf{x}^\dagger \to \mathbf{y}^\dagger, \theta)\mathbf{x}^\dagger : \theta \in [0, \cos^{-1}(\mathbf{x}^{\dagger T}\mathbf{y}^\dagger)]\}.$$

By Lemma 7.5, we have for any $\theta \in [0, \cos^{-1}(\mathbf{x}^{\dagger T}\mathbf{y}^\dagger)]$,

$$\begin{aligned} f^{-1}(\gamma(\theta)) &= \mathbf{R}(\mathbf{v})^T \begin{bmatrix} \sin(r)\mathbf{Q}(\mathbf{x}^\dagger \to \mathbf{y}^\dagger, \theta)\mathbf{x}^\dagger \\ \cos(r) \end{bmatrix} \\ &= \mathbf{R}(\mathbf{v})^T \begin{bmatrix} \mathbf{Q}(\mathbf{x}^\dagger \to \mathbf{y}^\dagger, \theta) & \mathbf{0}_{1\times m} \\ \mathbf{0}_{m\times 1} & 1 \end{bmatrix} \mathbf{R}(\mathbf{v})\mathbf{R}(\mathbf{v})^T \begin{bmatrix} \sin(r)\mathbf{x}^\dagger \\ \cos(r) \end{bmatrix} \\ &= \mathbf{Q}(\mathbf{x}_p \to \mathbf{y}_p, \theta)\mathbf{x}, \end{aligned}$$

where

$$\mathbf{x}_p = \mathbf{R}(\mathbf{v})^T \begin{bmatrix} \mathbf{x}^\dagger \\ 0 \end{bmatrix} = \frac{\mathbf{x} - \cos(r)\mathbf{v}}{\sin(r)},$$

and $\mathbf{y}_p$ is defined similarly. One can check that $\rho_m(\mathbf{x}_p, \mathbf{y}_p) = \rho_{m-1}(\mathbf{x}^\dagger, \mathbf{y}^\dagger)$ and $\mathbf{Q}(\mathbf{x}_p \to \mathbf{y}_p)\mathbf{x} = \mathbf{y}$. Thus the arc $\{\mathbf{Q}(\mathbf{x}_p \to \mathbf{y}_p, \theta)\mathbf{x} : \theta \in [0, \cos^{-1}(\mathbf{x}_p^T\mathbf{y}_p)]\}$ joins $\mathbf{x}$ to $\mathbf{y}$ and is minimal in A_{m-1} because it is isomorphic to the minimal geodesic Γ. Note that Lemma 7.4(d) leads to $\mathbf{Q}(\mathbf{x}_p \to \mathbf{y}_p, \theta)\mathbf{v} = \mathbf{v}$ for all $x, y \in A_{m-1}$, θ.

The difference between ρ_m and ρ^*_{m-1} is due to the fact that the minimal arc for ρ^*_{m-1} is not a geodesic in S^m. If $r < \pi/2$, then the geodesic segment joining $\mathbf{x},\mathbf{y}$ is always shorter than the minimal arc in A_{m-1}. Since the difference is relatively small for close points (by Proposition 7.7(d-e)), this difference does not obscure much the underlying structure of the points in S^m.

7.8.3 Geometry of Nested Spheres

We now define a sequence of *nested spheres* $\{\mathcal{A}_{d-1}, \mathcal{A}_{d-2}, \ldots, \mathcal{A}_1\}$ of S^d, $d \geqslant 2$, with decreasing intrinsic dimensions. We first introduce a sequence of subspheres $A_{d-1}, A_{d-2}, \ldots, A_1$ of S^d, which are in different spaces. The $d-1$ dimensional subsphere A_{d-1} of S^d, defined in Definition 7.1, is in $S^d \in \mathbb{R}^{d+1}$. The second subsphere A_{d-2} is defined from the isomorphic space S^{d-1} of A_{d-1}. Similarly, the lower dimensional subspheres are defined recursively.

Definition 7.4. A sequence $\{A_{d-1}, A_{d-2}, \ldots, A_1\}$ of subspheres is defined recursively as follows:

(i) A_{d-1} is defined as the subsphere with $\mathbf{v}_1 \in \mathbb{R}^{d+1}$, $r \in (0, \pi/2]$ by Definition 7.1.

(ii) For each $k = 2, \ldots, d-1$, A_{d-k} is the subsphere defined with $\mathbf{v}_k \in \mathbb{R}^{d-k+2}$, $r_k \in (0, \pi/2]$ from S^{d-k+1}, which is isomorphic to A_{d-k+1}.

We also define transformations $f_k : A_{d-k} \longrightarrow S^{d-k}$ with $(\mathbf{v}_k, r_k)$, for $k = 1, \ldots, d-1$, as done in Proposition 7.7, i.e.

$$\begin{aligned} f_k(\mathbf{x}) &= \frac{1}{\sin(r_k)}\mathbf{R}^{-}(\mathbf{v_k})\mathbf{x}, \ \mathbf{x} \in A_{d-k}, \\ f_k^{-1}(\mathbf{x}^\dagger) &= \mathbf{R}^T(\mathbf{v_k}) \begin{bmatrix} \sin(r_k)\cdot \mathbf{x}^\dagger \\ \cos(r_k) \end{bmatrix}, \ \mathbf{x}^\dagger \in S^{d-k}. \end{aligned}$$

Definition (7.2)**.** A $d-k$ *dimensional nested sphere* $\mathcal{A}_{d-k}$ of S^d is defined as

$$\mathcal{A}_{d-k} = \begin{cases} f_1^{-1} \circ \cdots \circ f_{k-1}^{-1}(A_{d-k}) & \text{if } k = 2, \ldots, d-1 \\ A_{d-1} & \text{if } k = 1 \end{cases}$$

The geometric interpretation and hierarchical structure of the nested spheres are illustrated in Figure 7.2 and 7.12. The nested sphere $\mathcal{A}_{d-k}$ can be understood as a shifted $(d-k)$-sphere, which is orthogonal to k orthogonal directions in the sense of Lemma 7.8. The following properties summarize some geometric facts of the nested spheres. $\mathbf{x}_{p,k}$ in the lemma can be understood as the projection of $\mathbf{x}$ onto the subspace that is orthogonal to $\mathbf{v}_1^*, \ldots, \mathbf{v}_{d-k}^*$.

