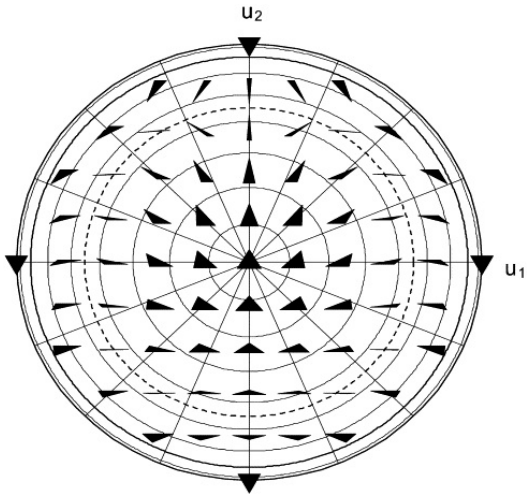
EIDÖGENSSISCHE TECHNISCHE HOCHSCHULE LAUSANNE
POLITECNICO FEDERALE DI LOSANNA
SWISS FEDERAL INSTITUTE OF TECHNOLOGY LAUSANNE

SCHOOL OF BASIC SCIENCES | MATHEMATICS INSTITUTE FOR ANALYSIS AND APPLICATIONS | Chair of Mathematical Statistics

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Statistics on Manifolds applied to Shape Theory

SEMESTER PROJECT (Spring 2013)
Department of Mathematics



Author :
Matthieu SIMEONI
matthieu.simeoni@epfl.ch

Under the supervision of :
Prof. Victor PANARETOS, SMAT, EPFL.

June 22, 2013

Abstract : In this report, we use a variety of tools from differential geometry to propose a nonlinear extension of the principal components analysis (PCA) into manifolds setting. This extension, that we shall call principal geodesics analysis (PGA), attempts to find analogs of the principal components by introducing the principal geodesic components. We then construct the shape space of triangles Σ_2^3 and find a convenient parametrization of it. Finally, we apply the PGA procedure previously designed to analyze the variability of a sample of shapes, randomly chosen onto the shape space of triangles.

Table of Contents

Introduction

The statistical shape theory is a mathematical theory aiming to design tools to efficiently analyze shapes of objects. This theory, introduced by D.G. Kendall in 1984, has known a considerable growth in interest from the scientific community with the advances in technology of the last decades. In fact, shape analysis appears to be of great interest in a wide range of disciplines, such as biology, medicine, image analysis, archaeology, geology, agriculture or even genetics (see Small [1] et Dryden [3]).

The word 'shape' is very commonly used in everyday language and usually refers to the appearance of an object. However, such a vague description isn't satisfactory for mathematical purposes. Therefore, the definition of shape that we consider is the intuitive one proposed by Kendall :

Definition 0.1 (*Shape*)

*The **shape** of an object can be define as the total of all information that is invariant under translations, rotations and rescaling (invariance under similarity transformations).*

Roughly speaking, we remove all information concerning *location, scale and orientation*. Thus, two objects can be said to have the same shape if they are similar in the sense of Euclidean geometry. With this definition, shape really is a *geometrical* invariant property of an object, and topological considerations aren't enough information to determine its shape. In practice, objects rarely have the same shape, due to measurement errors. In such cases, the variation in shape can often be the subject of statistical analysis.

Then, a primary goal of shape theory is to describe the variability of a population of geometric objects. An efficient tool to perform such an analysis would be the principal components analysis (PCA). However, this descriptive technique is linear in essence, and requires the data to be analyzed to lie in an Euclidean space. Unfortunately, complex representations of shape do not fulfill such a requirement, forcing us to re-design the technique for nonlinear situations. In fact, we will see that we can associate to a population of shapes a **shape space**, which can be seen in general as a *Riemannian manifold*.

This legitimates our interest for manifolds in Chapter 1, where we review some background concepts from differential geometry and briefly introduce topological and differentiable manifolds. In Chapter 2, we first recall the theory of principal components analysis, before investigating a possible extension of this tool into manifold setting : the principal geodesics analysis. Finally, we present in Chapter 3 the construction of the shape space of triangles Σ_2^3, and apply to this example the tools previously designed for manifolds.

CHAPTER

1 Introduction to Manifolds

In this chapter we review some background concepts and basic tools from differential geometry, in order to shortly introduce topological and differentiable manifolds, which will play a key role in the context of shape analysis.

1. Background concepts of differential geometry

1.1. Differentiability of a multivariate function

Let U and V be two open sets $U \subset \mathbb{R}^p$, $V \subset \mathbb{R}^q$. We first begin by generalizing the concepts of derivative and differentiability of univariate analysis to multivariate functions.

Definition 1.1 (*Differentiability*)

*A function $f : U \to V$ is said to be differentiable at a point $x_0 \in U$ if there exists a **linear map** $df_{x_0} : \mathbb{R}^p \to \mathbb{R}^q$ (called the **derivative** of f in x_0) such that :*

$$\lim_{h \to 0} \frac{f(x_0 + h) - f(x_0) - df_{x_0}(h)}{\|h\|} = 0$$

*Therefore, if f is differentiable in x_0, f admits a **first order taylor expansion** which is, for h in the neighborhood of 0 :*

$$f(x_0 + h) = f(x_0) + df_{x_0}(h) + o(\|h\|)$$

Remarks :
- f is differentiable in $x_0 \iff f$ admits a first order taylor expansion in x_0.
- if f is a linear map from U to V, then f is differentiable and $\forall x_0 \in U, df_{x_0} = f$.
- if $p = 1$, $x_0 \in \mathbb{R}$ and f differentiable in x_0, then we have $f'(x_0) = df_{x_0}(1)$.

The first remark allow us to formulate a first result about differentiable functions, making the link between differentiability and continuity.

Proposition 1.1

$f : U \to V$ and $x_0 \in U$. If f is differentiable in x_0 then f is continuous in x_0.

Proof : As f is differentiable, f admits a first order taylor expansion. Therefore, $f(x_0 + h) = f(x_0) + df_{x_0}(h) + o(\|h\|)$. Letting h tend to 0 we obtain :

$$\lim_{h \to 0} f(x_0 + h) = f(x_0),$$

which achieves the proof. Note that we've used the continuity of the derivative df_{x_0} to state the result, which is indeed continuous, as a linear function between two finite dimensional spaces and therefore lipschitz and so continuous. ■

Definition 1.1 doesn't provide us any systematic way to calculate the derivative of a multivariate function. This wish to find a general expression for the derivative of a multivariate function make us introduce the concepts of **directional derivative** and **partial derivative**.

Definition 1.2 (*Directional Derivative / Gâteaux differential*)

*Let f be a mapping from U to V, $x_0 \in U$ and $v \in \mathbb{R}^p$. If it exists, we define the **directional derivative** (or Gâteaux differential) of f at x_0 in the direction v provided that :*

$$\forall x \in \mathbb{R}^p, \quad D_v f(x_0) = \lim_{t \to 0} \frac{f(x_0 + tv) - f(x_0)}{t}.$$

Looking at definition 1.2 make us guess the underlying link between directional derivative and derivative of a multivariate function. Intuitively, we can think of the directional derivative as a radial information on the variations of f, while the derivative of f is carrying the whole information on the variations on f, not only along a direction. The following proposition formally states the link between these two quantities :

Proposition 1.2

Let $f : U \to V$ and $x_0 \in U$. If f is differentiable in x_0 then $\forall v \in \mathbb{R}^p$, f admits a directional derivative in the direction v and we have :

$$D_v f(x_0) = df_{x_0}(v).$$

Proof : f is differentiable so :

$$\begin{aligned} f(x_0 + tv) &= f(x_0) + df_{x_0}(tv) + o(t\|v\|) \\ &= f(x_0) + t df_{x_0}(v) + o(t). \end{aligned}$$

So

$$D_v f(x_0) = \lim_{t \to 0} \frac{f(x_0 + tv) - f(x_0)}{t} = df_{x_0}(v).$$ ■

Remark : ⚠ The converse is not true : the existence of all directional derivatives do not imply differentiability. In fact if we consider the following function :

$$f : \mathbb{R}^2 \to \mathbb{R}, (x, y) \mapsto \begin{cases} \frac{y^2}{x} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases},$$

then all the directional derivatives exist in 0, but it's not a differentiable function, as it's not even continuous ($\lim_{\epsilon \to 0} f(\epsilon^2, \epsilon) = 1 \neq 0$).

We now introduce **partial derivatives** as special case of directional derivatives.

Definition 1.3 (*Partial derivative*)

*Let $f : U \to V$, $x_0 \in U$, $e = \{e_1, ..., e_p\}$ a basis of $\mathbb{R}^p$ and $j \in \mathbb{N}$. If it exists, we call **jth partial derivative** the directional derivative of f at x_0 in the direction of v. We have :*

$$D_j f(x_0) = \frac{\partial f}{\partial x_j}(x_0) = D_{e_j} f(x_0) = df_{x_0}(e_j).$$

Partial derivatives allow us to calculate derivative of a multivariate function, as states the following proposition :

Proposition 1.3

Let $f : U \to V$, $x_0 \in U$, f differentiable in x_0. Then, all partial derivatives exist and for every $h = \sum_{j=1}^{p} h_j e_j$, we have :

$$df_{x_0}(h) = \sum_{j=1}^{p} h_j df_{x_0}(e_j) = \sum_{j=1}^{p} h_j \frac{\partial f}{\partial x_j}(x_0).$$

We can rewrite this proposition using matrix notation. To this aim, we introduce the **Jacobian matrix** defined as follows.

Definition 1.4 (*Jacobian Matrix*)

*Let $x_0 \in U$. If $f : U \to V$ is a **smooth function** (i.e. all first order partial derivatives exist) then we define the **Jacobian matrix** $J_f(x_0) \in \mathbb{R}^{p \times q}$ provided that :*

$$J_f(x_0) = \left(\frac{\partial f_i}{\partial x_j}(x_0) \right)_{i,j}, \quad 1 \leq i \leq q,\ 1 \leq i \leq p.$$

*When p=q the Jacobian matrix is a square matrix and we call the determinant of this matrix the **Jacobian**.*

Remarks :

- a function doesn't need to be differentiable for the Jacobian matrix to exist, since only the partial derivatives are required to exist.
- the Jacobian matrix generalizes the gradient of a scalar valued function of multiple variables (the gradient can be regarded as a special case of the Jacobian matrix).
- **Geometric interpretation :** the Jacobian measures the rate of change of volume induced by the transformation $x \mapsto f(x)$ locally around x (the reader can convince itself about it by looking at the substitution formula for multiple variables functions).

Therefore, using the definition 1.4, proposition 1.3 becomes :

$$df_{x_0}(h) = J_f(x_0) \times h,$$

which gives us a coordinate representation of the derivative of f.

► **Geometric Interpretation**

As df_{x_0} is the best linear approximation of f, we can see the linear space Vect$\{\frac{\partial f}{\partial x_1}(x_0), \ldots, \frac{\partial f}{\partial x_n}(x_0)\}$generated by the columns of f as the best linear subspace to approximate the graph of f.
More precisely, it can be shown that $J_f(x_0)$ gives the **orientation** of the **tangent space** to the graph of f in x_0. In fact, as we've seen, each **tangent vector** $D_v f(x_0)$ to the graph in x_0 can be written as a linear combination of the columns of the Jacobian matrix :

$$D_v f(x_0) = df_{x_0}(v) \overset{1.3}{=} \sum_{j=1}^{p} h_j \frac{\partial f}{\partial x_j}(x_0).$$

To conclude this subsection, we present a result allowing us to characterize the notion of differentiability :

Theorem 1.1

Let $f : U \to V$ and $x_0 \in U$. If f is $\mathcal{C}^1$ then f is differentiable.

✧✧✧

1.2. $\mathcal{C}^r$-functions

We define here the class of $\mathcal{C}^r$-functions for multivariate functions.

Definition 1.5

*We say that $f : U \to V$ is a $\mathcal{C}^r$-**function** on U for any $r \in \mathbb{N}$ if f has $\mathcal{C}^0$ partial derivatives :*

$$\frac{\partial^{r_1+\cdots+r_p} f}{\partial x_1^{r_1} \cdots \partial x_p^{r_p}},$$

for all non negative integers $r_1, r_2, \ldots, r_p$ such that $r_1 + \cdots + r_p \leq r$.

Clearly, a $\mathcal{C}^r$-function on U is also $\mathcal{C}^s$ for all $s \leq r$. If f is a $\mathcal{C}^r$-function $\forall r \geq 1$, than we say that f is a $\mathcal{C}^\infty$-**function**.

✧✧✧

1.3. Homeomorphisms and Diffeormorphisms

One of the last concept we need in order to be able to fully describe a manifold, is the concept of **homeomorphism**. Roughly speaking, we can see homeomorphisms as isomorphisms between **topological spaces** : they are mappings which preserves the topological properties of the spaces they put in correspondence.
For now, we will suppose that U and V are two open sets of $\mathbb{R}^p$.

Definition 1.6 (*Homeomorphisms*)

*Let f a bijective mapping from U to V. Then, f is said to be a **homeomorphism** from U to V provided that both f and f^{-1} are continuous.*
*When an homeomorphism can be established between U and V then we say that U and V are **homeomorphic**.*

There exists a very convenient result allowing us to **characterize** homeomorphisms without using the topological definition of continuity which can be a little cumbersome to work with. We provide this result without any proof :

Proposition 1.4 (*Characterization of homeomorphisms*)

*A function $f : U \to V$ is an homeomorphism **iif** it's a continuous, open (i.e the image of an open set of U is an open set of V) and bijective function.*

We now introduce a special class of homeomorphisms, called **diffeomorphisms**. These are basically homeomorphisms with some additional constraints on the smoothness of f and f^{-1}.

Definition 1.7 (*$\mathcal{C}^r$-diffeomorphism*)

*A homeomorphism f is called a $\mathcal{C}^r$ **-diffeomorphism** between U and V if both f and f^{-1} are $\mathcal{C}^r$-functions.*
*We will refer to a $\mathcal{C}^\infty$-diffeomorphism simply as a diffeormorphism. When a diffeomorphism can be established between U and V , we say that they are **diffeomorphic**.*

The apparent symmetry in f and f^{-1} of the above definition motivates the following proposition :

Proposition 1.5

If $f : U \to V$ is a $\mathcal{C}^r$ diffeomorphism then f^{-1} is a $\mathcal{C}^r$ diffeomorphism and we have :

$$\forall a \in U, \quad J_{f^{-1}}(f(a)) = J_f(a).$$

We terminate this section by a pretty convenient theorem, allowing us to characterize diffeomorphisms.

Theorem 1.2 (*Global Inversion*)

Let $f : U \to \mathbb{R}^p$ an injective $\mathcal{C}^1$-function. f is a $\mathcal{C}^1$-diffeomorphism from $U \to f(U)$ if :

$$\forall a \in U, \quad det(J_f(a)) \neq 0.$$

Proof (Idea) : We won't discuss the detailed proof of this theorem here, but we will only give to the reader some intuition about this result, focusing on the one-dimensional case. Let f be a real valued function as in the previous theorem. Looking at the following differentiation formula : $(f^{-1})' = \frac{1}{f' of^{-1}}$, we notice that the only thing we need for the derivative of f^{-1} to be continuous is that $f'(a) \neq 0, \ \forall t \in U$. In multivariate analysis, the derivative becomes the Jacobian matrix, which allow us to make an analogy with the one dimensional case. ■

2. Introduction to topological and differential manifolds

A differential manifold is a generalization of our basic understanding of a curved surface of the Euclidean space $\mathbb{R}^3$. An informal definition of this mathematical object, could be the following : *a manifold is a space that locally resembles the Euclidean space*. This local similarity with the Euclidean space will appear to be very convenient, as it will allow us to extend all the tools and constructions of multivariate calculus we previously introduced for the Euclidean space to any differential manifold. Therefore, we will focus on this local resemblance to define and describe manifolds.