Lemma 7.8. *Let $\mathcal{A}_{d-1}, \ldots, \mathcal{A}_1$ be nested spheres of S^d from a sequence of subspheres $A_{d-k}(\mathbf{v}_k, r_k)$. Then, there exists an orthogonal basis $\mathbf{v}_1^*, \ldots, \mathbf{v}_{d-1}^* \in \mathbb{R}^{d+1}$ such that for each $k = 1, \ldots, d-1$,*

$$\begin{tikzcd}
\text{Isomorphic space of NS} & \text{Nested Subspheres} & \text{Sqeuence of Nested Spheres} \\
S^3 \arrow[ddr] & & S^3 \\
& & \cup \\
S^2 \arrow[r, phantom, "\cong"] \arrow[ddr] & A_2(\mathbf{v}_1, r_1) \arrow[r, phantom, "\cong"] & \mathfrak{A}_2 \\
& & \cup \\
S^1 \arrow[r, phantom, "\cong"] & A_1(\mathbf{v}_1, r_1) \arrow[r, phantom, "\cong"] & \mathfrak{A}_1
\end{tikzcd}$$

Figure 7.12: Hierarchical structure of the sequence of nested spheres of the 3-sphere.

(a) $(\mathbf{x}-\mathbf{y})^T\mathbf{v}_i^* = 0$ *for all* $i = 1, \ldots, k$, $\mathbf{x}, \mathbf{y} \in \mathcal{A}_{d-k}$,

(b) $\mathbf{x} \in \mathcal{A}_{d-k}$ *if and only if* $\mathbf{x}_{p,k}^T\mathbf{v}_j^* = 0$, *for all* $j = 1, \ldots, k$, *and* $\|\mathbf{x}_{p,k}\| = \prod_{i=1}^k \sin(r_i)$ *where*

$$\mathbf{x}_{p,k} = \mathbf{x} - \cos(r_1)\mathbf{v}_1^* - \sin(r_1)\cos(r_2)\mathbf{v}_2^* - \cdots - \prod_{i=1}^{k-1}\sin(r_i)\cos(r_k)\mathbf{v}_k^*.$$

Moreover, an explicit expression for $\mathbf{v}_j^*$ *can be obtained from* $\mathbf{v}_1, \ldots, \mathbf{v}_j$ *as*

$$\mathbf{v}_j^\dagger = f_1^{-1} \circ \cdots \circ f_{j-1}^{-1}(\mathbf{v}_j) \in \mathcal{A}_{d-j+1}, \tag{7.5}$$

$$\mathbf{v}_j^* = \prod_{i=1}^{j-1}\sin^{-1}(r_i)\{\mathbf{v}_j^\dagger - \cos(r_1)\mathbf{v}_1^* - \sin(r_1)\cos(r_2)\mathbf{v}_2^* - \cdots - \prod_{i=1}^{j-2}\sin(r_i)\cos(r_{j-1})\mathbf{v}_{j-1}^*\} \tag{7.6}$$

A direct consequence of this lemma is that for any nested sphere $\mathcal{A}_{d-k}$ of S^d can be understood as the intersection of a hyperplane $\mathcal{H}_k$ and S^d. The hyperplane $\mathcal{H}_k$ is a $d-k$ dimensional affine subspace that is orthogonal to $\mathbf{v}_1^*, \ldots, \mathbf{v}_{d-k}^*$.

Proposition 7.9. *Let* $\mathcal{A}_{d-1}, \ldots, \mathcal{A}_1$ *be nested spheres of* S^d *from subspheres* $A_{d-k}(\mathbf{v}_k, r_k)$. *Then,*

(a) $\mathcal{A}_1 \subsetneq \mathcal{A}_2 \subsetneq \cdots \subsetneq \mathcal{A}_{d-1} \subsetneq S^d$, *where* $A \subsetneq B$ *means that* A *is a proper subset of* B,

(b) Let $\rho_{d-k}^*(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^k \sin(r_i)\rho_{d-k}(\mathbf{x}', \mathbf{y}')$, *where* $\mathbf{x}' = f_k \circ \cdots \circ f_1(\mathbf{x})$. *Then* ρ_{d-k}^* *is a metric on* $\mathcal{A}_{d-k}$.

(c) $(\mathcal{A}_{d-k}, \rho_{d-k}^*)$ *is isometric to* $(S^{d-k}, \prod_{i=1}^k \sin(r_i)\rho_{d-k})$.

The ρ_d and ρ_{d-k}^* are indeed equivalent metrics. Moreover, one can show that $\rho_{d-k}^*(\mathbf{x}, \mathbf{y})$ is the length of a minimal arc in $\mathcal{A}_{d-k}$ that joins $\mathbf{x}$ and $\mathbf{y}$.

7.9 Proofs and Additional Lemmas

Proofs for Appendix 7.8 will be given first. We then return to give proofs for the main results.

Proof of Lemma 7.5. (a) Let $\mathbf{a}_0 = \mathbf{Ra}$, $\mathbf{b}_0 = \mathbf{Rb}$ and $\mathbf{c}_0 = \{\mathbf{b}_0 - \mathbf{a}_0(\mathbf{a}_0^T\mathbf{b}_0)\}/\left\|\mathbf{b}_0 - \mathbf{a}_0(\mathbf{a}_0^T\mathbf{b}_0)\right\|$. Then $\mathbf{c}_0 = \mathbf{Rc}$, where $\mathbf{c} = \{\mathbf{b} - \mathbf{a}(\mathbf{a}^T\mathbf{b})\}/\left\|\mathbf{b} - \mathbf{a}(\mathbf{a}^T\mathbf{b})\right\|$, since $\mathbf{RR}^T = \mathbf{R}^T\mathbf{R} = \mathbf{I}_m$. Then,

$$\mathbf{Q}(\mathbf{Rb} \to \mathbf{Ra}, \theta) = \mathbf{R}\{\mathbf{I}_d + \sin(\theta)(\mathbf{ac}^T - \mathbf{ca}^T) + (\cos(\theta) - 1)(\mathbf{aa}^T + \mathbf{cc}^T)\}\mathbf{R}^T = \mathbf{RQ}(\mathbf{b} \to \mathbf{a}, \theta)\mathbf{R}^T.$$

(b) Let $\mathbf{c}'$ be defined similarly for $\mathbf{a}'$, $\mathbf{b}'$. We have

$$\mathbf{a}'\mathbf{a}'^T = \begin{bmatrix} \mathbf{a} \\ 0 \end{bmatrix} \begin{bmatrix} \mathbf{a}^T & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{aa}^T & \mathbf{0}_{m\times 1} \\ \mathbf{0}_{1\times m} & 0 \end{bmatrix},$$

and $\mathbf{c}'\mathbf{c}'^T$,$\mathbf{a}'\mathbf{c}'^T$ and $\mathbf{c}'\mathbf{a}'^T$ can be expressed in a similar fashion. Then the expression of $\mathbf{Q}$ in Lemma 1(b) gives the desired result. □

Proof of Lemma 7.6. For $\mathbf{x} \in \mathbb{R}^m$ such that $\|\mathbf{x}\| = 1$, $\mathbf{x} \in A_{m-1}$ if and only if $\rho_m(\mathbf{v}, \mathbf{x}) = \cos^{-1}(\mathbf{v}^T\mathbf{x}) = r$. This is equivalent to $\mathbf{v}^T\mathbf{x} - \cos(r)\mathbf{v}^T\mathbf{v} = 0$ since $\mathbf{v}^T\mathbf{v} = 1$. This proves (b). Write $\mathbf{x} - \mathbf{y} = (\mathbf{x} - \cos(r)\mathbf{v}) - (\mathbf{y} - \cos(r)\mathbf{v})$, then the result (a) follows from (b). □

Proof of Proposition 7.7. We first show that f is a well-defined bijective function. Proofs for (b-e) will follow. (a) is then given by (c) and (d).

First note that since $\sin(r) > 0$, f is well defined. For any $\mathbf{x} \in A_{m-1}$, let $\mathbf{x}^\dagger = f(\mathbf{x})$. Then $\mathbf{x}^\dagger \in \mathbb{R}^m$, and since $\mathbf{R}^-(\mathbf{v})^T\mathbf{R}^-(\mathbf{v}) = \mathbf{I}_{m+1} - \mathbf{vv}^T$, we get

$$\left\|\mathbf{x}^\dagger\right\|^2 = \frac{1}{\sin^2(r)}\left\|\mathbf{R}^-(\mathbf{v})\mathbf{x}\right\|^2 = \frac{1}{\sin^2(r)}\{\mathbf{x}^T\mathbf{x} - (\mathbf{x}^T\mathbf{v})^2\} = \frac{1}{\sin^2(r)}\{1 - \cos^2(r)\} = 1$$

Thus, $\mathbf{x}^\dagger \in S^{m-1}$. Conversely, for any $\mathbf{x}^\dagger \in S^{m-1}$, let $\mathbf{x} = f^{-1}(\mathbf{x}^\dagger)$. Then $\|\mathbf{x}\| = 1$ and

$$\mathbf{v}^T\mathbf{x} = (\mathbf{R}(\mathbf{v})\mathbf{v})^T \begin{bmatrix} \sin(r)\mathbf{x}^\dagger \\ \cos(r) \end{bmatrix} = \cos(r).$$

By Lemma 7.6(b), $\mathbf{x} \in A_{m-1}$. One can easily show that $f \circ f^{-1}(\mathbf{x}^\dagger) = \mathbf{x}^\dagger$, $f^{-1} \circ f(\mathbf{x}) = \mathbf{x}$.

Therefore, f is a well defined bijective function.

Since ρ_{m-1} is a metric and $\sin(r) > 0$, the metric ρ^*_{m-1} is nonnegative and symmetric, and the triangle inequality holds. In addition, since f is bijective, we have $\rho^*_{m-1}(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$. This proves (b). With the metric ρ^*_{m-1}, f is an isometry and (c) follows.

To prove (d) and (e), the difference between two metrics for a fixed $r \in (0, \pi/2]$ is given by

$$\rho^*_{m-1}(\mathbf{x}, \mathbf{y}) - \rho_m(\mathbf{x}, \mathbf{y}) = \sin(r) \cos^{-1}\left(\frac{\cos(\rho_m(\mathbf{x}, \mathbf{y})) - \cos^2(r)}{\sin^2(r)}\right) - \rho_m(\mathbf{x}, \mathbf{y}) := h_r\{\rho_m(\mathbf{x}, \mathbf{y})\},$$

for any $\mathbf{x}, \mathbf{y} \in A_{m-1}$. Note that $\max_{\mathbf{x},\mathbf{y}} \rho_m(\mathbf{x}, \mathbf{y}) = 2r$. Then h_r is a strictly increasing function on $[0, 2r]$ with minimum $h_r(0) = 0$ and the maximum $h_r(2r) = \pi \sin(r) - 2r$. This proves (e) and leads to the first inequality of (d). The second inequality is obtained from observing that

$$\frac{\pi \sin(r)}{2r} \rho_m(\mathbf{x}, \mathbf{y}) - \rho^*_{m-1}(\mathbf{x}, \mathbf{y})$$

is nonnegative and is zero if and only if $\rho_m(\mathbf{x}, \mathbf{y}) = 0$ or $2r$. □

The following lemmas are useful to prove Lemma 7.8 and also could be of independent interest.

Lemma 7.10. *Let* $\mathbf{v}^\dagger_j$ *and* $\mathbf{v}^*_j$ *be as defined in (7.5-7.6). For any* $\mathbf{x} \in S^d$ *and* $k = 1, \dots, d-1$, *the following are equivalent:*

(i) $\mathbf{x} \in \mathcal{A}_{d-k}$.

(ii) $\mathbf{v}_k^T [f_k \circ \cdots \circ f_1(\mathbf{x})] = \cos(r_k)$.

(iii) For all $j = 1, \dots, k$,

$$\mathbf{x}^T \mathbf{v}^\dagger_j = \prod_{i=1}^{j-1} \sin^2(r_i) \cos(r_j) + \prod_{i=1}^{j-2} \sin^2(r_i) \cos^2(r_{j-1}) + \cdots + \cos^2(r_1).$$

(iv) For all $j = 1, \dots, k$,

$$\mathbf{x}^T \mathbf{v}^*_j = \prod_{i=1}^{j-1} \sin(r_i) \cos(r_j).$$

Proof of Lemma 7.10. [*(i)* $\Leftrightarrow$ *(ii)*] By Definition 7.2 and since each f_i is bijective, $\mathbf{x} \in \mathcal{A}_{d-k}$ is equivalent to $f_k \circ \cdots \circ f_1(\mathbf{x}) \in A_{d-k}$. By Lemma 7.6(b), this is also equivalent to (ii).

[*(i)* $\Leftrightarrow$ *(iii)*] First note that for any $k = 1, \dots, d-1$, for $\mathbf{y} \in S^{d-k}$,

$$\begin{aligned} f_1^{-1} \circ \cdots \circ f_k^{-1}(\mathbf{y}) &= \mathbf{R}^T(\mathbf{v}_1) \begin{bmatrix} \sin(r_1)\{f_2^{-1} \circ \cdots \circ f_k^{-1}(\mathbf{y})\} \\ \cos(r_1) \end{bmatrix} \\ &= [\mathbf{R}(\mathbf{v}_1, \dots, \mathbf{v}_k)]^T \begin{bmatrix} \prod_{i=1}^{k} \sin(r_i)\mathbf{y} \\ \prod_{i=1}^{k-1} \sin(r_i)\cos(r_{k-1}) \\ \vdots \\ \cos(r_1) \end{bmatrix}, \end{aligned}$$

where $\mathbf{R}(\mathbf{v}_1, \dots, \mathbf{v}_k)$ is a rotation matrix defined as

$$\mathbf{R}(\mathbf{v}_1, \dots, \mathbf{v}_k)^T = \mathbf{R}^T(\mathbf{v}_1) \begin{bmatrix} \mathbf{R}^T(\mathbf{v}_2) & \mathbf{0}_{d\times 1} \\ \mathbf{0}_{1\times d} & 1 \end{bmatrix} \cdots \begin{bmatrix} \mathbf{R}^T(\mathbf{v}_k) & \mathbf{0}_{(d+2-k)\times 1} \\ \mathbf{0}_{1\times(d+2-k)} & 1 \end{bmatrix}.$$