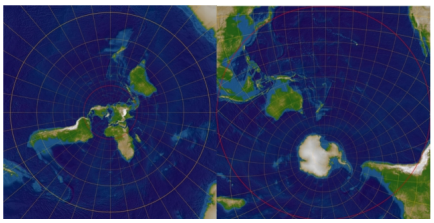


Figure 1.1.: *At least two charts are needed to fully describe the surface of the earth (stereographic projections).* ***©Wikipédia****.*

The first thing we need to do is to precise in a mathematical way what we mean by *"local resemblance"*. To this end, we will introduce the notion of **atlas**, a collection of **charts** which will help us to describe a manifold. This designation is inspired from our own description of the earth (which is indeed a manifold) : being a sphere, the earth isn't an Euclidean space, but still can be locally charted trough projective maps (the Mercator projection is the most famous projection of the earth).

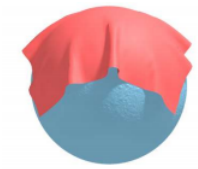


Figure 1.2.: *It's impossible to cover a sphere with a sheet without wrapping it.*

However, in order to include every point of the earth in this description, one map isn't enough (see fig. 1.1). The reader can convince itself about this fact by tempting to cover a sphere by a sheet : it is impossible to cover the whole sphere under this sheet without wrapping the sheet, and therefore destroying is Euclidean structure (see fig. 1.2). Analogously, we might use several charts regrouped in an atlas to describe a manifold. As these charts are covering the whole manifold, it might occurs that they overlap in some places. In this case, we'd like that the different descriptions provided by the charts on the region of overlap be coherent : they might not be identical, but at least *compatible*. This notion of compatibility will be precised using homeomorphisms and diffeormorphisms, which will provide a way to pass from one description to the other.
Most of the results presented below, are adapted from the chapter 2 of Small [1].

✧✧✧

2.1. Topological and differential manifolds

Let M^p be a **topological space**[1] with a collection of open subsets $\{U_\alpha : \alpha \in A\}$ such that[2] $\cup_\alpha U_\alpha = M^p$.

[1] We don't recall here the definition of a topology. Nevertheless, any reader not familiar with the subject should familiarize himself with these notions in the literature before continuing the reading of this report.

[2] Note that the collection isn't necessarily the whole topology.

We also assume the existence of a collection of functions :

$$c_\alpha : U_\alpha \to \mathbb{R}^p,$$

that are all **homeomorphisms** onto the open subsets $c_\alpha(U_\alpha) \subset \mathbb{R}^p$.

Definition 2.1 (*Charts*)

*We say that the functions c_α are **charts** on M^p provided that :*

$$c_\beta \circ c_\alpha^{-1} : c_\alpha(U_\alpha \cap U_\beta) \to c_\beta(U_\alpha \cap U_\beta), \tag{1.1}$$

*is a **homeomorphism** from $c_\alpha(U_\alpha \cap U_\beta)$ to $c_\beta(U_\alpha \cap U_\beta)$, $\forall \alpha, \beta \in A$.*

► Interpretation (see also fig. 1.3)

- We can think of the charts $\{c_\alpha\}_{\alpha \in A}$ as providing local coordinate systems on M^p. This a formal translation of the assertion : a manifold locally resembles the Euclidean space. The charts provide this local Euclidean structure we wish to work with.
- The patching criterion eq. (1.1) provides a formal characterization of what we ingeniously called *compatibility* of two charts on the region of overlapping : the coordinate systems provided by the two different charts can be mapped together in a topologically consistent way ($c_\alpha(U_\alpha \cap U_\beta)$ and $c_\beta(U_\alpha \cap U_\beta)$ are homeomorphic).

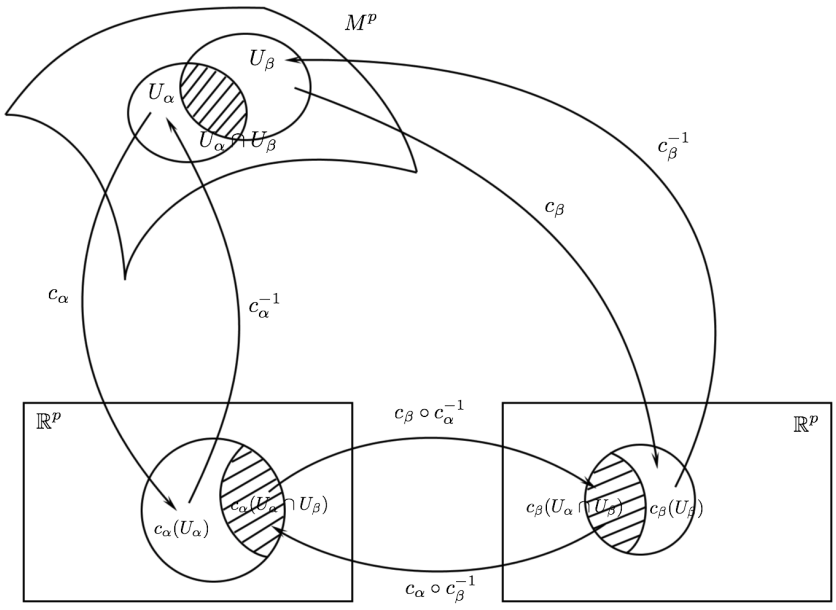


Figure 1.3.: *Charts provide local coordinate systems on M^p. The patching criterion eq.* (1.1) *ensures the compatibility of two coordinate systems on a region of overlapping.*

Definition 2.2 (*Atlas and Topological Manifold*)

*The collection of subsets $\{(U_\alpha, c_\alpha)\}_{\alpha\in A}$ is said to form an **atlas** on M^p. The set M^p together with its atlas is called a **topological manifold of dimension** p.*

Remark (augmenting the Atlas) : As in $\mathbb{R}^p$ it's common to change coordinates for convenience of calculations, the same is true for topological manifolds. Here, a change of coordinates correspond of a change of charts. Therefore, we can add charts to the original atlas under the condition that they respect the patching criterion eq. (1.1).

The next proposition offers us a different interpretation of our basic understanding that a topological manifold is locally homeomorphic to the Euclidean space. Here, we forget the geometric identification through coordinate systems provided by charts and focus on the topologies of the two spaces. The following result allow us to characterize elements of the topology of M^p through elements of the topology of $\mathbb{R}^p$:

Proposition 2.1 (*Characterization of open subsets of M^p*)

If $c_\alpha(V \cap U_\alpha)$ is an open subset of $\mathbb{R}^p$, $\forall\alpha \in A$, then V is an open subset of M^p.

Proof : c_α is a homeomorphism so it's bijective and c_α^{-1} is continuous so, by the topological definition of continuity, $\forall\alpha \in A$:

$$c_\alpha^{-1}(c_\alpha(V \cap U_\alpha)) = V \cap U_\alpha$$

is an open subset of M^p. Thus, we have $\cup_{\alpha\in A} V \cap U_\alpha$ open subset of M^p as union of opens subsets of M^p and $\cup_{\alpha\in A} U_\alpha = M^p$, which gives us : $V \cap M^p = V$ open subset of M^p. ∎

By strengthening the smoothness of the patching criterion we obtain a special class of topological manifolds, called differential manifolds.

Definition 2.3 (*$\mathcal{C}^r$-differential Manifold*)

*If the functions $c_\beta \circ c_\alpha^{-1}$ are also required to be $\mathcal{C}^r$-diffeormorphisms then the topological manifold M^p is said to be a **$\mathcal{C}^r$-differential manifold**.*
*We shall refer to $\mathcal{C}^\infty$-differential manifolds simply as **differential manifolds**.*

Now let M^p and N^q be two differential manifolds of dimension p and q. We wish to extend the notion of differentiability introduced in multivariate analysis to a function between two manifolds. Let see how we can exploit the local resemblance of manifolds with the Euclidean space to do so.

Definition 2.4 (*Differentiability on Manifolds*)

A continuous function

$$h : M^p \to N^q$$

*is said to be **differentiable** if for every $x \in M^p$ there exists a chart (U_α, c_α) on M^p and a chart (V_β, c_β) on N^q such that: $x \in U_\alpha$, $h(x) \in V_\beta$, and the mapping :*

$$c_\beta \circ h \circ c_\alpha^{-1} : c_\alpha(h^{-1}(V_\beta) \cap U_\alpha) \to \mathbb{R}^q$$

is differentiable.

► Idea

The idea here is to define the notion of differentiability of a function between two manifolds using the knowledge we already have of the Euclidean case, and the local resemblance between a manifold and the Euclidean space (see fig. 1.4).

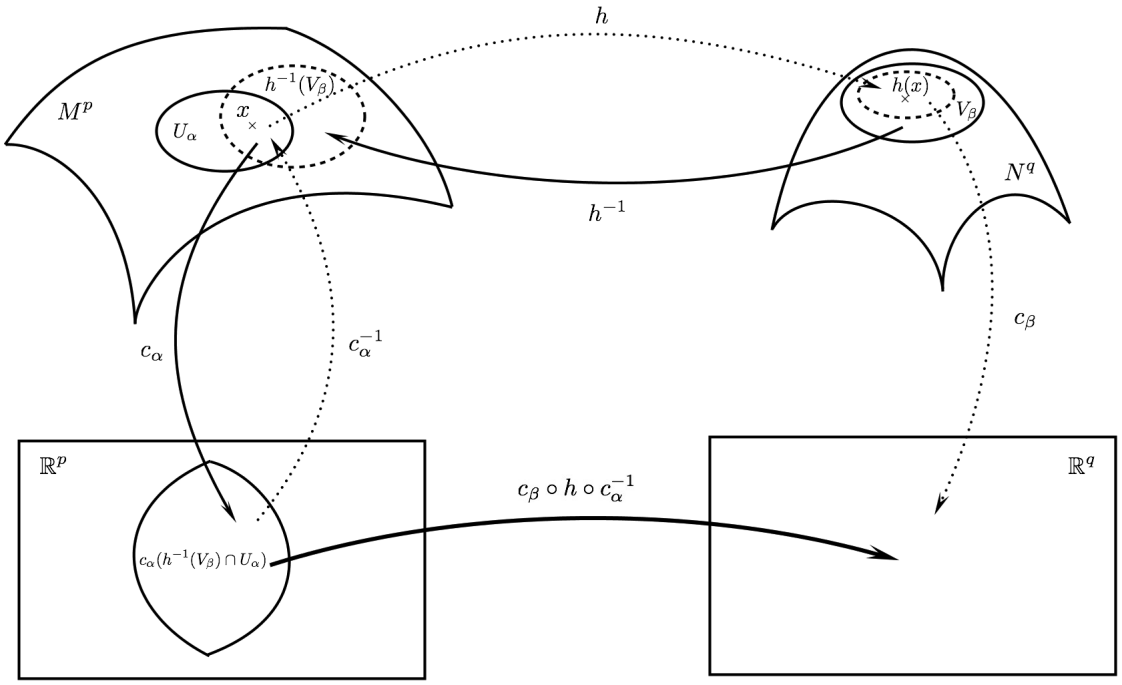


Figure 1.4.: *Defining differentiability of functions on manifolds through the local resemblance with the Euclidean space. The plain-line arrows show how we get back to the euclidean case. The dotted-line arrows show the steps of the mapping $c_\beta \circ h \circ c_\alpha^{-1}$.*

Similarly, we say that h is a $\mathcal{C}^r$**-function** provided that $c_\beta \circ h \circ c_\alpha^{-1}$ is a $\mathcal{C}^r$-function. If $p = q$ and h a 1-1 map, then h is called $\mathcal{C}^r$**-diffeomorphism** if both h and h^{-1} are $\mathcal{C}^r$. When $\mathcal{C}^r$-diffeomorphismcan be established between two manifolds M^p and N^q then they are said to be $\mathcal{C}^r$**-diffeomorphic**.

Notation : the charts provide a system of coordinates for the manifold. For convenience, we will say that x has coordinates $(x_1, x_2, \ldots, x_p)$ rather than the more precise statement that these coordinates belongs to $c_\alpha(x)$.

Definition 2.5 (*Intrinsic/extrinsic properties*)

We say that a property of a manifold is ***intrinsic*** *if it's invariant under a change of coordinates compatible with the differential structure. Respectively, a property is said* ***extrinsic*** *if it depends on the coordinate system.*

✧✧✧

2.2. Tangent vectors and tangent spaces

We now wish to generalize the definition of **tangent vectors** and **tangent spaces** for surfaces in $\mathbb{R}^3$ to any differential manifold. One first attempt could be to try to generalize these notions trough the Jacobian matrix. In fact, if we consider a surface $\Sigma \subset \mathbb{R}^3$ implicitly defined by $\Sigma = \{(x, y, z) \in \mathbb{R}^3 : h(x, y, z) = 0\}$ with $h : \mathbb{R}^3 \to \mathbb{R}$ sufficiently smooth, one could define the tangent space to the surface in $(x_0, y_0, z_0) \in \Sigma$ through the gradient $\nabla h(x_0, y_0, z_0) : v \in \mathbb{R}^3$ *is a tangent vector to* Σ *at* (x_0, y_0, z_0) *if* $\langle v, \nabla h(x_0, y_0, z_0) >= 0$. Therefore, a natural choice should be to try to generalize tangent vectors through the generalization of the gradient itself, which is the Jacobian matrix. But unfortunately, this approach isn't satisfactory. First of all, we'd like a definition which doesn't require any parametrization for the manifold : the fact that a vector is tangent to a manifold shouldn't depend on the existence or not of a parametrization for this manifold. Secondly, we'd like a definition that doesn't depend on the embedding of the manifold in a greater Euclidean space[3]: the property of tangency to a manifold is an *intrinsic* property of this manifold. The definition we propose here fulfills both of these criteria, defining tangent vectors through **equivalence classes**. Let M^p be a differential manifold. Let $x(t)$ and $y(t)$ be **two smooth paths**[4] in M^p passing through a common point x_0 at $t = 0$. Let us suppose that we have a coordinate system through a chart (U_α, c_α) around x_0 such that the paths have the coordinates :

$$\begin{aligned} x(t) &= (x_1(t), \dots, x_p(t)), \\ y(t) &= (y_1(t), \dots, y_p(t)), \\ x_0 &= (x_{01}, \dots, x_{0p}). \end{aligned}$$

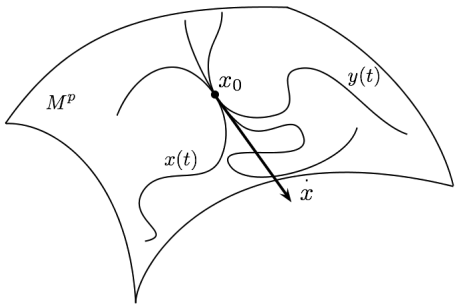


Figure 1.5.: *Tangent vector* $\dot{x}$ *at* x_0*, seen as the equivalence class of all smooth paths tangent in* x_0*.*

Then, the paths $x(t)$ and $y(t)$ are said to be **tangent** at x_0 if :

$$\frac{dx_j}{dt}(0) = \frac{dy_j}{dt}(0), \quad \forall j = 1, ..., p. \tag{1.2}$$

If we call $\mathcal{S}$ the set of all smooth paths passing through x_0 at $t = 0$, we are able to define an **equivalence relation** $R \subset \mathcal{S} \times \mathcal{S}$: two paths $x(t)$, $y(t) \in \mathcal{S}$ are said to be **equivalent** if they are tangent, in the sense of eq. (1.2). We write $x(t) \sim_R y(t)$. It's easy to show that this relation is indeed an equivalence relation. This equivalence relation allow us to define tangent vectors as equivalence classes :

Definition 2.6 (*Tangent Vector*)

We define the ***tangent vector*** $\dot{x}$ *to the path* $x(t)$ *at the point* $x_0 = x(0)$ *to be the* ***equivalence class*** *of* $x(t)$ *:*

$$\dot{x} = \{y(t) \in \mathcal{S} : x(t) \sim_R y(t)\}.$$

[3] the scalar product between v and ∇h is a property of the Euclidean space $\mathbb{R}^3$ in which Σ is embedded.
[4] $x(t)$ and $y(t)$ are said to be smooth if their coordinates are differentiable functions of t.

Then, the **tangent space** will simply be defined as the set of all tangent vectors :

Definition 2.7 (*Tangent Space*)

*The set of all tangent vectors to the manifold M^p at x_0 is called **tangent space** at x_0 and is denoted by $T_{x_0}(M^p)$.*

Remark : In term of the equivalence relation R, we have : $T_{x_0}(M^p) = \mathcal{S}/_{\sim_R}$. In fact, $\mathcal{S}/_{\sim_R}$ is the set of all equivalence classes which is exactly the same as the set of all tangent vectors.

At this point, we have formally defined tangent vectors and tangent space at a point $x_0 \in M^p$, through equivalence classes. During this process, we used meaningful mathematical terminology such as *vector* to designate equivalence classes, which can seem a little peculiar at first glance. Thus, we'd like to show that those equivalence classes deserve this designation, and can be equipped with a linear space structure.
To do so, we must define an addition between two vectors and a multiplication of a vector by a scalar. Once those basic operations defined, proving that the tangent space is indeed a linear space is straightforward.

Definition 2.8 (*Addition of two tangent vectors*)

*We define the **vector sum** $\dot{x} + \dot{z}$ to be the tangent vector at x_0 to the path whose coordinates are :*

$$(x_1(t) + z_1(t) - x_{01}, \dots, x_p(t) + z_p(t) - x_{0p}),$$

which passes through x_0 at $t = 0$.

Remark : ⚠ We can add tangent vectors at the same point x_0, but we cannot add tangent vectors to the manifold at different points.

To be well-defined, this sum must be independent of the system of coordinates and of the choice of the representatives paths of the equivalence classes. This is not obvious. To show that this sum is well independent of the system of coordinates, it's sufficient to notice that the diffeomorphic images of two tangent paths in $\mathbb{R}^p$ will also be tangent. Let now show that this sum is independent of equivalence classes representative :
Let $x(t)$,$y(t)$,$z(t)$ and $w(t)$ be four smooth paths passing trough x_0 at $t = 0$ and such that $\dot{x} = \dot{y}$, $\dot{z} = \dot{w}$. If we call $\gamma(t) = (x_1(t) + z_1(t) - x_{01}, \dots, x_p(t) + z_p(t) - x_{0p})$ and $\mu(t) = (y_1(t) + w_1(t) - x_{01}, \dots, y_p(t) + w_p(t) - x_{0p})$ then, showing that $\dot{x} + \dot{z} = \dot{y} + \dot{w}$ is equivalent to show that $\gamma(t)$ and $\mu(t)$ are tangent in $t = 0$:

$$\frac{d\gamma_j}{dt}(0) = \frac{dx_j}{dt}(0) + \frac{dz_j}{dt}(0) = \frac{dy_j}{dt}(0) + \frac{dw_j}{dt}(0) = \frac{d\mu_j}{dt}(0), \quad \forall j = 1, \dots, p,$$

and therefore $\mu(t)$ and $\gamma(t)$ are tangent.
Similarly, we define the multiplication by a scalar $\lambda \in \mathbb{R}$.

Definition 2.9 (*Multiplication by a scalar*)

Let $\dot{x}$ be a tangent vector, $\lambda \in \mathbb{R}$, then $\lambda\dot{x}$ is defined as the tangent vector at x_0 to the following path :

$$(\lambda(x_1(t) - x_{01}) + x_{01}, \dots, \lambda(x_p(t) - x_{0p}) + x_{0p}),$$

which passes through x_0 at $t = 0$.

One more time, it can be shown that this definition is well-defined.

Having defined these two basic operations on our tangent vectors, we provided our tangent space a linear structure[5] . It can be shown that the tangent space $T_{x_0}(M^p)$ has the same dimension as M^p [6] . Therefore, $T_{x_0}(M^p)$ is linearly isomorphic to $\mathbb{R}^p$, and we can construct a set of basis vectors for this linear space $\mathcal{B} = \{\partial_1(x_0), \dots, \partial_p(x_0)\}$, defined as tangent vectors to the paths :

$$\forall j = 1, \dots, p : \quad t \mapsto (x_{01}, \dots, x_{0j} + t, \dots, x_{0p}),$$

defined in a neighborhood of x_0 around $t = 0$. Since the cardinality of the basis is equal to the dimension of $T_{x_0}(M^p)$, to show that this is well a basis it is sufficient to show linear independence between the elements of the basis. Let suppose that there exist $\alpha_1, \dots, \alpha_p \in \mathbb{R}$ such that :

$$\sum_{i=1}^{p} \alpha_i \partial_i(x_0) = 0,$$

with 0 the null tangent vector, defined as the tangent vector to the constant path : $c(t) : t \mapsto (x_{01}, \dots, x_{0p})$. Thus, if we call $\beta(t)$ one possible representative of the vector $\sum_{i=1}^{p} \alpha_i \partial_i(x_0)$, then $\beta(t)$ and $c(t)$ are tangent at x_0 and we have :

$$\begin{aligned} \frac{d\beta}{dt}(0) &= \frac{dc}{dt}(0), \\ \iff \sum_{i=1}^{p} \alpha_i(0, \dots, 1, \dots, 0) &= (0, \dots, 0), \end{aligned}$$

which implies that all the α_i must be null, and therefore $\mathcal{B}$ is indeed a basis.

This basis allow us to give a coordinate representation of tangent vectors :

$$\forall v \in T_{x_0}(M^p), \quad v(x_0) = \sum_{i=1}^{p} \alpha_i \partial_i(x_0).$$

[5] To be more accurate, the definition of the two basic operations aren't enough to provide linear structure, and we should check all the other properties, but here they are all verified as we use the standard multiplication and addition laws to define the paths.

[6] This is reassuring, at least intuitively : we tried to build definition of tangent vectors and tangent space as intrinsic objects to the manifold, i.e objects which do not require the manifold to be embedded in a greater space to exist. Therefore, it would have been peculiar if the resulting tangent space of our construction were of greater dimension than the manifold itself.

Therefore, we can write, for each given time :

$$\dot{x}(t) = \sum_{i=1}^{p} \dot{x}_i \partial_i(x(t)), \tag{1.3}$$

where $\dot{x}_i = \frac{dx_i}{dt}(t)$ (by unicity of the representation in the considered basis).
It should be noted that the definition of the basis vector depends on the coordinate system used : for a different coordinate system we obtain different bases, but they both span $T_{x_0}(M^p)$.

✧✧✧

2.3. Metric tensors and Riemannian Manifolds

In the previous section, we showed that we could provide our tangent space a linear structure. Here we investigate wether we can provide it a metric structure, through the definition of an inner product between tangent vectors.
Suppose that :

$$g(x) = \begin{pmatrix} g_{11}(x) & \cdots & g_{1p}(x) \\ \vdots & \ddots & \vdots \\ g_{p1}(x) & \cdots & g_{pp}(x) \end{pmatrix}$$

is a positive definite matrix for all $x \in M^p$. We additionally suppose that the functions g_{ij} are smooth functions on M^p, $\forall i, j$. If $a(x) = \sum_j a_j(x)\partial_j(x)$ and $b(x) = \sum_j b_j(x)\partial_j(x)$, then $g(x)$ defines an **inner product** on $T_x(M^p)$:

$$\langle a(x), b(x) \rangle = a(x)^T g(x) b(x) = \sum_{j=1}^{p} \sum_{k=1}^{p} g_{jk}(x) a_j(x) b_k(x).$$

It's easy to show that this is well an inner product. Therefore, the existence of $g(x)$ allowed us to provide our tangent space a metric structure (more precisely it's a prehilbertian space). This gives us a new class of differential manifolds :

Definition 2.10 (*Metric tensor and Riemannian Manifold*)

*The **inner product** $\langle ., . \rangle$ defined on the tangent space $T_x(M^p)$ is said to be a **Riemannian metric tensor**, or simply a metric tensor on M^p. A differential manifold endowed with a smooth metric tensor is said to be a **Riemannian manifold**.*

✧✧✧

2.4. Geodesic paths and geodesic distance

In this section, we use the metric structure we just provided to our tangent space to locally extend it to the differential manifold itself, by defining length of paths on it.
Consider a smooth path $x(t)$ on a Riemannian manifold M^p. Then, using eq. (1.3), we can write the tangent vector to the path at a time t as :

$$\dot{x}(t) = \sum_{i=1}^{p} \frac{dx_i}{dt}(t)\partial_i(x(t)),$$

where $x_i(t)$ is the i-th coordinate of $x(t)$. For any t, we define the norm of $\dot{x}(t)$ as the norm induced by the inner product $\langle ., . \rangle$:

$$v(t) = \|\dot{x}(t)\| = \sqrt{\langle \dot{x}(t), \dot{x}(t) \rangle},$$
$$\Leftrightarrow v(t) = \sqrt{\sum_{j=1}^{p}\sum_{k=1}^{p} g_{jk}(x(t))\dot{x_j}(t)\dot{x_k}(t)},$$

where $g_{jk}(x(t))$ is the value of the metric tensor at $x(t)$. Then, analogously with the Euclidean case, the length ds of the path from $x(t)$ to $x(t + dt)$ will be defined as :

$$ds = v(t)dt.$$

Therefore, the length of $x(.)$ from t_0 to t_1 is :

$$L = \int_{t_0}^{t_1} ds = \int_{t_0}^{t_1} v(t)dt. \tag{1.4}$$

Remark : metric tensor determine length of arcs and it's also itself determined by the arc length. In fact, it can be shown that if we are able to determine that if we are able to calculate ds for any increment of a smooth path from $x(t)$ to $x(t + dt)$ then there exists at most one metric tensor compatible with this definition.

Having defined lengths of arcs on a differential manifold, a natural problem would be the seek of shortest paths between two points of the manifold. This investigation leads us to the notion of **geodesics paths** which are a generalization of straight lines in the Euclidean space :

Definition 2.11 (*Geodesic path*)

*A **geodesic path**, or simply **geodesic**, is a smooth path $x(t)$ in a Riemannian manifold which is* locally *the shortest.*

► Geometric Characterization

Geometrically speaking, a geodesic is a path which can be broken up into pieces such that the paths connecting the endpoints of the pieces are all the shortest paths. Therefore, the property of being a geodesic can be investigated locally.

The geometric characterization above gives us good intuition about geodesics, but isn't very convenient in practice and we would prefer an analytic characterization, such as a differential equation verified by such a path. Tus, let $\gamma : [a, b] \to M^p$ be a geodesic

path. Using the definition 2.11, $\gamma(t)$ is locally the shortest path. This means we can find a small enough open subset $]t_0, t_1[$ of $[a, b]$ such that $\gamma(t)$ is the path minimizing the length eq. (1.4) among all possible smooth paths from t_0 to t_1. Therefore, looking for a geodesic path is equivalent to look for extrema of the following function :

$$L(\epsilon) = \int_{t_0}^{t_1} v_\epsilon(t)dt = \int_{t_0}^{t_1} F(t, g_\epsilon, \dot{g_\epsilon})dt,$$

with $g_\epsilon(t) = \gamma(t) + \epsilon\eta(t)$ for any smooth path $\eta(t)$ vanishing on the boundaries ($\eta(t_0) = \eta(t_1) = 0$) and $F(t, g_\epsilon, \dot{g_\epsilon}) = \sqrt{\langle \dot{g_\epsilon}(t), \dot{g_\epsilon}(t)\rangle}$. If we want $\gamma(t)$ to be the shortest path on $]t_0, t_1[$ then we must have :

$$\frac{dL}{d\epsilon}(\epsilon)\bigg|_{\epsilon=0} = 0, \text{for all smooth path } \eta(t) \text{ vanishing on the boundaries.} \tag{1.5}$$

The above condition is in fact the directional derivative (see definition 1.2) of L (considered as a function of γ) at γ in the direction η. If γ is well an extremum, then we should have all the directional derivatives null at this point.

This yields :

$$\begin{aligned}
&\int_{t_0}^{t_1} \frac{d}{d\epsilon}(F(t, g_\epsilon, \dot{g_\epsilon}))dt = 0,\\
\Leftrightarrow &\int_{t_0}^{t_1} \left(\frac{\partial g_\epsilon}{\partial \epsilon}\right)^T \frac{\partial F}{\partial g_\epsilon}(t, g_\epsilon, \dot{g_\epsilon}) + \left(\frac{\partial \dot{g_\epsilon}}{\partial \epsilon}\right)^T \frac{\partial F}{\partial \dot{g_\epsilon}}(t, g_\epsilon, \dot{g_\epsilon})dt = 0,\\
\Leftrightarrow &\int_{t_0}^{t_1} (\eta(t))^T \frac{\partial F}{\partial g_\epsilon}(t, g_\epsilon, \dot{g_\epsilon}) + (\dot{\eta}(t))^T \frac{\partial F}{\partial \dot{g_\epsilon}}(t, g_\epsilon, \dot{g_\epsilon})dt = 0.
\end{aligned}$$

Futhermore, when $\epsilon = 0$, we have $g_\epsilon(t) = \gamma(t)$ and $\dot{g_\epsilon}(t) = \dot{\gamma}(t)$, which gives us :

$$\int_{t_0}^{t_1} (\eta(t))^T \frac{\partial F}{\partial \gamma}(t, \gamma, \dot{\gamma}) + (\dot{\eta}(t))^T \frac{\partial F}{\partial \dot{\gamma}}(t, \gamma, \dot{\gamma})dt = 0.$$

As $\dot{\eta}(t) = \sum_{i=1}^{p} \frac{d\eta_i}{dt}(t)\partial_i(\eta(t))$ we can integrate by parts on each component which yields :

$$\begin{aligned}
&\int_{t_0}^{t_1} (\eta(t))^T \frac{\partial F}{\partial \gamma}(t, \gamma, \dot{\gamma}) + (\dot{\eta}(t))^T \frac{\partial F}{\partial \dot{\gamma}}(t, \gamma, \dot{\gamma})dt = 0,\\
\Leftrightarrow &\int_{t_0}^{t_1} (\eta(t))^T \frac{\partial F}{\partial \gamma}(t, \gamma, \dot{\gamma})dt + \left[(\eta(t))^T \frac{\partial F}{\partial \dot{\gamma}}(t, \gamma, \dot{\gamma})\right]_{t_0}^{t_1} - \int_{t_0}^{t_1} (\eta(t))^T \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{\gamma}}(t, \gamma, \dot{\gamma})\right)dt = 0,\\
\Leftrightarrow &\int_{t_0}^{t_1} (\eta(t))^T \left[\frac{\partial F}{\partial \gamma}(t, \gamma, \dot{\gamma}) - \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{\gamma}}(t, \gamma, \dot{\gamma})\right)\right]dt = 0,
\end{aligned}$$

as $\eta(t_0) = \eta(t_1) = 0$. This equality being verified for all path $\eta(t)$ vanishing on the boundaries we must have :

$$\frac{\partial F}{\partial \gamma}(t,\gamma,\dot{\gamma}) - \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{\gamma}}(t,\gamma,\dot{\gamma})\right) = 0,$$

which is equivalent to

$$\frac{\partial F}{\partial \gamma_i}(t,\gamma_i,\dot{\gamma}_i) - \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{\gamma}_i}(t,\gamma_i,\dot{\gamma}_i)\right) = 0, \quad \forall i = 1,\dots,p.$$

This provides us the following characterization :

► **Analytic Characterization (Euler-Lagrange equations)**

A smooth path $\gamma(t)$ is a geodesic provided the **Euler-Lagrange equations** are satisfied, namely that

$$\frac{\partial F}{\partial \gamma_i}(t,\gamma_i,\dot{\gamma}_i) - \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{\gamma}_i}(t,\gamma_i,\dot{\gamma}_i)\right) = 0, \quad \forall i = 1,\dots,p, \tag{1.6}$$

with $F(t,\gamma_i,\dot{\gamma}_i) = \|\dot{\gamma}(t)\|$.

The above Euler-Lagrange equations can be exploited to derive the following existence and uniqueness theorem on geodesics, which will be very useful later in the definition of the exponential map, providing local diffeomorphism between the tangent space and the manifold (see section 2.5).