Then

$$\begin{aligned} \mathbf{x}^T\mathbf{v}_j^\dagger &= \left(f_1^{-1} \circ \cdots \circ f_{j-1}^{-1}\{f_{j-1} \circ \cdots \circ f_1(\mathbf{x})\}\right)^T f_1^{-1} \circ \cdots \circ f_{j-1}^{-1}(\mathbf{v}_j) \\ &= \prod_{i=1}^{j-1} \sin^2(r_i)\{f_{j-1} \circ \cdots \circ f_1(\mathbf{x})\}^T\mathbf{v}_j + \prod_{i=1}^{j-2} \sin^2(r_i)\cos^2(r_{j-1}) + \cdots + \cos^2(r_1) \end{aligned}$$

and the result follows from (ii).

[*(i)* $\Rightarrow$ *(iv)*] Since $\mathbf{x} \in \mathcal{A}_{d-k}$, we have $\mathbf{x}^T\mathbf{v}_1^* = \cos(r_1)$ by definition. Suppose $\mathbf{x}^T\mathbf{v}_\jmath^* = \prod_{i=1}^{\jmath-1} \sin(r_i)\cos(r_\jmath)$ for all $\jmath = 1, \dots, j-1$, then (iii) and canceling terms give

$$\begin{aligned} \mathbf{x}^T\mathbf{v}_j^* &= \mathbf{x}^T \left(\mathbf{v}_j^\dagger - \cos(r_1)\mathbf{v}_1^* - \cdots - \prod_{i=1}^{j-2} \sin(r_1)\cos(r_{j-1})\mathbf{v}_{j-1}^*\right) \prod_{i=1}^{j-1} \sin^{-1}(r_i) \\ &= \prod_{i=1}^{j-1} \sin(r_1)\cos(r_j). \end{aligned}$$

Thus by induction, (iv) holds.

[*(iv)* $\Rightarrow$ *(iii)*] Suppose (iv) holds, then for $j = 1, \ldots, k$,

$$\begin{aligned}
&\mathbf{x}^T \mathbf{v}_j^* - \prod_{i=1}^{j-1} \sin(r_i)\cos(r_j) \\
&= \prod_{i=1}^{j-1} \sin^{-1}(r_i) \left(\mathbf{x}^T \mathbf{v}_j^\dagger - \cos^2(r_1) - \cdots - \prod_{i=1}^{j-2} \sin^2(r_i)\cos^2(r_{j-1}) - \prod_{i=1}^{j-1} \sin^2(r_i)\cos(r_j) \right),
\end{aligned}$$

which equals to zero if and only if (iii) holds. □

Proof of Lemma 7.8. We first show that $\{\mathbf{v}_i^*; i = 1, \ldots, d-1\}$ is an orthonormal basis. Note that $\mathbf{v}_1^* = \mathbf{v}_1$, and $\mathbf{v}_2^* = \sin^{-1}(r_1)\{\mathbf{v}_2^\dagger - \cos(r_1)\mathbf{v}_1^*\}$. Since $\mathbf{v}_2^\dagger \in \mathcal{A}_{d-1}$, by Lemma 7.10, we have

$$\mathbf{v}_2^{*T}\mathbf{v}_1 = \sin^{-1}(r_1)\{\mathbf{v}_2^{\dagger T}\mathbf{v}_1 - \cos(r_1)\} = 0,$$

and

$$\mathbf{v}_2^{*T}\mathbf{v}_2^* = \sin^{-1}(r_1)\mathbf{v}_2^{*T}\mathbf{v}_2^\dagger = \sin^{-1}(r_1)\{\mathbf{v}_2^{\dagger T}\mathbf{v}_2^\dagger - \cos(r_1)\mathbf{v}_2^\dagger\mathbf{v}_1^*\} = 1.$$

Suppose $\mathbf{v}_i^{*T}\mathbf{v}_j^* = 0$ and $\|\mathbf{v}_i^*\| = \|\mathbf{v}_i^*\| = 1$ for $1 \leqslant i < j \leqslant k-1$. Since $\mathbf{v}_k^\dagger \in \mathcal{A}_{d-k+1}$, by Lemma 7.10, we have

$$\mathbf{v}_j^{*T}\mathbf{v}_k^* = \prod_{i=1}^{k-1} \sin^{-1}(r_i)\mathbf{v}_j^{*T}\{\mathbf{v}_k^\dagger - \prod_{i=1}^{j-1} \sin(r_i)\cos(r_j)v_j^*\} = 0,$$

and

$$\begin{aligned}
\|\mathbf{v}_k^*\| &= \prod_{i=1}^{k-1} \sin^{-1}(r_i)\mathbf{v}_k^{*T}\mathbf{v}_k^\dagger \\
&= \prod_{i=1}^{k-1} \sin^{-2}(r_i)\mathbf{v}_k^{\dagger T}\{\mathbf{v}_k^\dagger - \cos(r_1)\mathbf{v}_1^* - \cdots - \prod_{i=1}^{k-2} \sin(r_i)\cos(r_{k-1})\mathbf{v}_{k-1}^*\} = 1.
\end{aligned}$$

Thus, by induction, $\mathbf{v}_i^*, i = 1, \ldots, d-1$, are orthonormal.

Now for (b), suppose first that $\mathbf{x} \in \mathcal{A}_{d-k}$. Then by Lemma 7.10, we get for all $j = 1, \ldots, k$

$$\mathbf{x}_{p,k}^T\mathbf{v}_j^* = \mathbf{x}^T\mathbf{v}_j^\dagger - \prod_{i=1}^{j-1} \sin(r_i)\cos(r_j) = 0$$

and

$$\begin{aligned}\|\mathbf{x}_{p,k}\|^2 &= \mathbf{x}^T\mathbf{x}_{p,k} = \mathbf{x}^T\{\mathbf{x} - \cos(r_1)\mathbf{v}_1^* - \cdots - \prod_{i=1}^{k-1}\sin(r_i)\cos(r_k)\mathbf{v}_k^*\} \\ &= 1 - \cos^2(r_1) - \cdots - \prod_{i=1}^{k-1}\sin^2(r_i)\cos^2(r_k).\end{aligned}$$

Thus by rearranging terms, $\|\mathbf{x}_{p,k}\| = \prod_{i=1}^{k}\sin(r_i)$.