Theorem 2.1 (*Existence and uniqueness of geodesics*)

Geodesics on a Riemannian manifold ***exist and are unique****. More precisely, for all* $x \in M^p$, $v \in T_x(M^p)$, *there exists a unique geodesic* $\gamma : I \to M$ *such that :*

$$\begin{cases} \gamma(0) = x \\ \dot{\gamma}(0) = v \end{cases},$$

with $0 \in I \subset \mathbb{R}$.

Proof (Idea) : Expliciting $F(t,\gamma_i,\dot{\gamma}_i)$ in eq. (1.6) we obtain, using Einstein summation convention :

$$\begin{aligned}
&\frac{\partial}{\partial \gamma_i}\left(\sqrt{g_{jk}\dot{\gamma}^j\dot{\gamma}^k}\right) - \frac{d}{dt}\left(\frac{\partial}{\partial \dot{\gamma}_i}\left(\sqrt{g_{jk}\dot{\gamma}^j\dot{\gamma}^k}\right)\right) = 0,\\
\Leftrightarrow &\frac{1}{2\|\dot{\gamma}\|}\frac{dg_{jk}}{d\gamma_i}\dot{\gamma}^j\dot{\gamma}^k - \frac{d}{dt}\left(\frac{1}{\|\dot{\gamma}\|}g_{ik}\dot{\gamma}^k\right) = 0,\\
\Leftrightarrow &\frac{1}{2}\frac{dg_{jk}}{d\gamma_i}\dot{\gamma}^j\dot{\gamma}^k - \frac{d}{dt}\left(g_{ik}\dot{\gamma}^k\right) = 0\\
\Leftrightarrow &\frac{1}{2}\frac{dg_{jk}}{d\gamma_i}\dot{\gamma}^j\dot{\gamma}^k - \left(\frac{dg_{ik}}{d\gamma_j}\dot{\gamma}^j\dot{\gamma}^k + g_{ik}\ddot{\gamma}^k\right) = 0,\\
\Leftrightarrow &g_{ik}\ddot{\gamma}^k = \left(\frac{1}{2}\frac{dg_{jk}}{d\gamma_i} - \frac{dg_{ik}}{d\gamma_j}\right)\dot{\gamma}^j\dot{\gamma}^k.
\end{aligned}$$

Thus the Euler-Lagrange equations (1.6) can be rewritten as a second order differential equation and the existence and uniqueness come from the Picard-Lindelöf theorem for solutions of differential equations with prescribed initial conditions. Moreover, the same theorem tells us that γ depends smoothly on both x and v. ■

Having define the concept of a geodesic path in a Riemannian manifold, we are now ready to define the concept of **geodesic distance** between two points in the manifold.

Definition 2.12 (*Geodesic distance*)

*Suppose that a Riemannian manifold M^p is **pathwise connected**, in the sense that for any two points x, y M^p there exists a smooth path $x(t)$ such that $x(t_0) = x$ and $x(t_1) = y$.*
*We define the **geodesic distance** from x to y to be the length of the shortest path from x to y.*

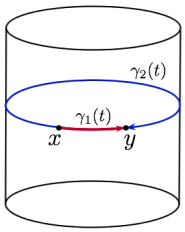


Figure 1.6.: *γ_1 and γ_2 are both geodesic paths from x to y but only γ_1 is the shortest path from x to y.*

With this definition, a pathwise connected Riemannian manifold M^p becomes a metric space.
It can be shown that the shortest path from x to y is a geodesic. However, the converse is not true and there exists geodesics between two points such that their length is strictly greater to the length of the shortest path (see fig. 1.6).

✧✧✧

2.5. The Exponential Map

Geodesics and the metric structure they induce on a path wise connected Riemannian manifold will allow us to define a diffeomorphic map between the manifold and the tangent plane. This map is called the **exponential map** and is defined as follows :

Definition 2.13 (*Exponential Map*)

*Let M^p be a Riemannian manifold, $x \in M^p$, $v \in T_x(M^p)$ and $\gamma_v(t)$ the unique geodesic such that $\gamma(0) = x$ and $\dot{\gamma}(0) = v$. Then, we define the **exponential map** as :*

$$Exp_x(v) = \gamma_v(1).$$

Terminology : The designation "exponential map" can be better understood by looking at one of the most famous property of the classic exponential function : $exp(x + y) = exp(x)exp(y)$. Thus, the exponential function is mapping a linear space with a non linear space, as the exponential map is mapping the tangent space with the manifold.

Thanks to theorem 2.1 the exponential map is well defined, as we have existence and uniqueness of the geodesic $\gamma_v(t)$. Intuitively, given $x \in M^p$ and $v \in T_x(M^p)$, the exponential map runs along a geodesic with constant velocity v, for one unit of time. Thus, the distance traveled on the geodesic depends on the velocity v, and we have the

following property :

$$\text{Exp}_x(tv) = \gamma_{tv}(1) = \gamma_v(t),$$

for $t \in I \subset \mathbb{R}$, such that $\gamma_v(t)$ is defined on I. If $I = \mathbb{R}$, then the manifold is said **geodesically complete** (which means we can maximally extend each geodesic to $\mathbb{R}$). Moreover, we have the following property on the exponential map :

Proposition 2.2 (*Local diffeomorphism*)

*For any $x \in M^p$ there exist a neighborhood U of 0 in $T_x(M^p)$ and a neighborhood V of x in M^p such that $Exp_x : U \to V$ is a **diffeomorphism**.*

Proof : The exponential map is mapping the prescribed initial conditions $(x, v) \in M^p \times T_x(M^p)$ with the corresponding solution $\gamma_v(t)$ of the Euler-Lagrange equations 1.6, which can be re-expressed as second order differential equations as seen in the proof of theorem 2.1. Thus, the Picard-Lindelöf theorem ensures that the solution $\gamma_v(t)$ smoothly depends on the data (x, v) and therefore the exponential map is a smooth map.

As $T_x(M^p)$ is a linear space, we can identify $T_0(T_x(M^p))$ with $T_x(M^p)$. Then, for any $v \in T_0(T_x(M^p)) = T_x(M^p)$, we have :

$$d(\text{Exp}_x)_0(v) = \frac{d}{dt}\left(\text{Exp}_x(tv)\right)\Big|_{t=0} = \left(\frac{d}{dt}\gamma_v(t)\right)\Big|_{t=0} \overset{(1.3)}{=} \dot{\gamma}_v(0) = v.$$

Therefore, we have $d(\text{Exp}_x)_0 = Id_x$, which is a bicontinuous isomorphism. Then, using the local inverse theorem, there exists an open neighborhood U of $0 \in T_x(M^p)$ and a neighborhood V of $x \in M^p$ such that $\text{Exp}_x : U \to V$, is a diffeomorphism. ■

This local diffeomorphism between the tangent space and the manifold allow us to define the logarithmic map, as the local inverse of the exponential map :

Definition 2.14 (*Logarithmic Map*)

*Let $U \subset T_x(M^p)$ and $V \subset M^p$ the biggest neighborhood such that $Exp_x : U \to V$ defines a diffeomorphism (the radius of U is called the **injectivity radius**). Then, we define the **logarithmic map** as the inverse of the exponential map :*

$$Log_x : \begin{cases} V & \to U, \\ y & \mapsto {Exp_x}^{-1}(y) = v, \end{cases}$$

with $\gamma_v(t)$ the unique geodesic passing through x at $t = 0$ such that $\gamma_v(1) = y$ and $\dot{\gamma}(0) = v$.

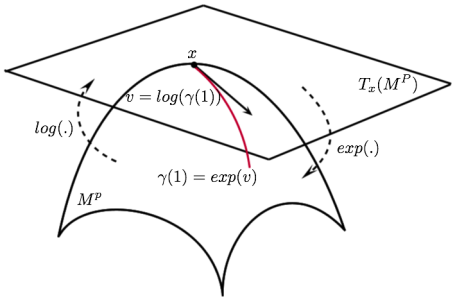


Figure 1.7.: *The exponential map allow us to locally carry the Euclidean structure of the tangent plane to the differential manifold.*

The logarithmic map provides a very efficient way to compute geodesic distance. In fact, if we assume that geodesics are parametrized in such a way that they have constant velocity[7], then we have :

[7] which we can always achieve through the arc-lenght parametrization

$$\forall x \in M^p, \forall v \in T_x(M^p), \quad d(x, \text{Exp}_x(v)) = \|v\|,$$

with $\|\cdot\|$ the norm on the tangent plane induced by the metric tensor.
Thus, $\forall x, y \in U$, with U as in definition definition 2.14 we have :

$$d(x, y) = \|\text{Log}_x(y)\|. \tag{1.7}$$

This result shows that the metric structure provided to the manifold by the geodesic distance is directly inherited from the tangent space metric structure.

CHAPTER

2 Nonlinear Extension of Principal Components Analysis

In this chapter, we first recall the theory of principal components analysis, a powerful statistical tool aiming to de-correlate and re-express data sets in an optimal way. Then, we investigate a possible extension of this tool to nonlinear situation, which will provide us an efficient way to compare shapes one another.

1. Introduction to the Principal Components Analysis

The origins of **principal components analysis (PCA)** lies in multivariate data analysis, where it became crucial to provide the analyst efficient ways to deal with huge amounts of data varying among multiple possibly correlated variables. To this end, the principal component analysis seeks to de-correlate the original data by finding independent directions, called principal components, in which the variance is maximized. By doing so, principal components analysis re-express the data as a sum of uncorrelated components. This is exactly the same philosophy as Fourier analysis[1] , which aims to represent periodic functions as the sum of simpler trigonometric functions. Such a decomposition is (among other things) widely used in signal processing, as it can help to identify main frequencies composing the signal. By focusing on this main frequencies, the analyst removes possible noise and therefore reduces the dimensionality of his problem, simplifying the analyze (see the Fourirer analysis of the FTSE index, fig. 2.1).
The same is true for PCA : by projecting the data on the main directions maximizing the variance, one may reduce the dimension of the problem, interpreting the data with less but more relevant variables (the principal components). Then, identifying trends, patterns or outliers in the data can be performed far more easily than with the original variables. Therefore, PCA is a purely descriptive technique, allowing us to perform efficient exploratory analysis of large data sets. PCA also has a wide range of other possible applications, in particular in linear regression or in image compressing.
The introduction on principal components analysis that we present here is inspired from the work of Joliffe in [4].

[1] in fact, the philosophy of PCA and Fourier analysis are so similar, that one could even wonder why performing a PCA rather than a Fourier transform of the data. Putting aside the technical differences of the two approaches, there still remains a fundamental difference between PCA and Fourier analysis. In fact, in some situation, it's embarrassingly weird to perform a Fourier analysis. Let's take the example of cars features data. One could number each features and then take the Fourier transform of the numbered features but this just seems crazy ! There is no legitimate reason of performing such an analysis. We will see that there is no such embarrassment for PCA.

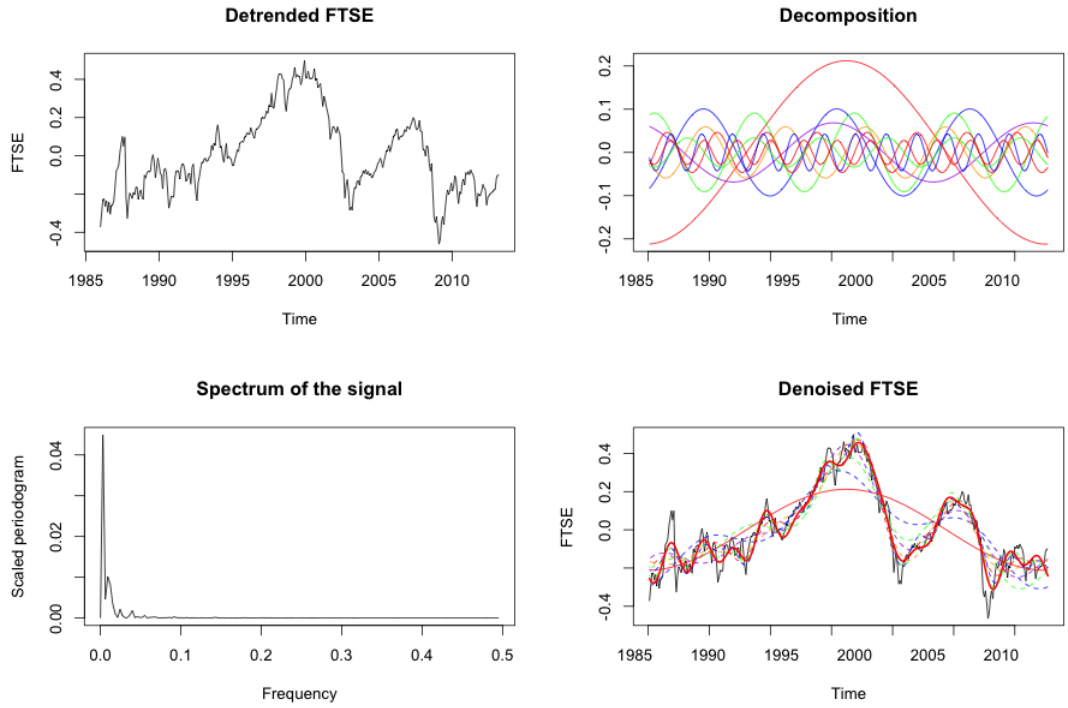


Figure 2.1.: *Fourier analysis of the Financial Times Stock Exchange (FTSE) index from 1985 to nowadays. The spectral analysis of the signal allow us to perform a decomposition of the signal in main frequencies (here we only took in considerations the first nine greatest frequencies). By summing the corresponding trigonometric functions, we obtain a de-noised version of the original signal, allowing us a better analysis (for example, we can easily detect financial crisis).*

1.1. Principal Components for a known covariance matrix Σ

Suppose we have a vector of random variables $X = (X_1, \dots, X_p) \in \mathbb{R}^p$ with $\mathbb{E}X = 0$ (without loss of generality) and known covariance matrix Σ. As we have seen previously the aim of principal components analysis is to find a new basis that re-express the random vector *optimally*. Algebraically speaking, we are seeking for a matrix of a change of basis P such that :

$$Y = PX,$$

with Y the random vector expressed in the new basis. Then, the rows of P are simply the principal components.

To find these principal components, we now need to address the issue of what this new basis should be, indeed what is the best way to re-express the random vector X. The principal component analysis pretends that such an optimal basis can be obtained by reducing the correlation between random variables and maximizing individual variances. Let see how we could achieve such a thing.

We start by defining the first component w_1 as :

$$w_1 = \underset{\|w\|=1}{\operatorname{argmax}} \mathrm{Var}(w^T X) = \underset{\|w\|=1}{\operatorname{argmax}} w^T \mathrm{Var}(X) w = \underset{\|w\|=1}{\operatorname{argmax}} w^T \Sigma w.$$

We additionally imposed $\|w\| = 1$ in order to obtain unicity of the basis. A simple way to obtain de-correlated variables in the new basis is to construct an orthogonal basis : this ensures the new variables to be linearly independent, an therefore with null covariance (as the covariance is only catching *linear dependencies* between two variables). Thus, we subtract to X all linear dependency with the first component :

$$\hat{X}_2 = X - ({w_1}^T X) w_1,$$

and w_2 will be defined as :

$$w_2 = \underset{\|w\|=1}{\operatorname{argmax}} \mathrm{Var}(w^T \hat{X}_2). \tag{2.1}$$

As $\hat{X}_2 \in \mathrm{span}(w_1)^\perp$ we necessarily have $w_2 \in \mathrm{span}(w_1)^\perp$ (as w_2 is maximizing eq. (2.1)) and therefore the basis we are building is indeed orthogonal. We can extend this procedure to all principal components :

Definition 1.1 (*Principal Components*)

Let $X \in \mathbb{R}^p$ a random vector with known covariance matrix Σ. We define the p ***principal components*** *as :*

- *First principal component :*

$$w_1 = \underset{\|w\|=1}{\operatorname{argmax}} \mathrm{Var}(w^T X).$$

- *k-th principal component, $2 \leq k \leq p$:*

$$w_k = \underset{\|w\|=1}{\operatorname{argmax}} \mathrm{Var}(w^T \hat{X}_k),$$

with $\hat{X}_k = X - \sum_{i=1}^{k-1} ({w_i}^T X) w_i$.