Conversely, suppose that $\mathbf{x}_{p,k}^T\mathbf{v}_j^* = 0$ for all $j = 1, \ldots, k$ and $\|\mathbf{x}_{p,k}\| = \prod_{i=1}^{k}\sin(r_i)$. Then since $\mathbf{x}_{p,k}, \mathbf{v}_1^*, \ldots, \mathbf{v}_k^*$ are orthogonal to each other,

$$\begin{aligned}\|\mathbf{x}\|^2 &= \left\|\mathbf{x}_{p,k} + \cos(r_1)\mathbf{v}_1^* + \cdots + \prod_{i=1}^{k-1}\sin(r_i)\cos(r_k)\mathbf{v}_k^*\right\|^2 \\ &= \mathbf{x}_{p,k}^T\mathbf{x}_{p,k} + \cos^2(r_1) + \cdots + \prod_{i=1}^{k-1}\sin^2(r_i)\cos^2(r_k) = 1.\end{aligned}$$

One can check that for all $j = 1, \ldots, k$

$$\mathbf{x}^T\mathbf{v}_j^* = \{\mathbf{x}_{p,k} + \prod_{i=1}^{j-1}\sin(r_i)\cos(r_j)\mathbf{v}_j^*\}^T\mathbf{v}_j^* = \prod_{i=1}^{j-1}\sin(r_i)\cos(r_j),$$

and again by Lemma 7.10, the result follows. (a) is directly obtained from (b). □

Proof of Proposition 7.9. (a) is readily derived by either Lemma 7.8 or the fact that $A_{m-1} \subsetneq S^m$ for all $m = 2, \ldots, d$.

For (b) and (c), it can be easily checked that $f_k \circ \cdots \circ f_1 : \mathcal{A}_{d-k} \longrightarrow S^{d-k}$ is a well defined bijective function. Since ρ_{d-k} is a metric and $\sin(r_i) > 0$, the metric ρ_{d-k} is nonnegative and symmetric, and the triangle inequality holds. In addition, since f is a bijection, $\rho_{d-k}^*(x, y) = 0$ if and only if $x = y$. This proves (b). Then by the definition of ρ_{d-k}^*, $f_k \circ \cdots \circ f_1$ is an isometry and (c) follows. □

Proof of Theorem 7.1. Let the singular value decomposition of the $(d + 1) \times n$ data matrix $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_n]$ be

$$\mathbf{X} = \sum_{i=1}^{n}\lambda_i\mathbf{u}_i\mathbf{v}_i^T,$$

where λ_is are the singular values, $\mathbf{V} = [\mathbf{v}_1 \cdots \mathbf{v}_n]$ is such that $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_n$, and $\mathbf{U} =$

$[\mathbf{u}_1 \cdots \mathbf{u}_n \mathbf{u}_{n+1} \cdots \mathbf{u}_{d+1}]$ is such that $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}_{d+1}$. Then $\mathcal{U} = \{\mathbf{u}_{n+1}, \ldots, \mathbf{u}_{d+1}\}$ is a an orthogonal basis set that complements $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$. For any $\mathbf{u} \in \mathcal{U}$, $\mathbf{u}^T\mathbf{X} = [\mathbf{u}^T\mathbf{x}_1 \cdots \mathbf{u}^T\mathbf{x}_n] = \mathbf{0}$. Therefore, $\rho(u, x_i) = \cos^{-1}(u_i^x) = \pi/2$ for all $i = 1, \ldots, n$. Write the orthogonal basis $\mathcal{U}$ as $\{v_1^*, \ldots, v_{d-n+1}^*\}$ and let $r_1 = \cdots = r_{d-n+1} = \pi/2$. Then by Lemma 7.8(b), there exist $\mathcal{A}_{d-1} \supset \cdots \supset \mathcal{A}_{n-1}$ such that $\mathbf{x}_i \in \mathcal{A}_{n-1}$. □

Proof of Theorem 7.2. Note that $\mathbf{w}^T\mathbf{w}^* = \mathbf{w}^T\mathbf{M}\mathbf{w} = 0$, and for all $\mathbf{z} \in S^d$ such that $\mathbf{w}^T\mathbf{z} \geqslant 0$ and $\mathbf{w}^T\mathbf{M}\mathbf{z} = 0$, $\mathbf{z}^T\mathbf{w}^* = \mathbf{z}^T\mathbf{M}\mathbf{w} = \mathbf{w}^T\mathbf{M}\mathbf{z} = 0$. Thus $\rho_d(\mathbf{w}, \mathbf{w}^*) = \cos^{-1}(\mathbf{w}^T\mathbf{w}^*) = \pi/2$ and $\rho_d(\mathbf{z}, \mathbf{w}^*) = \pi/2$. Moreover, since $\mathbf{w}^T\mathbf{z} \geqslant 0$, we have $\mathbf{w}, \mathbf{z} \in hA_{d-1}$. □

Proof of Theorem 7.3. Let $\mathcal{V}_\mathbf{W} = \{\mathbf{W}\Lambda : \Lambda^T = -\Lambda\}$, which is known as the vertical subspace of $\mathbf{W}$ (Kendall et al., 1999, p. 109), where such Λ is the $m \times m$ skew-symmetric matrix. Observe that the assumption (b) is equivalent to

$$\mathrm{tr}(\mathbf{Z}_i^T\mathbf{W}\Lambda) = 0, \quad \text{for all } \Lambda \text{ s.t. } \Lambda^T = -\Lambda.$$

Then for any $\mathbf{X} \in \mathcal{V}_\mathbf{W}$, $\mathrm{tr}(\mathbf{Z}_i^T\mathbf{X}) = \mathrm{tr}(\mathbf{Z}_i^T\mathbf{W}\Lambda) = 0$, i.e. $\mathcal{V}_\mathbf{W}$ is orthogonal to any $\mathbf{Z}_i$. The assumptions (c-d) assure that $\mathcal{V}_\mathbf{W}$ is $\frac{m(m-1)}{2}$ dimensional subspace of $\mathbb{R}^{m(k-1)}$ (Kendall et al., 1999, p. 109).

Let $\mathbf{v}_1^*, \ldots, \mathbf{v}_{m(m-1)/2}^*$ be a basis set of $\mathcal{V}_\mathbf{W}$, and let $r_1 = \cdots = r_{m(m-1)/2} = \pi/2$. Then by Lemma 7.8(b), there exists $\mathcal{A}_d = \mathcal{A}_{k(m-1)-1-m(m-1)/2}$ such that $\mathbf{Z}_i \in \mathcal{A}_d$. By checking that $\mathbf{W}$ is also orthogonal to $\mathcal{V}_\mathbf{W}$, $\mathbf{W} \in \mathcal{A}_d$. It is clear by the assumption (a) that $\mathbf{W}, \mathbf{Z}_i \in h\mathcal{A}_d$. □

Bibliography

Acker, A. F. (1974), "Absolute continuity of eigenvectors of time-varying operators," *Proc. Amer. Math. Soc.*, 42, 198–201.

Ahn, J., Marron, J. S., Muller, K. M., and Chi, Y.-Y. (2007), "The high-dimension, low-sample-size geometric representation holds under mild conditions," *Biometrika*, 94, 760–766.

Amaral, G. J. A., Dryden, I. L., and Wood, A. T. A. (2007), "Pivotal bootstrap methods for k-sample problems in directional statistics and shape analysis," *J. Amer. Statist. Assoc.*, 102, 695–707.

Aydin, B., Pataki, G., Wang, H., Bullitt, E., and Marron, J. S. (2009), "A principal component analysis for trees," *to appear in Annals of Applied Statistics.*

Baik, J., Ben Arous, G., and Péché, S. (2005), "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices," *Ann. Probab.*, 33, 1643–1697.

Baik, J. and Silverstein, J. W. (2006), "Eigenvalues of large sample covariance matrices of spiked population models," *J. Multivariate Anal.*, 97, 1382–1408.

Banerjee, A., Dhillon, I. S., Ghosh, J., and Sra, S. (2005), "Clustering on the Unit Hypersphere using von Mises-Fisher Distributions," *Journal of Machine Learning Research*, 6, 1345–1382.

Basser, P. J., Mattiello, J., and Le Bihan, D. (1994), "MR diffusion tensor spectroscopy and imaging," *Biophysics Journal*, 66, 259–267.

Bates, D. M. and Watts, D. G. (1988), *Nonlinear regression analysis and its applications*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, New York: John Wiley & Sons Inc.

Bhattacharjee, A., Richards, W., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J., and Meyerson, M. (2001), "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proc. Natl. Acad. Sci. USA*, 98(24):13790-5.

Bhattacharya, R. and Patrangenaru, V. (2003), "Large sample theory of intrinsic and extrinsic sample means on manifolds. I," *Ann. Statist.*, 31, 1–29.

— (2005), "Large sample theory of intrinsic and extrinsic sample means on manifolds. II," *Ann. Statist.*, 33, 1225–1259.

Bickel, P. J. and Levina, E. (2004), "Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations," *Bernoulli*, 10, 989–1010.

Billera, L. J., Holmes, S. P., and Vogtmann, K. (2001), "Geometry of the space of phylogenetic trees," *Adv. in Appl. Math.*, 27, 733–767.

Bingham, C. and Mardia, K. V. (1978), "A small circle distribution on the sphere," *Biometrika*, 65, 379–389.

Bookstein, F. L. (1986), "Size and Shape Spaces for Landmark Data in Two Dimensions," *Statist. Sci.*, 1, 181–242.

— (1991), *Morphometric tools for landmark data*, Cambridge: Cambridge University Press, geometry and biology, Reprint of the 1991 original.

Boothby, W. M. (1986), *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, 2nd ed.

Bradley, R. C. (2005), "Basic properties of strong mixing conditions. A survey and some open questions," *Probab. Surv.*, 2, 107–144 (electronic), update of, and a supplement to, the 1986 original.

Brillinger, D. R. and Stewart, B. S. (1998), "Elephant-seal movements: Modelling migration," *Canad. J. Statist.*, 26, 431–443.

Bullitt, E. and Aylward, S. (2002), "Volume rendering of segmented image objects," *IEEE Trans. Medical Imaging*, 21, 998–1002.

Buss, S. R. and Fillmore, J. P. (2001), "Spherical Averages and Applications to Spherical Splines and Interpolation," *ACM Transactions on Graphics*, 20, 95–126.

Casella, G. and Hwang, J. T. (1982), "Limit expressions for the risk of James-Stein estimators," *Canad. J. Statist.*, 10, 305–309.

Chernov, N. (2010), *Circular and Linear Regression: Fitting Circles and Lines by Least Squares*, Chapman & Hall.

Chikuse, Y. (2003), *Statistics on special manifolds*, vol. 174 of *Lecture Notes in Statistics*, New York: Springer-Verlag.

Churchill, R. V. and Brown, J. W. (1984), *Complex variables and applications*, New York: McGraw-Hill Book Co., 4th ed.

Cootes, T. F. and Taylor, C. J. (2001), "Statistical models of appearance for medical image analysis and computer vision," in *Proc. of SPIE Medical Imaging*, vol. 4322, pp. 236–248.

Cox, D. R. (1969), "Some Sampling Problems in Technology," in *New Developments in Survey Sampling*, eds. Johnson, U. L. and Smith, H., New York: Wiley Interscience.

Dryden, I. L. (2005), "Statistical analysis on high-dimensional spheres and shape spaces," *Ann. Statist.*, 33, 1643–1665.

Dryden, I. L. and Mardia, K. V. (1998), *Statistical shape analysis*, Wiley Series in Probability and Statistics: Probability and Statistics, Chichester: John Wiley & Sons Ltd.

Eaton, M. L. and Tyler, D. E. (1991), "On Wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix," *Ann. Statist.*, 19, 260–271.

El Karoui, N. (2008), "Spectrum estimation for large dimensional covariance matrices using random matrix theory," *Ann. Statist.*, 36, 2757–2790.

Elandt, R. C. (1961), "The folded normal distribution: two methods of estimating parameters from moments," *Technometrics*, 3, 551–562.

Fan, J. and Lv, J. (2010), "A selective overview of variable selection in high dimensional feature space," *Statistica Sinica*, 20, 101–148.

Fisher, N. I. (1993), *Statistical analysis of circular data*, Cambridge: Cambridge University Press.

Fisher, N. I., Lewis, T., and Embleton, B. J. J. (1993), *Statistical analysis of spherical data*, Cambridge: Cambridge University Press, revised reprint of the 1987 original.

Fisher, R. (1953), "Dispersion on a sphere," *Proc. Roy. Soc. London. Ser. A.*, 217, 295–305.

Fletcher, P. T. (2004), "Statistical Variability in Nonlinear spaces: Application to Shape Analysis and DT-MRI," Ph.D. thesis, University of North Carolina at Chapel Hill.

Fletcher, P. T. and Joshi, S. (2007), "Riemannian geometry for the statistical analysis of diffusion tensor data," *Signal Processing*, 87, 250–262.

Fletcher, P. T., Lu, C., Pizer, S. M., and Joshi, S. (2004), "Principal geodesic analysis for the study of nonlinear statistics of shape," *IEEE Trans. Medical Imaging*, 23, 995–1005.

Fletcher, R. (1971), "A modified Marquardt subroutine for non-linear least squares," Tech. Rep. AERE-R 6799.