Thus, our matrix of change of basis P will be $P = \begin{pmatrix} {w_1}^T \\ \vdots \\ {w_p}^T \end{pmatrix}$. As the basis $\{w_1, \ldots, w_p\}$ is orthonormal, P is an orthogonal matrix. The expression of X in the new basis is :

$$Y = PX = \sum_{i=1}^{p} ({w_i}^T X) w_i.$$

The covariance matrix Λ in the new basis is given by :

$$\Lambda = P \Sigma P^T,$$

as $P^{-1} = P^T$. Therefore, Σ and Λ are similar. Moreover, in the new basis all the random variables are linearly independent (by construction), and therefore the covariance matrix Λ is diagonal. By the spectral theorem[2], there exists a unique orthonormal basis in which Σ is diagonal, and thus principal components are simply the eigenvectors of the matrix Σ, sorted in order of decreasing eigenvalues.

Thus, to compute principal components, one should simply find an orthonormal basis in which Σ is diagonal, and sort the eigenvectors in order of decreasing eigenvalues. Existence and unicity of such a basis are guaranteed by the spectral theorem applied to the covariance matrix Σ, which is indeed a symmetric matrix.

The following property provides a better understanding of the above statement, as it reformulate the problem in a geometrical point of view :

Proposition 1.1 (*Ellipsoids and Principal Components*)

Let X be a random vector of known covariance matrix Σ. Consider the family of p-dimensional ellipsoids :

$$X^T \Sigma^{-1} X = c, \tag{2.2}$$

with c a constant. Then, the principal components define the directions of the principal axes of these ellipsoids.

Proof : the expression of X in the basis defined by the principal components is $Y = PX$, with P defined as before. As P is orthogonal, we have $X = P^T Y$. Substituting in eq. (2.2) we obtain :

$$\begin{aligned}
&(P^T Y)^T \Sigma^{-1} (P^T Y) = c, \\
\Longleftrightarrow &Y^T (P \Sigma P^T)^{-1} Y = c, \\
\Longleftrightarrow &Y^T \Lambda^{-1} Y = c, \\
\Longleftrightarrow &\sum_{i=1}^{p} \frac{{y_i}^2}{\lambda_i} = c, \\
\Longleftrightarrow &\sum_{i=1}^{p} \frac{{y_i}^2}{\left(\sqrt{c\lambda_i}\right)^2} = 1,
\end{aligned} \tag{2.3}$$

and eq. (2.3) is the equation for an ellipsoid referred to its principal axes. Equation (2.3) also implies that the half-lengths of the principal axes are, from the greatest to the smallest axis, $\sqrt{c\lambda_1}, \ldots, \sqrt{c\lambda_p}$, if $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. ■

One way to understand this property, is to consider the random vector X as a "distortion" of the Euclidean space $\mathbb{R}^p$: all the points haven't got the same probability, and therefore, if we draw a series of observations from this random vector, the spread of the resulting scatter plot won't be the same in all directions. However, this spread is entirely determined by the covariance matrix Σ, which contains the average information about the dispersion[3] and joint dispersion of each random variable composing X. Thus,

[2] that we can invoke as the covariance matrix Σ is symmetric positive semi-definite .

[3] in all that follows, dispersion should be understood as dispersion in term of second order momentum (variance)

one could be interested in creating a measurement of dispersion, using the information contained in Σ. But how could we achieve such a thing ? As usual, let's get back to the well known univariate case, when X is simply a random variable in $\mathbb{R}$ with null mean and variance σ^2. Then, given an observation x of this random variable, one possible way to measure the dispersion of this observation could be to evaluate the quotient x^2/σ^2. Then, if $x^2/\sigma^2 \geq 1$, we would say that the observation is more dispersive than the average dispersion σ^2, and respectively less dispersive if $x^2/\sigma^2 \leq 1$.
Analogously with the univariate case, we can now generalize this measurement to multivariate statistics, through a quadratic form defined on $\mathbb{R}^p$, using the inverse of the covariance matrix $\Sigma : |w| = w^T\Sigma^{-1}w, \forall w \in \mathbb{R}^p$. Then, all the $w \in \mathbb{R}^p$ on the ellipsoids $w^T\Sigma^{-1}w = cste$ will have the same relative dispersion around the mean. Through this measure of dispersion, we can quantify the distortion of the space $\mathbb{R}^p$ induced by the distribution of X : the ellipsoids will be shaped according to this distortion (see fig. 2.2), and as stated by proposition 1.1, the greatest axis of the ellipsoids will correspond to the first principal component, the second greatest principal axis to the second principal component and so on...

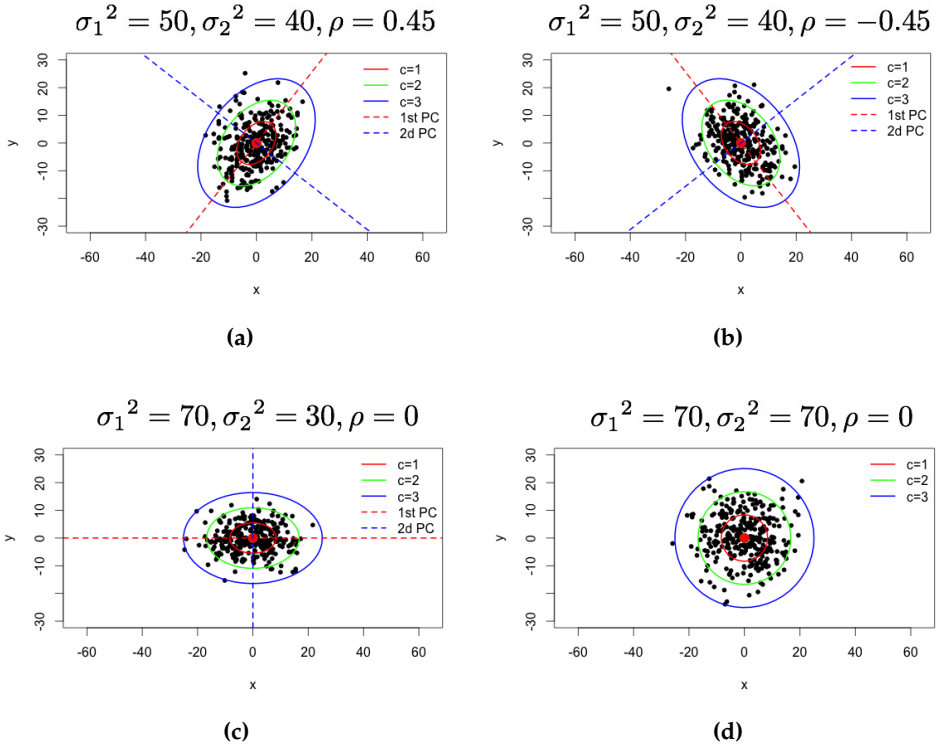


Figure 2.2.: *Illustration of the concept of dispersion ellipsoids for a bivariate normal distribution with mean* $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ *and covariance matrix* $\Sigma = \begin{pmatrix} {\sigma_1}^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & {\sigma_2}^2 \end{pmatrix}$. *We drawn 300 points from this bivariate normal distribution for different values of* σ_1,σ_2 *and* ρ *and plotted three ellipsoids for* $c = 1, 2, 3$ *and the first and second principal components. We can see that the shape of the ellipsoids are determined by the covariance matrix: the width depends on* σ_1, *the height depends on* σ_2 *and the inclination on the parameter* ρ *(the correlation).*

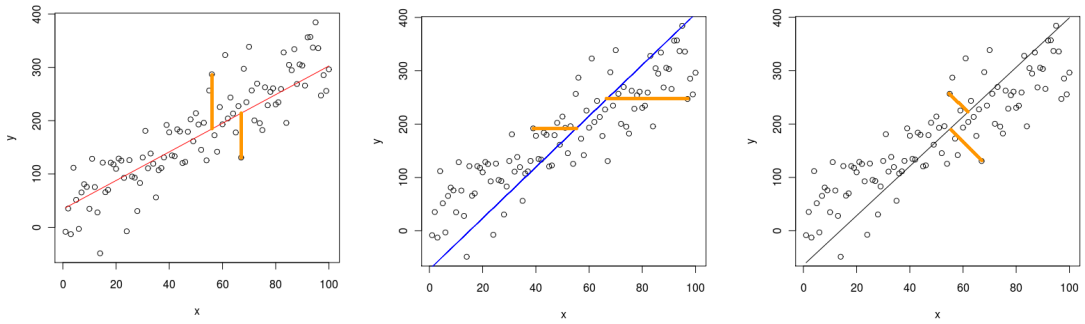
► PCA vs Linear Regression

One could wonder how different is PCA from linear regression. Despite the fact that they obviously don't have the same purpose, they appear to be *conceptually* different : while linear regression is trying to explain one variable through multiple explicative variables, PCA doesn't impose such an asymmetry among variables, and tries to optimally re-express the whole data. However, when we only have two variables, one could wonder if the direction of the regression line could be the same as the one provided by the first component. We investigated this issue on fig. 2.3. To understand the results, we must remember the definition of the first component :

$$w_1 = \underset{\|w\|=1}{\operatorname{argmax}} \operatorname{Var}(w^T X) = \underset{\|w\|=1}{\operatorname{argmax}} \mathbb{E}\left[(w^T X)^2\right] - \underbrace{\left(\mathbb{E}\left[w^T X\right]\right)^2}_{=0} = \underset{\|w\|=1}{\operatorname{argmax}} \mathbb{E}\left[(w^T X)^2\right],$$

as we imposed $\mathbb{E}X = 0$. From this expression, we can deduce that PCA is minimizing the sum of square of *orthogonal errors* to the first component line[a] (see fig. 2.3c). On the opposite, we know that for Gaussian linear models the best estimator of β is the least square estimator which itself minimizes the square *vertical distances* to the regression line (see fig. 2.3a and fig. 2.3b). Therefore, except for some rare case, there is no theoretical reason that the first component and the line regression have same direction.

[a] we can show that this is also true for all principal components.



(a) *Linear regression $y \sim x$. The sum of square of vertical distances to the regression line are minimized.*

(b) *Linear regression $x \sim y$. This time, y is the explanatory variables, so the horizontal distances to the regression line are minimized.*

(c) *First principal component direction. The sum of square of orthogonal errors are minimized.*

Figure 2.3.: *Differences between linear regression and PCA. In this example, x and y are linked by the following relationship : $y = 20 + 3x + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 60)$. We then drawn a hundred values of y for x varying in $[1, 100]$, and performed linear regressions $y \sim x$, $x \sim y$ and PCA.*

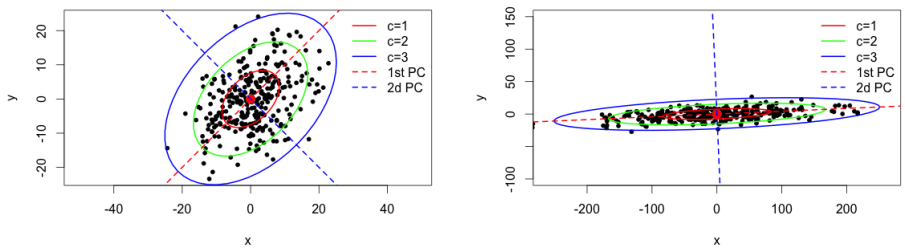
✧✧✧

1.2. Principal Components Analysis using correlation matrix

The above derivation of principal components is based on the eigenvectors and eigenvalues of the covariance matrix Σ. However, this approach has some limitation, the main one being its *huge sensitivity* to the units of measurement used for each elements of the vector X, as illustrated in the following artificial example. Let say we only have two variables x and y, x being a length, which can be measured either in centimeters or in millimeters. The second variable y is a weight, in grams for example. The covariance matrices in the two cases are respectively[4] :

$$\Sigma_1 = \begin{pmatrix} 70 & 30 \\ 30 & 70 \end{pmatrix}, \quad \text{and} \quad \Sigma_2 = \begin{pmatrix} 7000 & 300 \\ 300 & 70 \end{pmatrix}.$$

In the first situation the first principal component is $0.707x+0.707y$, with approximatively 71.43 % of the total variation [5]. In the second situation the change of unit results in a radically different first principal component : $0.999x + 0.043y$, with 99.19% of the total variation ! While in the first situation the first PC gives equal weight to x and y, in the second situation the first PC is almost entirely dominated by the variable x. This can also be seen through the shape of the respective dispersion ellipsoids (see fig. 2.4). This



(a) *PCs and dispersion ellipsoids for the covariance matrix Σ_1. The two variables x and y have the same weight in the first PC, and the dispersion ellipsoid is equally shaped.*

(b) *PCs and dispersion ellipsoids for the covariance matrix Σ_2. Here, the variable x is dominating y in the first PC, and the dispersion ellipsoid is stretched in this direction.*

Figure 2.4.: *Sensitivity of the PCs and the dispersion ellipsoids to the units of measurement, in the context of the above example. For the simulation, we supposed a bivariate normal distribution for the joint distribution of x and y.*

example illustrates the problematic behavior of PCA in case of wide differences between the variances of each variable. One remedy to this huge sensitivity could be to use the correlation matrix Σ^* instead of the covariance matrix Σ. The **correlation matrix** Σ^* is

[4] using the fact that $\text{Var}(10x) = 100 \times \text{Var}(x)$ and $\text{cov}(10x, y) = 10 \times \text{cov}(x, y)$.

[5] to compute the percentage of total variation accounted to the first component, one must evaluate the ratio : $\lambda_1/(\lambda_1 + \lambda_2)$, with λ_i the eigenvalues of Σ, sorted in decreasing order.

the covariance matrix of a *standardized* version of the vector X:

$$X^* = \begin{pmatrix} x_1/\sigma_1 \\ \vdots \\ x_p/\sigma_p \end{pmatrix},$$

with x_i the i-th random variable composing the random vector X, and σ_i^2 its corresponding variance, i-th term of the diagonal of Σ. Then, all the standardized random variables of X^* are dimensionless, preventing us from the possible issues of scale dependency of the PCA. However, this normalization is arbitrary, and one could decide to re-weight the variables by weights w_i different from σ_i, corresponding to the analyst intuition of the relative importance of each variables. But in practice, it is rare that obvious re-weighting suggests itself, explaining our preference for the correlation matrix.

Finally, it must be noted that there is no simple relationship between the PCs obtained by the covariance matrix and the correlation matrix : in fact, it can be shown that the PCs are preserved under orthogonal transformation, and the normalization of X in X^* isn't an orthogonal transformation.

✧✧✧

1.3. Principal Components for an unknown covariance matrix (Sample Principal Components)

The derivation of principal components is based on the assumed knowledge of the covariance (or correlation) matrix of the random vector X. But in practice, this assumption is pretty unrealistic, and we could have to deal with data from an unknown distribution and therefore unknown theoretical covariance matrix. Thus, we would have to use a *proxy* of the covariance matrix, called the **sample covariance matrix**.