Fréchet, M. (1944), "L'intégrale abstraite d'une fonction abstraite dúne variable abstraite et son application à la moyenne d'un élément aléatoire de nature quelconque," *Revue Scientifique*, 483512.

— (1948), “Les éléments aléatoires de nature quelconque dans un espace distancié,” *Ann. Inst. H. Poincaré*, 10, 215310.

Gaydos, T. L. (2008), “Data Representation and Basis Selection to Understand Variation of Function Valued Traits,” Ph.D. thesis, University of North Carolina at Chapel Hill.

Gerig, G., Styner, M., and Szekely, G. (2004), “Statistical shape models for segmentation and structural analysis,” *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)*, I, 467–473.

Gray, N. H., Geiser, P. A., and Geiser, J. R. (1980), “On the least-squares fit of small and great circles to spherically projected orientation data,” *Mathematical Geology*, 12, 173–184.

Hall, P., Marron, J. S., and Neeman, A. (2005), “Geometric representation of high dimension, low sample size data,” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67, 427–444.

Hastie, T. and Stuetzle, W. (1989), “Principal curves,” *J. Amer. Statist. Assoc.*, 84, 502–516.

Helgason, S. (2001), *Differential geometry, Lie groups, and symmetric spaces*, vol. 34 of *Graduate Studies in Mathematics*, American Mathematical Society.

Hotelling, H. (1933), “Analysis of a complex of statistical variables into principal components,” *J. Educ. Psychol.*, 24, 417–441, 498–520.

Huang, H., Liu, Y., and Marron, J. S. (2010), “Bi-Directional Discrimination with application to data visualization,” *submitted to J. Amer. Statist. Assoc.*

Huckemann, S., Hotz, T., and Munk, A. (2010), “Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric lie group actions,” *Statistica Sinica*, 20, 1–58.

Huckemann, S. and Ziezold, H. (2006), “Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces,” *Adv. in Appl. Probab.*, 38, 299–319.

Jeong, J.-Y., Stough, J. V., Marron, J. S., and Pizer, S. M. (2008), “Conditional-Mean Initialization Using Neighboring Objects in Deformable Model Segmentation,” in *SPIE Medical Imaging.*

John, S. (1971), “Some optimal multivariate tests,” *Biometrika*, 58, 123–127.

— (1972), “The distribution of a statistic used for testing sphericity of normal distributions,” *Biometrika*, 59, 169–173.

Johnson, N. L. (1962), “The folded normal distribution: accuracy of estimation by maximum likelihood,” *Technometrics*, 4, 249–256.

Johnstone, I. M. (2001), “On the distribution of the largest eigenvalue in principal components analysis,” *Ann. Statist.*, 29, 295–327.

Jolliffe, I. (2002), *Principal Component Analysis*, Springer, New York.

Joshi, S., Davis, B., Jomier, M., and Gerig, G. (2004), "Unbiased diffeomorphic atlas construction for computational anatomy," *Neuroimage*, 23, S151–160.

Joshi, S. C. and Miller, M. I. (2000), "Landmark matching via large deformation diffeomorphisms," *IEEE Trans. Image Process.*, 9, 1357–1370.

Jung, S. (2008), "Additional web materials," URL `http://www.unc.edu/~sungkyu`.

Jung, S., Dryden, I. L., and Marron, J. S. (2011a), "Analysis of Principal Nested Spheres," *Submitted in Biometrika.*

Jung, S., Foskey, M., and Marron, J. S. (2011b), "Principal Arc Analysis on direct product-manifolds,," *to appear in Annals of Applied Statistics.*

Jung, S. and Marron, J. S. (2009), "PCA consistency in high dimension, low sample size context," *Ann. Statist.*, 37, 4104–4130.

Jung, S., Sen, A., and Marron, J. S. (2011c), "Boundary behavior in high dimension, low sample size asymptotics of PCA in a spiked covariance model," *submitted to Ann. Statist.*

Karcher, H. (1977), "Riemannian center of mass and mollifier smoothing," *Comm. Pure Appl. Math.*, 30, 509–541.

Kendall, D. G. (1984), "Shape manifolds, Procrustean metrics, and complex projective spaces," *Bull. London Math. Soc.*, 16, 81–121.

Kendall, D. G., Barden, D., Carne, T. K., and Le, H. (1999), *Shape and shape theory*, Wiley Series in Probability and Statistics, Chichester: John Wiley & Sons Ltd.

Kendall, W. S. (1990), "Probability, convexity, and harmonic maps with small image. I. Uniqueness and fine existence," *Proc. London Math. Soc. (3)*, 61, 371–406.

Kenobi, K., Dryden, I. L., and Le, H. (2010), "Shape curves and geodesic modelling," *To appear in Biometrika.*

Kent, J. T., Dryden, I. L., and Anderson, C. R. (2000), "Using circulant symmetry to model featureless objects," *Biometrika*, 87, 527–544.

Kolmogorov, A. N. and Rozanov, Y. A. (1960), "On Strong Mixing Conditions for Stationary Gaussian Processes," *Theory Probab. Appl.*, 5, 204–208.

Krantz, S. G. (1999), *Handbook of complex variables*, Boston, MA: Birkhäuser Boston Inc.

Kume, A., Dryden, I. L., and Le, H. (2007), "Shape-space smoothing splines for planar landmark data," *Biometrika*, 94, 513–528.

Le, H. (2001), “Locating Fréchet means with application to shape spaces,” *Adv. in Appl. Probab.*, 33, 324–338.

Le, H. and Kume, A. (2000), “The Fréchet mean shape and the shape of the means,” *Adv. in Appl. Probab.*, 32, 101–113.

Lee, J. (1997), *Riemannian Manifolds: An Introduction to Curvature*, Springer.

Lee, M. H. (2007), “Continuum direction vectors in high dimensional low sample size data,” Ph.D. thesis, University of North Carolina.

Lee, S., Zou, F., and Wright, F. A. (2010), “Convergence and Prediction of Principal Component Scores in High-Dimensional Settings,” *to appear in Ann. Statist.*

Leone, F. C., Nelson, L. S., and Nottingham, R. B. (1961), “The folded normal distribution,” *Technometrics*, 3, 543–550.

Levina, E. and Bickel, P. (2005), “Maximum Likelihood Estimation of Intrinsic Dimension,” *Advances in NIPS*, 17.

Liu, Y., Hayes, D. N., Nobel, A., and Marron, J. S. (2008), “Statistical significance of clustering for high dimension low sample size data.” *J. Amer. Statist. Assoc.*, 103, 1281–1293.

Lu, S. and Kulkarni, V. G. (2008), “Geodesic mean of directional data on the circle,” Manuscript.