Let $X \in \mathbb{R}^p$ be a random vector and $\{(x_{1,i}, \dots, x_{p,i})^T : i = 1, \dots n\}$ n observations from this random vector. We store these observations in a so-called **design matrix** $\tilde{W} \in \mathbb{R}^{p \times n}$:

$$\tilde{W} = \begin{pmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{p,1} & \dots & x_{p,n} \end{pmatrix} = \begin{pmatrix} \tilde{W}_1 \\ \vdots \\ \tilde{W}_p \end{pmatrix},$$

where $\tilde{W}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,n}) \in \mathbb{R}^n$, is a vector composed by the n occurrences of the random variable X_i, for $i = 1, \dots, p$. As we have seen before, in order to perform a proper PCA, we must have $\mathbb{E}X = 0$. Here, as we don't have direct access to the mean, we center the design matrix in order to ensure that all sample means of the new variables W_i be null :

$$W = \begin{pmatrix} \tilde{W}_1 - \tilde{\mu}_1 \\ \vdots \\ \tilde{W}_p - \tilde{\mu}_p \end{pmatrix} = \begin{pmatrix} W_1 \\ \vdots \\ W_p \end{pmatrix},$$

with $\tilde{\mu}_i = \frac{1}{n}\sum_{k=1}^{n} x_{i,k}$, for $i = 1, \ldots, p$. Then, the **sample variance**[6] of the series W_i will be :

$$\hat{\sigma}_i^2 = \frac{1}{n-1} W_i W_i^T = \frac{1}{n-1} \sum_{k=1}^{n} {w_{i,k}}^2.$$

Similarly, the **sample covariance** between W_i and W_j is :

$$\hat{\sigma}_{ij} = \frac{1}{n-1} W_i W_j^T = \frac{1}{n-1} \sum_{k=1}^{n} w_{i,k} w_{j,k}.$$

More generally, we can define a sample covariance matrix gathering all sample covariances and variances :

Definition 1.2 (*Sample Covariance matrix*)

*Let $X \in \mathbb{R}^p$ be a random vector an $W \in \mathbb{R}^{p \times n}$ the centered design matrix defined above of n observations from X. Then we define the **sample covariance matrix** $\hat{\Sigma}_W \in \mathbb{R}^{p \times p}$ as :*

$$\hat{\Sigma}_W = \frac{1}{n-1} W W^T = \frac{1}{n-1} \begin{pmatrix} W_1 W_1^T & \cdots & W_1 W_p^T \\ \vdots & \ddots & \vdots \\ W_p W_1^T & \cdots & W_p W_p^T \end{pmatrix} = \begin{pmatrix} \hat{\sigma}_1^2 & \cdots & \hat{\sigma}_{1p} \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_{p1} & \cdots & \hat{\sigma}_p^2 \end{pmatrix}.$$

This unbiased estimator of the covariance matrix will help us to re-express the data optimally, in PCA fashion. This time, the aim is to find a matrix P such that :

$$V = PW,$$

with V the design matrix of the re-expressed data, with null sample correlations and maximized individual sample variances. Having noticed that

$$\hat{\Sigma}_V = (PW)(PW)^T = PWW^T P^T = P\hat{\Sigma}_W P^T,$$

we see that we can still find principal components through the spectral decomposition of $\hat{\Sigma}_W$. Then, the procedure is exactly the same as discussed in the previous sections. The principal components obtained trough PCA on sample covariance matrix are called **sample principal components**. Moreover, all the properties previously derived for a known covariance matrix remains true when using the sample covariance matrix to perform PCA.

Finally, to avoid issues of sensitivity to the units of measurement developed in section 1.2, one can also define a **sample correlation matrix** by taking the sample covariance matrix of the standardized version of W :

$$W^* = \begin{pmatrix} W_1/\hat{\sigma}_1 \\ \vdots \\ W_p/\hat{\sigma}_p \end{pmatrix} = \begin{pmatrix} W_1^* \\ \vdots \\ W_p^* \end{pmatrix}.$$

[6] we choose here the unbiased estimator of the variance as we divide by $n-1$

1.4. Example : Web ranking of Universities in Mathematics

In this example, we ranked 23 top world universities in mathematics using PCA.
Each university has been evaluated in 19 different domains in mathematics (analysis, algebra, statistics, probability...), using the Google data base. The score in each domain is simply the number of pages referencing both the considered domain and university. A university is therefore associated with a vector in $\mathbb{R}^{19}$, containing all those scores.
Considering this high number of criteria, a full and accurate comparison of the universities seems impossible. A solution to this issue would be to reduce the dimensionality of the problem, by agglomerating the scores into one scale. To do so, we perform PCA with unknown variance matrix as described in the previous section, and project the data onto the first PC, representing 47 % of total variability. Table 2.1 presents the ranking obtained this way.

	Universities	Score
1	MIT	2.67
2	Harvard	2.23
3	Columbia University	0.46
4	University College London	0.42
5	Johns Hopkins University	0.28
6	Cornell University	0.27
7	Princeton University	0.20
8	Stanford University	0.09
9	Brown University	0.01
10	University of Michigan	-0.12
11	Imperial College	-0.16
12	**ETH**	-0.17
13	California institute of technology	-0.23
14	University of Oxford	-0.23
15	**EPFL**	-0.25
16	Kyoto University	-0.27
17	University of Toronto	-0.30
18	Ecole normale superieure	-0.30
19	University of Cambridge	-0.31
20	University of California Los Angeles	-0.32
21	Mcgill university	-0.35
22	The university of Melbourne	-0.37
23	University of Hong Kong	-0.39

Table 2.1.: *Web ranking of top universities in Mathematics*

✧✧✧

2. Nonlinear Statistics on Manifolds

A primary goal of shape theory is to describe the variability of a population of geometric objects. An efficient tool to perform such an analysis would be the principal component analysis, described above (see section 1). However, this technique is linear in essence, as it involves various results of linear algebra. Thus, to perform PCA, the data must lie in an Euclidean space. Unfortunately, complex representations of shape do not fulfill such a requirement, forcing us to re-design the technique to nonlinear situations. More precisely, we will consider the data to lie in a Riemannian manifold and present the method of **principal geodesics analysis** to analyze its variability , in PCA fashion. The first step in generalizing statistical tools to manifolds is to define the notion of mean. The methodology presented here is based on the work of Fletcher in [5], that we modified and adapted to our purpose.

2.1. Means on Manifolds

As usual, to generalize means to manifolds, we will wisely borrow the knowledge we already have about the Euclidean case.

Given a set of points $x_1, \dots, x_N \in \mathbb{R}^d$, the mean $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ is the point minimizing the sum-of-squared Euclidean distance to the data :

$$\bar{x} = \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{N} \|x - x_i\|^2.$$

This is a widely known result from M-Estimation theory, which generalizes this approach by providing a whole family of means estimator depending on the considered distance. Since the manifold may not form a linear space, we may not be able to compute additive means. A possible solution to this issue could be the use of the above characterization as a natural extension of the mean to nonlinear situations. However, such a definition strongly depends on the definition of the distance, which should be consequently carefully chosen to obtain the most natural way of extending means to manifolds.
A naive approach would consist in embedding the manifold in a greater Euclidean space and use the inherited metric structure to define the distance. Such an approach leads to the so-called **extrinsic mean**.

The Extrinsic Mean

Given an embedding $\Phi : M^p \to \mathbb{R}^d$, we define :

Definition 2.1 (*Extrinsic mean*)

*For every set of points $x_1, \dots, x_N \in M^p$, we define the **extrinsic mean** as :*

$$\mu_\Phi = \underset{x \in M^p}{\operatorname{argmin}} \sum_{i=1}^{N} \|\Phi(x) - \Phi(x_i)\|^2, \tag{2.4}$$

with $\|\cdot\|$ the Euclidean norm on $\mathbb{R}^p$.

This is equivalent to computing the classic arithmetic mean of the embedded points $\Phi(x_i)$ and then project this mean onto the manifold M^p. To see this, one must define a projection mapping $\pi : \mathbb{R}^d \to M^p$ in the following way :

$$\pi(x) = \underset{y \in M^p}{\operatorname{argmin}} \|\Phi(y) - x\|^2, \quad \forall x \in \mathbb{R}^p,$$

and then we have :

$$\mu_\Phi = \pi\left(\frac{1}{N}\sum_{i=1}^{N}\Phi(x_i)\right),$$

which is indeed the projection of the arithmetic mean of the data onto the manifold. On fig. 2.5 we present one example of computation of the extrinsic mean, using the gradient descent algorithm to minimize the sum-of-square distance function (2.4). However, such a definition of the mean isn't totally satisfactory, as it requires the manifold to be embedded in a greater Euclidean space. Such a requirement is embarrassing as the manifold exists wether it is embedded in an Euclidean space or not, and so an *extrinsic* definition of the mean depending on an additional embedding of the manifold doesn't seem to be the most natural definition of the mean. This concern leads us to consider a more appropriate distance to define the mean, providing us an *intrinsic* definition of the mean.

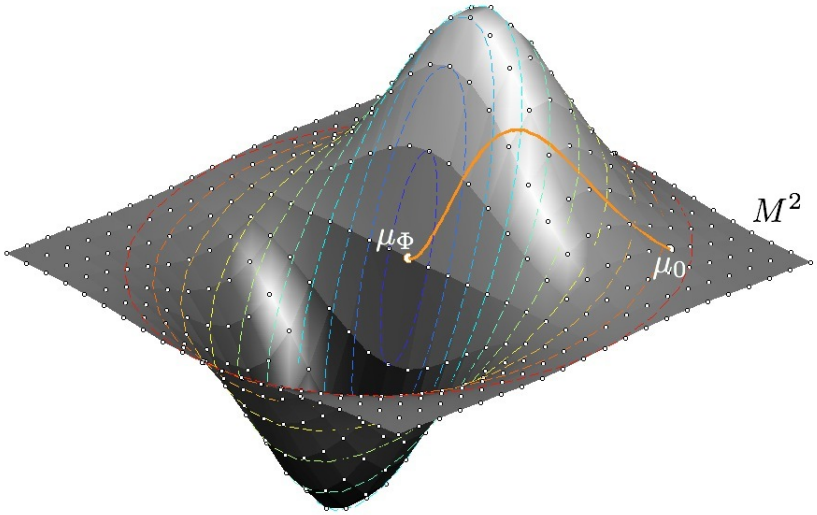


Figure 2.5.: ***Computing the extrinsic mean with the gradient descent algorithm.*** *In this example we computed the extrinsic mean μ_Φ of the withe dots on the manifold M^2 embedded in $\mathbb{R}^3$. The embedding is : $\Phi : M^2 \to \mathbb{R}^3, \quad u \mapsto \Phi(u) = \left(x, y, x \cdot e^{-(x^2+y^2)}\right)$. Starting from an arbitrary estimate of the extrinsic mean μ_0, we minimize the sum-of-square distance function (2.4), by taking successive steps (in orange on the graphic) in the negative gradient direction of the function at the current estimate of the mean. We additionally draw (dotted lines on the graphic) the equipotential levels of the objective function (from red to blue on the graphic as the intensity decreases).*

The Intrinsic Mean

To obtain a natural definition of means on Riemannian manifolds, one must exploit the natural distance provided by the existence of a metric tensor on it. Such a distance is the geodesic distance defined in 2.12, which depends only on the intrinsic geometry of the manifold.

Definition 2.2 (*Intrinsic Mean*)

*Let M^p be a pathwise connected Riemannian manifold and $d(\cdot,\cdot)$ the natural geodesic distance defined on it. Then, the **intrinsic mean** of a set of points $x_1,\dots,x_N \in M^p$ is defined by :*

$$\mu = \underset{x\in M^p}{\operatorname{argmin}} \sum_{i=1}^{N} d(x,x_i)^2.$$

Remark : Since the intrinsic mean is defined as an optimization problem, its existence and uniqueness aren't ensured. However, Kendall has shown that the intrinsic mean exists and is unique if the data is **well-localized** (i.e. localized in a sufficiently small neighborhood).

Computing the intrinsic mean is then equivalent to minimizing the sum-of-square distance function :

$$f(x) = \frac{1}{2N}\sum_{i=1}^{N} d(x,x_i)^2, \tag{2.5}$$

with $x_1,\dots,x_n \in M^p$ lying in a sufficiently small neighborhood such that existence and unicity of the solution to the problem is guaranteed.
To solve the above optimization problem, we will use the widely-known **gradient descent algorithm**. This algorithm, also known as the **steepest descent algorithm**, finds local minimum of a given function by taking successive steps in the negative gradient direction of the function at the current point (see fig. 2.5 for one application of the gradient descent algorithm to the computation of the extrinsic mean). In our case, it's possible to express the gradient of the function $f(x)$ through the logarithmic map (see definition 2.14).

Proposition 2.1 (*Expression of the gradient*)

Let $x \in M^p$ and $x_1,\dots,x_N \in M^p$ lie in a sufficiently small neighborhood U of x, such that the exponential map is a diffeomorphism on it. Then, the gradient of the sum-of-square distance function (2.5) *is :*

$$\nabla f(x) = -\frac{1}{N}\sum_{i=1}^{N} Log_x(x_i),$$

with $Log_x(\cdot)$ the logarithmic map at x.

Proof (Heuristic) : We won't discuss here the detailed proof of this result provided by Karcher in [6], as it implies some additional background concepts from manifolds theory, that we didn't

introduce in this report. Thus, we only provide here some intuition about the result. Then, we proceed heuristically in the following.

If $x_1, \ldots, x_N$ lie in a neighborhood U such that the exponential map is a diffeomorphism on it, then we can re-express the geodesic distance through the logarithmic map (see eq. (1.7)). We have, $\forall x, y \in U$:

$$d(x, y) = \|\text{Log}_x(y)\|.$$

This gives us a new expression of f, $\forall x \in M^p$:

$$f(x) = \frac{1}{2N} \sum_{i=1}^{N} \|\text{Log}_x(x_i)\|^2.$$

Let $p \in M^p$ be sufficiently close to x. Then, we can approximate the above expression as follows :

$$f(x) \simeq \frac{1}{2N} \sum_{i=1}^{N} \|\text{Log}_p(x) - \text{Log}_p(x_i)\|^2 = \frac{1}{2N} \sum_{i=1}^{N} \|\tilde{x} - \tilde{x}_i\|^2 = \tilde{f}(\tilde{x}), \tag{2.6}$$

with $\tilde{x} = \text{Log}_p(x)$ and $\tilde{x_i} = \text{Log}_p(x_i)$.

Then, as $T_p(M^p)$ is linearly isomorphic to $\mathbb{R}^p$, we know that the arithmetic mean $\tilde{\mu} \in T_p(M^p)$ is minimizing the right-hand side :

$$\tilde{\mu} = \frac{1}{N} \sum_{i=1}^{N} \text{Log}_p(x_i) = \frac{1}{N} \sum_{i=1}^{N} \tilde{x}_i = \underset{\tilde{x} \in \mathbb{R}^p}{\text{argmin}} \frac{1}{2N} \sum_{i=1}^{N} \|\tilde{x} - \tilde{x}_i\|^2.$$

Then, one could exploit the approximation (2.11) and use $\tilde{\mu}$ as a direction from p to the minimum of the function f. Then, to obtain $\nabla f(x)$, we take the opposite[7] of the parallel transport of $\tilde{\mu}$ along the minimal geodesic linking p and x.

$$\nabla f(x) = -\nabla_v \left(\frac{1}{N} \sum_{i=1}^{N} \text{Log}_p(x_i) \right),$$

with ∇_v denoting the parallel transport[8] of $\tilde{\mu}$ along the minimal geodesic linking x and p.

Finally, as p is close enough to x, we can make the following approximation :

$$\nabla f(x) = -\nabla_v \left(\frac{1}{N} \sum_{i=1}^{N} \text{Log}_p(x_i) \right) \simeq -\frac{1}{N} \sum_{i=1}^{N} \text{Log}_x(x_i). \quad \blacksquare$$

Then, given a current estimate μ_j for the extrinsic mean, the equation for updating the mean by taking one step in the negative direction is (see also fig. 2.6):

$$\mu_{j+1} = \text{Exp}_{\mu_j} \left(\frac{\tau}{N} \sum_{i=1}^{N} \text{Log}_{\mu_j}(x_i) \right),$$

[7] the vector $\tilde{\mu}$ is pointing in the direction of the minimum of the scalar field f, while the gradient is pointing to the greatest rate of increase of it, thus we must take the opposite of $\tilde{\mu}$.

[8] for more information on parallel transport of vectors along geodesics, see *affine connexion*.