Mardia, K. V. and Gadsden, R. J. (1977), “A circle of best fit for spherical data and areas of vulcanism,” *J. Roy. Statist. Soc. Ser. C*, 26, 238–245.

Mardia, K. V. and Jupp, P. E. (2000), *Directional statistics*, Wiley Series in Probability and Statistics, John Wiley & Sons Ltd.

Marron, J. S., Jung, S., and Dryden, I. L. (2010), “Speculation on the Generality of the Backward Stepwise View of PCA,” in *Proceedings of MIR 2010: 11th ACM SIGMM International Conference on Multimedia Information Retrieval, Association for Computing Machinery, Inc., Danvers, MA, 227-230.*

Marron, J. S., Todd, M. J., and Ahn, J. (2007), “Distance-weighted discrimination,” *J. Amer. Statist. Assoc.*, 102, 1267–1271.

Merck, D., Tracton, G., Saboo, R., Levy, J., Chaney, E., Pizer, S., and Joshi, S. (2008), “Training models of anatomic shape variability,” *Med. Phys.*, 35, 3584–3596.

Moakher, M. (2002), “Means and averaging in the group of rotations,” *SIAM J. Matrix Anal. Appl.*, 24, 1–16 (electronic).

Muirhead, R. J. (1982), *Aspects of multivariate statistical theory*, New York: John Wiley & Sons Inc., wiley Series in Probability and Mathematical Statistics.

Nadler, B. (2008), "Finite sample approximation results for principal component analysis: a matrix perturbation approach," *Ann. Statist.*, 36, 2791–2817.

Paul, D. (2007), "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model," *Statistica Sinica*, 17, 1617–1642.

Pennec, X. (1999), "Probabilities and Statistics on Riemannian Manifolds: Basic Tools for Geometric measurements," in *International Workshop on Nonlinear Signal and Image Processing.*

Pennec, X., Fillard, P., and Ayache, N. (2006), "A Riemannian Framework for Tensor Computing," *International Journal of Computer Vision*, 66, 41–66.

Pesarin, F. and Salmaso, L. (2010a), "Finite-sample consistency of combination-based permutation tests with application to repeated measures designs," *J. Nonparametr. Stat.*, 22, 669–684.

— (2010b), *Permutation tests for complex data : theory, applications and software*, Chichester, U.K.: Wiley.

Pizer, S., Fletcher, P., Fridman, Y., Fritsch, D., Gash, A., Glotzer, J., Joshi, S., Thall, A., Tracton, G., Yushkevich, P., and Chaney, E. (2003), "Deformable M-Reps for 3D Medical Image Segmentation," *International Journal of Computer Vision - Special UNC-MIDAG issue*, 55, 85–106.

Pizer, S. M., Broadhurst, R. E., Levy, J., Liu, X., Jeong, J.-Y., Stough, J., Tracton, G., and Chaney, E. L. (2007), "Segmentation by Posterior Optimization of M-reps: Strategy and Results," *Unpublished*.

Qiao, X., Zhang, H. H., Liu, Y., Todd, M., and Marron, J. S. (2010), "Weighted Distance Weighted Discrimination and its asymptotic properties," *J. Amer. Statist. Assoc.*, 105, 401–414.

Rao, C. R. (1973), *Linear statistical inference and its applications*, John Wiley & Sons, New York-London-Sydney, 2nd ed., wiley Series in Probability and Mathematical Statistics.

Rao, N. R., Mingo, J. A., Speicher, R., and Edelman, A. (2008), "Statistical eigen-inference from large Wishart matrices," *Ann. Statist.*, 36, 2850–2885.

Rivest, L.-P. (1999), "Some linear model techniques for analyzing small-circle spherical data," *Canad. J. Statist.*, 27, 623–638.

Scales, L. E. (1985), *Introduction to nonlinear optimization*, New York: Springer-Verlag.

Schlkopf, B., Smola, A. J., and Mller, K.-R. (1998), "Nonlinear component analysis as a kernel Eigenvalue problem," *Neural Computation*, 10, 1299–1319.

Siddiqi, K. and Pizer, S. (2008), *Medial Representations: Mathematics, Algorithms and Applications*, Springer.

Srivastava, A. and Klassen, E. (2002), "Monte Carlo extrinsic estimators of manifold-valued parameters," *IEEE Trans. Signal Process.*, 50, 299–308.

Srivastava, A., Klassen, E., Joshi, S. H., and Jermyn, I. H. (2010), "Shape Analysis of Elastic Curves in Euclidean Spaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99.

Stewart, G. W. and Sun, J. G. (1990), *Matrix perturbation theory*, Computer Science and Scientific Computing, Boston, MA: Academic Press Inc.

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G., and Davis, R. W. (2005), "Significance analysis of time course microarray experiments," *Proc Natl Acad Sci USA.*, 102, 1283712842.

Thambisetty, M., Wan, J., Carass, A., An, Y., Prince, J. L., and Resnick, S. M. (2010), "Longitudinal changes in cortical thickness associated with normal aging," *NeuroImage*, 52, 1215 – 1223.

Tschirren, J., Palágyi, K., Reinhardt, J. M., Hoffman, E. A., and Sonka, M. (2002), "Segmentation, skeletonization and branchpoint matching—A fully automated quantitative evaluation of human intrathoracic airway trees," *Proc. Fifth International Conterence on Medical Image Computing and Computer-Assisted Intervention, Part II. Lecture Notes in Comput. Sci.*, 2489, 12–19.

Umbach, D. and Jones, K. N. (2003), "A few methods for fitting circles to data," *IEEE Transactions on Instrumentation and Measurement*, 52, 1881– 1885.

Vapnik, V. N. (1995), *The Nature of Statistical Learning Theory*, New York: Springer.

Wang, H. and Marron, J. S. (2007), "Object oriented data analysis: sets of trees," *Ann. Statist.*, 35, 1849–1873.

Wichers, L., Lee, C., Costa, D., Watkinson, P., and Marron, J. S. (2007), "A Functional Data Analysis Approach for Evaluating Temporal Physiologic Responses to Particulate Matter," Tech. Rep. 5, University of North Carolina at Chapel Hill, Department of Statistics and Operations Research.

Wilkinson, J. H. (1988), *The algebraic eigenvalue problem*, Monographs on Numerical Analysis, New York: The Clarendon Press Oxford University Press, oxford Science Publications.

Yata, K. and Aoshima, M. (2009), "PCA consistency for non-Gaussian data in high dimension, low sample size context," *Comm. Statist. Theory Methods*, 38, 2634–2652.

— (2010a), “Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix,” *to appear in Journal of Multivariate Analysis*.

— (2010b), “Intrinsic Dimensionality Estimation of High-Dimension, Low Sample Size Data with D-Asymptotics,” *Communications in Statistics - Theory and Methods*, 39, 1511–1521.