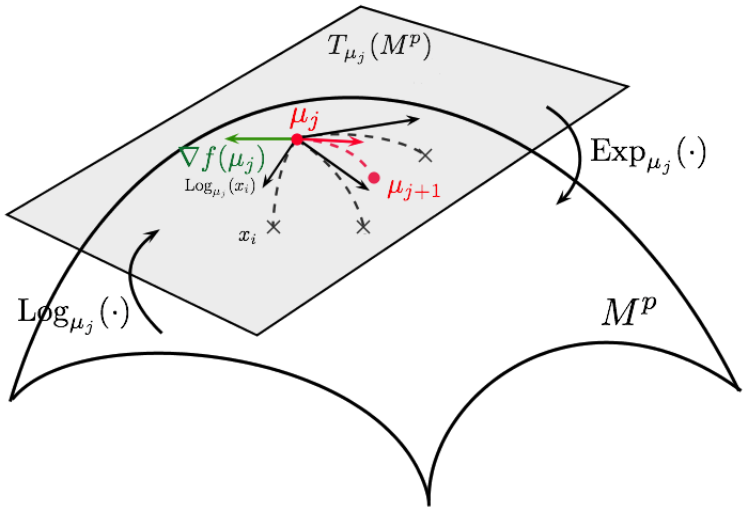


Figure 2.6.: ***One step of the gradient descent algorithm in the computation of the intrinsic mean.*** *Given a current estimate μ_j of the mean, the algorithm compute the gradient at this point and runs along the geodesic with opposed velocity for τ unit of time.*

where τ is the step size. Since the gradient descent algorithm only converges locally, one must choose carefully the initial estimate μ_0 for the intrinsic mean and the step size parameter τ. As the data is supposed to be well-localized, a reasonable choice for the initial estimate μ_0 is one of the data point, say x_1. The choice of τ is somewhat harder and depends on the manifold.
In summary, the algorithm to compute the intrinsic mean is the following :

Algorithm 1: Intrinsic mean

Data: A set of points $x_1, \dots, x_N \in M^p$, a step parameter τ and a precision ϵ.
Result: The intrinsic mean $\mu \in M^p$ of these points.
$\mu = x_1$;
while $\|\Delta\mu\| \geq \epsilon$ **do**
  $\Delta\mu = \frac{1}{N}\sum_{i=1}^{N} \mathrm{Log}_{\mu}(x_i)$;
  $\mu = \mathrm{Exp}_{\mu}(\tau \cdot \Delta\mu)$;
end

On the choice of τ : In some cases, it's possible to improve the convergence of the gradient descent algorithm by *optimally* choosing the step size parameter τ at each step. We present here an example of such a case and try to see if we could extend this approach to accelerate the computation of the intrinsic mean.

Let $x, b \in \mathbb{R}^p$ and $A \in \mathbb{R}^{p\times p}$. We want to solve the linear system $Ax = b$, using the gradient descent algorithm. We associate to this linear system the energy scalar field Φ:

$$\Phi(x) = \frac{1}{2}x^T Ax - x^T b. \tag{2.7}$$

Then, the gradient descent algorithm will indeed return the solution to the linear system, as

the minimum of the energy function (2.7) is such that : $\nabla\Phi(x) = Ax - b = 0$. Given a current estimate $x^{(j)}$ of the solution, the equation for updating the estimate is :

$$x^{(j+1)} = x^{(j)} - \tau_j \nabla\Phi(x^{(j)}),$$

with τ_j the step size parameter, which this time isn't chosen uniformly on j and can vary at each step. More precisely, we wish to choose τ_j *optimally* at each step such that the function $\Phi(x^{(j+1)}; \tau_j)$ is minimized with respect to τ_j. Differenciating with respect to τ_j and equalling to 0 leads to :

$$\tau_j = \frac{\nabla\Phi(x^{(j)})^T \nabla\Phi(x^{(j)})}{\nabla\Phi(x^{(j)})^T A \nabla\Phi(x^{(j)})}.$$

In the same fashion, one could try to modify the algorithm 1 by optimally chosing the step size parameter τ at each step. In this case, the energy scalar function becomes :

$$f(x) = \frac{1}{2N} \sum_{i=1}^{N} d(x, x_i)^2,$$

and the equation for updating the mean is :

$$\mu_{j+1} = \text{Exp}_{\mu_j} \left(\frac{\tau_j}{N} \sum_{i=1}^{N} \text{Log}_{\mu_j}(x_i) \right).$$

Then, the optimal τ_j is such that :

$$\begin{aligned} &\frac{d}{d\tau} \left(f(\mu_{j+1}) \right) \Big|_{\tau=\tau_j} = 0, \\ \iff &\nabla f \left(\text{Exp}_{\mu_j} \left(\frac{\tau}{N} \sum_{i=1}^{N} \|\text{Log}_{\mu_j}(x_i)\| \right) \right) \times \frac{d}{d\tau} \left[\text{Exp}_{\mu_j} \left(\frac{\tau}{N} \sum_{i=1}^{N} \|\text{Log}_{\mu_j}(x_i)\| \right) \right] \Bigg|_{\tau=\tau_j} = 0. \end{aligned}$$

If we call $v = \frac{1}{N} \sum_{i=1}^{N} \|\text{Log}_{\mu_j}(x_i)\|$, we obtain, using proposition 2.1 :

$$\begin{aligned} &\left(-\frac{1}{N} \sum_{i=1}^{N} \text{Log}_{\text{Exp}_{\mu_j}(\tau v)}(x_i) \right) \times \frac{d}{d\tau} \left(\gamma_v(\tau) \right) \Big|_{\tau=\tau_j} = 0, \\ \iff &\left(-\frac{1}{N} \sum_{i=1}^{N} \text{Log}_{\text{Exp}_{\mu_j}(\tau_j v)}(x_i) \right) \times \dot{\gamma}_v(\tau_j) = 0. \end{aligned}$$

Thus, as geodesics are paths with constant velocity we have $\dot{\gamma}_v(\tau_j) = v \neq 0$, and then the optimal τ_j is such that :

$$-\frac{1}{N} \sum_{i=1}^{N} \text{Log}_{\text{Exp}_{\mu_j}(\tau_j v)}(x_i) = \nabla f(\text{Exp}_{\mu_j}(\tau_j v)) = \nabla f(\mu_{j+1}) = 0.$$

In other words, such an optimal τ_j would make you run along the geodesic with initial velocity v, until you reach a point $\mu_{j+1} = \text{Exp}_{\mu_j}(\tau_j v)$ such that the gradient is null at this point. But then, the algorithm 1 would converge in exactly one step ! In fact, at step $j+1$ we would have $\Delta\mu = 0 < \epsilon$

which would end the while loop. Thus, finding an optimal τ_j at each step is computationally as difficult as computing the intrinsic mean itself.

This explains our choice of an uniform step size parameter τ in algorithm 1. However, our incapacity of finding an optimal τ makes the choice of it difficult in general, and strongly dependent on the considered manifold. For example, if M^p is a linear space, a choice of $\tau = 1$ is equivalent to linear averaging and would make the algorithm converge in one step. Finally, Buss and Fillmore have shown in [7] that for data lying on spheres, a value of $\tau = 1$ is sufficient.

✧✧✧

2.2. Variance on Manifolds

Another important concept from PCA that we need to extend into manifold setting is the concept of variance. It measures the expected dispersion of the data around the mean. For a real-valued random variable x with mean μ we have :

$$\sigma^2 = \mathbb{E}\left[(x-\mu)^2\right],$$

which naturally extend to the covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ for a real-valued random vector $x \in \mathbb{R}^p$:

$$\Sigma = \mathbb{E}\left[(x-\mu)(x-\mu)^T\right]. \tag{2.8}$$

Obviously, this definition is not valid for general manifolds since vector space operations do not exist for such spaces. However, this lack of linear structure can be compensated by the existence of a metric structure on general Riemannian manifolds. Thus, one idea might be to re-express the above expression (2.8) only in term of metric, which would ensure a better tractability of the concept to Riemannian manifolds.
To achieve it, let's get back to the one-dimensional case and see if it can give us some insight about this problem. If $x \in \mathbb{R}$ is a random variable, then we can show that the variance can be re-expressed as :

$$\text{Var}(x) = \mathbb{E}[x^2] - \left(\mathbb{E}[x]\right)^2.$$

Then, for $x \in \mathbb{R}^p$ one can show that the analog formula is :

$$\text{trace}\left(\text{Var}(x)\right) = \mathbb{E}\left[x^T x\right] - \mathbb{E}[x]^T\mathbb{E}[x].$$

As, $\text{Var}(x-\mu) = \text{Var}(x)$, we obtain, if $\mathbb{E}[x] = \mu$:

$$\begin{aligned}
& \text{trace}\left(\text{Var}(x)\right) = \mathbb{E}\left[(x-\mu)^T(x-\mu)\right] - \overbrace{\mathbb{E}[x-\mu]^T\mathbb{E}[x-\mu]}^{=0}, \\
\iff\quad & \text{trace}\left(\text{Var}(x)\right) = \mathbb{E}\left[\|x-\mu\|^2\right], \\
\iff\quad & \text{trace}\left(\text{Var}(x)\right) = \mathbb{E}\left[d(x,\mu)^2\right],
\end{aligned} \tag{2.9}$$

with $d(\cdot,\cdot)$ the Euclidean distance.
The above expression seems really interesting for our purpose, as it links the trace of the variance with the distance function. But what exactly does eq. (2.9) captures about the dispersion of the data around the mean ? In fact, we went from a $p \times p$ covariance matrix

describing the spread of the data around the mean to the trace of this same matrix, which lies in $\mathbb{R}$. Surely we lost some information during this process but does it still remain enough information to legitimate the use of eq. (2.9) as an inspiration for the variance extension into the manifold setting ? To answer this question, one must remark that :

$$\text{trace}(\text{Var}(x)) = \sum_{i=1}^{p} \lambda_i,$$

with λ_i the eigenvalues of the covariance matrix, which, in the eigenvectors basis, represent the dispersion of the data in each direction. In this sense, we can say that eq. (2.9) is capturing the *total variation* of the data, as the sum of the eigenvalues of $\text{Var}(x)$.
Then, it seems legit to propose the following definition of the variance on manifolds, which has been first proposed by Fréchet :

Definition 2.3 (*Variance*)

*Let M^p be a pathwise connected Riemannian manifold and $d(\cdot,\cdot)$ the natural geodesic distance defined on it. Then, the **variance** of a random variable $x \in M^p$ with mean μ is defined by :*

$$\sigma^2 = \mathbb{E}\left[d(x,\mu)^2\right].$$

Remark : More generally, this definition could be used to define the variance of a random variable in any metric space.

Finally, given a set of points $x_1, \dots, x_N \in M^p$ we define the **sample variance** of the data as :

Definition 2.4 (*Sample variance*)

*Let M^p be a pathwise connected Riemannian manifold and $d(\cdot,\cdot)$ the natural geodesic distance defined on it. Then, for any set of points $x_1, \dots, x_N \in M^p$ well-localized we define the **sample variance** of the data as :*

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} d(\mu, x_i)^2 = \frac{1}{N}\sum_{i=1}^{N} \|Log_\mu(x_i)\|^2.$$

✧✧✧

2.3. Geodesic submanifolds

The next step in generalizing PCA to manifolds is to define the notion of linear subspace. In fact, PCA aims to optimally re-express the data as a sum of uncorrelated components, which are one-dimensional linear subspaces. Then, we need to find analogs of those one-dimensional linear subspaces into manifold setting.
Geodesic curves seem to be the perfect candidates for such a purpose. In fact, we know that, as locally shortest paths between two points, they are natural extensions of

straight lines in Euclidean spaces, and thus natural extensions of one-dimensional linear subspaces.
Moreover, geodesics are clearly preserving the metric structure of M^p provided by the geodesic distance. This is an important feature as the principal components of PCA are maximizing the variance, which we extended into manifold setting as the average square distance to the intrinsic mean.
More generally, we will call a **geodesic submanifold** a submanifold $H \subset M^p$ which preserves the Riemannian distance :

Definition 2.5 (*Geodesic Submanifolds*)

*Let M^p be a Riemannian manifold. A submanifold H of M^p is said to be **geodesic** at $x \in H$ if all geodesics of H passing through x are also geodesics of M^p.*

Then, for any points $(x, y) \in H \times M^p$ with H a geodesic submanifold, we will have, with $d_H(\cdot, \cdot)$ the Riemannian distance on H:

$$\begin{aligned} d_H(x,y) &= min\{\text{lengths of geodesics of } H \text{ linking } x \text{ and } y\} \\ &= min\{\text{lengths of geodesics of } M^p \text{ linking } x \text{ and } y\} \\ &= d(x,y), \end{aligned}$$

and the Riemannian distance $d(\cdot, \cdot)$ is indeed preserved.
Thus, geodesic submanifolds containing the intrinsic mean will be the generalization of the linear subspaces of PCA.

✧✧✧

2.4. Projection operator

In PCA the data is orthogonally projected onto linear subspaces. We define here an orthogonal projection operator for geodesic submanifolds and show how it may be efficiently approximated.
Inspired from a classic result of linear algebra on orthogonal projections[9] , we will define the projection of a point $x \in M^p$ onto a geodesic submanifold $H \subset M^p$ as the point on H minimizing the Riemannian distance from x to any point of H :

Definition 2.6 (*Projection operator*)

*Let M^p be a Riemannian manifold, $x \in M^p$ and $H \subset M^p$ a geodesic submanifold of M^p. Then, we define the **projection operator** $\pi_H : M^p \to H$ as :*

$$\forall x \in M^p, \quad \pi_H(x) = \underset{y \in H}{\operatorname{argmin}}\, d(x,y)^2, \tag{2.10}$$

with $d(\cdot, \cdot)$ the Riemannian distance.

[9] The Pythagorean theorem allow us to alternatively characterize the orthogonal projection of a point $x \in \mathbb{R}^p$ onto a linear subspace $V \subset \mathbb{R}^p$ as the point minimizing the distance from x to any point of V.

Remark : One more time, since projection of x onto H is defined as an optimization problem, existence and uniqueness of it are not guaranteed. However, under the supposition that the data is well-localized, we can show that problem (2.10) admits a unique solution.

To solve the above optimization problem, one could use a gradient descent algorithm, as previously done for the computation of the intrinsic mean. However, in this very particular case, we can save a lot of computation time by considering the approximation of the projection operator in the tangent plane of M^p.
Let $x \in M^p$, $H \subset M^p$ be a geodesic submanifold and $p \in H$. In a sufficiently small neighborhood of x and p, we can approximate (2.10) :

$$\begin{aligned}\pi_H(x) &= \underset{y\in H}{\operatorname{argmin}}\, d(x,y)^2 \\ &= \underset{y\in H}{\operatorname{argmin}}\, \|\mathrm{Log}_x(y)\|^2 \\ &\simeq \underset{y\in H}{\operatorname{argmin}}\, \|\mathrm{Log}_p(x) - \mathrm{Log}_p(y)\|^2. \end{aligned} \tag{2.11}$$

Notice that $\mathrm{Log}_p(y)$ is simply a vector in $T_p(H)$ and thus we can rewrite (2.11) in terms of tangent vectors :

$$\mathrm{Log}_p\left(\pi_H(x)\right) \simeq \underset{v\in T_p(H)}{\operatorname{argmin}}\, \|\mathrm{Log}_p(x) - v\|^2. \tag{2.12}$$

But then, as $T_p(H)$ is a linear space, we know that the solution of the right hand side of eq. (2.12) is simply the orthogonal projection of $\mathrm{Log}_p(x)$ onto $T_p(H)$. Then, if $v_1, \cdots, v_k$ is an orthogonal basis of $T_p(H)$, we have the following approximation formula :

$$\mathrm{Log}_p\left(\pi_H(x)\right) \simeq \sum_{i=1}^{k} \langle \mathrm{Log}_p(x), v_i \rangle. \tag{2.13}$$

✧✧✧

2.5. Principal Geodesics Analysis

We are now ready to define the **Pincipal Geodesics Analysis (PGA)**, analog of the principal components analysis for manifolds. Given a set of points $x_1, \cdots, x_N \in M^p$, we wish to find two sequences of geodesic submanifolds :

- A sequence of *nested* geodesic submanifolds $H_1 \subsetneq H_2 \subsetneq \cdots \subsetneq H_p = M^p$. Those nested geodesic submanifolds, called **principal geodesic submanifolds**, will help us to efficiently analyze the variability of the data, by looking at the projection of the data on each of them.

- A sequence of *one-dimensional* geodesic submanifolds $V_1, \cdots, V_p \subset M^p$.Those geodesic submanifolds, that we will call the **principal geodesic components**, must be interpreted as an attempt to re-express the data as a sequence of "independent" components, analogously with PCA (they are analog of principal components).

Let $x_1, \cdots, x_N \in M^p$ be well-localized such that we can compute the intrinsic mean of the data μ. Let $U \subset T_\mu(M^p)$ be a neighborhood of 0 such that projection is well-defined (existence and unicity) for all geodesic submanifolds of $\text{Exp}_\mu(U)$, and such that $\text{Exp}_\mu : U \to \text{Exp}_\mu(U)$ is a local diffeomorphism.
Then, the **first principal geodesic submanifold** H_1 must be such that the variance of the data projected on it is maximized. Using the diffeomorphism provided by the exponential map and the fact that the projection must be well-defined, we have $H_1 = \text{Exp}_\mu(\text{span}\{v_1\} \cap U)$, and then $v_1 \in T_\mu(M^p)$ must be chosen such that :

$$v_1 = \underset{\|v\|=1}{\text{argmax}} \frac{1}{N} \sum_{i=1}^{N} \|\text{Log}_\mu(\pi_H(x_i))\|^2,$$

with $H = \text{Exp}_\mu(\text{span}\{v\} \cap U)$. We additionally impose $|v\| = 1$ in order to obtain unicity of the solution. The **first principal geodesic component** will be defined as $V_1 = H_1 = \text{Exp}_\mu(\text{span}\{v_1\} \cap U)$. We postpone the interpretation of the principal geodesic components.
Then, given the first principal geodesic submanifold $H_1 = \text{Exp}_\mu(\text{span}\{v_1\} \cap U)$, the **second principal geodesic submanifold** $H_2 = \text{Exp}_\mu(\text{span}\{v_1, v_2\} \cap U)$ will be such that the tangent vector $v_2 \in T_\mu(M^p)$ is maximizing the variance of the data projected onto the bi-dimensional geodesic submanifold $H = \text{Exp}_\mu(\text{span}\{v_1, v\} \cap U)$:

$$v_2 = \underset{\|v\|=1}{\text{argmax}} \frac{1}{N} \sum_{i=1}^{N} \|\text{Log}_\mu(\pi_H(x_i))\|^2. \tag{2.14}$$

The **second principal geodesic component** is defined as : $V_2 = \text{Exp}_\mu(\text{span}\{v_2\} \cap U)$.
Roughly speaking, H_2 is the best bi-dimensional geodesic submanifold to describe the data, in the sense that it maximizes the projected variance of the data on it, and thus the neglected variability of the data about H_2 is minimized.
However, the choice of v_2 provided by (2.14) surely isn't unique as multiple vs can generate the same submanifold H_2. Then, to obtain unicity of the definition of v_2, we will additionally impose that $v_2 \in \text{span}\{v_1\}^\perp$. Even if this choice can seem rather arbitrary, we can legitimate it as an attempt to "de-correlate" the data, analogously with PCA. Of course, as the correlation is meant to capture *linear dependencies* between two variables, there is no proper way to generalize this notion for variables lying into a manifold, which doesn't even possess such a linear structure in general. However, we'd like to meet at least a weak form of de-correlation, by retrieving a property of two linearly independent variables $X, Y \in \mathbb{R}$:

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

Transposed into manifold setting, and applied to the particular case of the first two principal geodesic components V_1 and V_2, this property becomes :

$$\frac{1}{N} \sum_{i=1}^{N} \|\text{Log}_\mu(\pi_{H_2}(x_i))\|^2 = \frac{1}{N} \sum_{i=1}^{N} \|\text{Log}_\mu(\pi_{V_1}(x_i))\|^2 + \frac{1}{N} \sum_{i=1}^{N} \|\text{Log}_\mu(\pi_{V_2}(x_i))\|^2, \tag{2.15}$$

with $H_2 = \text{Exp}_\mu(\text{span}\{v_1, v_2\} \cap U)$, $V_1 = \text{Exp}_\mu(\text{span}\{v_1\} \cap U)$, and $V_2 = \text{Exp}_\mu(\text{span}\{v_2\} \cap U)$. If we could constrain v_2 in (2.14) to verify this condition, and extend this construction to all principal geodesic submanifolds, then we could say that PGA is analogously with PCA, "re-expressing the data as a sum of independent components": for every pairs of principal geodesic components $V_i = \text{Exp}_\mu(\text{span}\{v_i\} \cap U)$ and $V_k = \text{Exp}_\mu(\text{span}\{v_k\} \cap U)$, the variance of the data projected onto the bi-dimensional manifold $H_{ik} = \text{Exp}_\mu(\text{span}\{v_i, v_k\} \cap U)$ is equal to the sum of the variances of the data projected respectively onto the two geodesic submanifolds V_i and V_k. In this sense, the principal geodesic components are analog to the principal components in PCA (see fig. 2.7).

Let now show that $v_2 \in \text{span}\{v_1\}^\perp$ is a sufficient condition to approximately verify (2.15). If $v_2 \in \text{span}\{v_1\}^\perp$, then $\{v_1, v_2\}$ is an orthonormal basis of $T_\mu(H_2)$, and thus we can exploit approximation (2.13). Under this approximation we have :

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} \|\text{Log}_\mu(\pi_{H_2}(x_i))\|^2 &\simeq \frac{1}{N}\sum_{i=1}^{N} (\langle v_1, \text{Log}_\mu(x_i)\rangle^2 + \langle v_2, \text{Log}_\mu(x_i)\rangle^2) \\
&= \frac{1}{N}\sum_{i=1}^{N} \langle v_1, \text{Log}_\mu(x_i)\rangle^2 + \frac{1}{N}\sum_{i=1}^{N} \langle v_2, \text{Log}_\mu(x_i)\rangle^2 \\
&\simeq \frac{1}{N}\sum_{i=1}^{N} \|\text{Log}_\mu(\pi_{V_1}(x_i))\|^2 + \frac{1}{N}\sum_{i=1}^{N} \|\text{Log}_\mu(\pi_{V_2}(x_i))\|^2,
\end{aligned}$$

and thus we almost fulfilled (2.15).

Therefore, taking $v_2 \in \text{span}\{v_1\}^\perp$ to obtain unicity of the definition seems reasonable in the light of the above discussion.

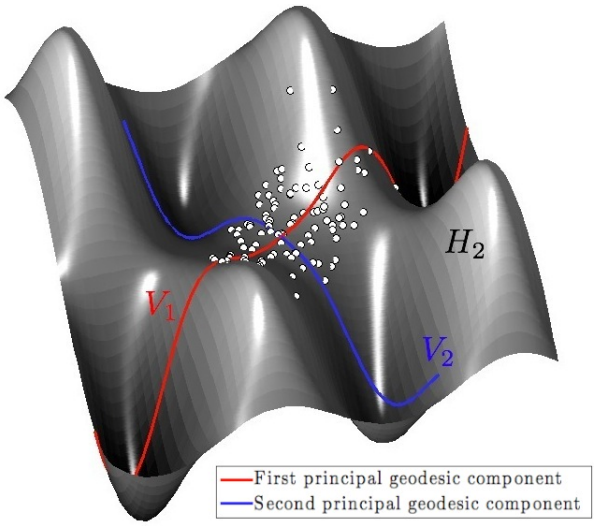


Figure 2.7.: ***Example of a principal geodesic analysis.*** *The first principal geodesic submanifold is $H_1 = V_1$, the second H_2 is the manifold itself. The principal geodesic components V_1, V_2 are analog of the principal components in PCA.*

We can extend this construction to all principal geodesic submanifolds and principal geodesic components to obtain :

Definition 2.7 (*Principal geodesic submanifolds/components*)

Let M^p be a Riemannian manifold, $x_1, \cdots, x_N \in M^p$ and μ be the intrinsic mean of those points. Let $U \subset T_\mu(M^p)$ be a neighborhood of 0 such that projection is well-defined for all geodesic submanifolds of $Exp_\mu(U)$, and such that $Exp_\mu : U \to Exp_\mu(U)$ is a local diffeomorphism. Then, we define the p ***principal geodesic submanifolds / components*** *as :*

- ***First principal geodesic submanifold / component :***

  Let $v_1 \in T_\mu(M^p)$ be such that :

  $$v_1 = \underset{\|v\|=1}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} \|Log_\mu(\pi_H(x_i))\|^2,$$

  with $H = Exp_\mu(span\{v\} \cap U)$.

  Then, the ***first principal geodesic submanifold*** *is :*

  $$H_1 = Exp_\mu(span\{v_1\} \cap U),$$

  and the ***first principal geodesic component*** *is :*

  $$V_1 = Exp_\mu(span\{v_1\} \cap U).$$

- ***k-th principal geodesic submanifold / component :***

  Let $v_k \in T_\mu(M^p)$ be such that :

  $$v_k = \underset{\substack{\|v\|=1 \\ v \in span\{v_1, \cdots, v_{k-1}\}^\perp}}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} \|Log_\mu(\pi_H(x_i))\|^2,$$

  with $H = Exp_\mu(span\{v_1, \cdots, v_{k-1}, v\} \cap U)$.

  Then, the ***k-th principal geodesic submanifold*** *is :*

  $$H_k = Exp_\mu(span\{v_1, \cdots, v_{k-1}, v_k\} \cap U),$$

  and the ***k-th principal geodesic component*** *is :*

  $$V_k = Exp_\mu(span\{v_k\} \cap U).$$

Remark : As we didn't exactly verified (2.15), the principal geodesic components are not strictly speaking the analogs of principal components in PCA. However, we will consider in this study that the approximation we made is good enough to make such an analogy.

Computing PGA

Let see now how we could efficiently compute PGA. The first thing we need to notice is that the geodesic H_k and V_k are entirely determined as soon as we know $v_k \in T_\mu(M^p)$. thus, we will only focus on the computation of v_k. Let say we already computed $v_1, \dots, v_{k-1} \in T_\mu(M^p)$ and we want to compute v_k. We have :

$$v_k = \underset{\substack{\|v\|=1 \\ v \in \text{span}\{v_1, \cdots, v_{k-1}\}^\perp}}{\text{argmax}} \frac{1}{N} \sum_{i=1}^{N} \|\text{Log}_\mu(\pi_H(x_i))\|^2,$$

with $H = \text{Exp}_\mu(\text{span}\{v_1, \cdots, v_{k-1}, v\} \cap U)$. As $\{v_1, \cdots, v_{k-1}, v\}$ is an orthogonal basis, we can one more time exploit the approximation (2.13). This yields :

$$\begin{aligned} v_k &\simeq \underset{\substack{\|v\|=1 \\ v \in \text{span}\{v_1, \cdots, v_{k-1}\}^\perp}}{\text{argmax}} \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=1}^{k-1} \langle \text{Log}_\mu(x_i), v_j \rangle^2 + \langle \text{Log}_\mu(x_i), v \rangle^2 \right), \\ \Leftrightarrow \quad v_k &\simeq \underset{\substack{\|v\|=1 \\ v \in \text{span}\{v_1, \cdots, v_{k-1}\}^\perp}}{\text{argmax}} \frac{1}{N} \sum_{i=1}^{N} \langle \text{Log}_\mu(x_i), v \rangle^2. \end{aligned} \tag{2.16}$$

As $T_\mu(M^p)$ is a liner space, $\frac{1}{N} \sum_{i=1}^{N} \langle \text{Log}_\mu(x_i), v \rangle^2$ is simply the classic sample variance of the data $\text{Log}_\mu(x_i)$ projected onto the one-dimensional linear subspace $\text{span}\{v\}$. Thus, we can translate (2.16) as follows : *find* $v_k \in$ *span*$\{v_1, \cdots, v_{k-1}\}^\perp$ *such that Var*$(\langle \text{Log}_\mu(x_i), v_k \rangle)$ *is maximized.*
But, this is simply classic principal component analysis computed on the data $\text{Log}_\mu(x_i)$, under the assumption that $v_1, \cdots, v_{k-1}$ are the $k-1$ first principal components ! Thus, we could show by induction that a good approximation for the v_ks are the principal components of the PCA computed on the data $\text{Log}_\mu(x_i)$.
This leads to the following algorithm to approximately compute PGA :

Algorithm 2: PGA computation

Data: A set of points $x_1, \dots, x_N \in M^p$, with intrinsic mean μ.
Result: Principal directions $v_k \in T_\mu(M^p)$
$u_i = \text{Log}_\mu(x_i)$;
$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} u_i u_i^T$;
$\{v_k\}$ = eigenvectors of $\hat{\Sigma}$, sorted in order of decreasing eigenvalues;

CHAPTER

3 The Shape Space of Triangles

In this chapter, we construct the shape space Σ_2^3 of triads of labelled planar points , and show that it's homeomorphic to the sphere S^2. We then apply some of the statistic tools we previously designed for manifolds on the particular example of the shape space of triangles.

1. The shape space of triangles

Shape theory tries to capture shapes of objects by looking at the total of all information that is invariant under *translations, rotations and rescaling* (invariance under similarity transformations). Roughly speaking, we remove all superfluous information concerning location, scale and orientation of the object, by taking the quotient of our Euclidean space under similarity transformations. Then, we define the remaining information as the shape of the object.

However, this cannot be achieved without a parametrization of the object, allowing us to manipulate it in a mathematical way. As a perfect mathematical description of an object is often impossible, we will only focus on special points called **landmarks** : they are points of special interest for the considered object, which are meant to provide a partial geometric description of it (see fig. 3.1). This approach has been first introduced by Kendall in 1977.

In the following developments, we will consider the simple case (but yet interesting), where we only have three landmarks $x_1, x_2, x_3 \in \mathbb{R}^2$ (as in fig. 3.1). Those triads of labelled planar points can be seen as triangles in the plane, which legitimate our study of the shape space of triangles. The construction we propose here is inspired both on the works of Small [1] and Stoyan [2].

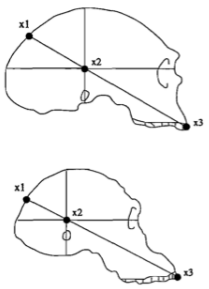


Figure 3.1.: *Landmarks of human skulls (Neanderthal and australopithecine).*

1.1. Construction of the shape space

Let x_1, x_2, x_3 lie in $\mathbb{R}^2$. In the following, we will identify $\mathbb{R}^2$ with the complex plane, so that similarity transformations can be easily interpreted of basic operations of complex multiplication and addition.

In language of group theory, the **shape** of a given triad $(x_1, x_2, x_3) \in \mathbb{C}^3$ is its *orbit* under the symmetry group generated by :

1. the *addition* of a same complex number z to each point x_i. This operation corresponds to a *translation* of the triad in the complex plane.

2. the *multiplication* of each point x_i by the same non-null complex number w. This operation corresponds to an *homothetic transformation* of the triad, or a combination of a rotation and a dilatation.

Thus, this orbit is the following set :

$$\mathcal{O}(x_1, x_2, x_3) = \{(wx_1 + z, wx_2 + z, wx_3 + z) \ : \ z \in \mathbb{C}, \ w \in \mathbb{C}^*\}.$$

Then, the **space shape** Σ_2^3 is simply the set of all orbits of triads in $\mathbb{C}^3$ under the symmetry group :

$$\Sigma_2^3 = \{\mathcal{O}(x_1, x_2, x_3) \subset \mathbb{C}^3 : (x_1, x_2, x_3) \in \mathbb{C}^3\}.$$

However, a more explicit description of the shape space is required for practical purposes. In the following, we will try to characterize topologically and geometrically the shape space by successively removing superfluous information about location, scale and orientation of the landmarks.
First, the location of a triad $(x_1, x_2, x_3) \in \mathbb{C}^3$ can be described through its uniquely defined **centroid** $\bar{x}$:

$$\bar{x} = \frac{x_1 + x_2 + x_3}{3}.$$

Then, any triad can be translated to a centered triad $(x_1 - \bar{x}, x_2 - \bar{x}, x_3 - \bar{x}) \in \mathbb{C}^3$ so that $((x_1 - \bar{x}) + (x_2 - \bar{x}) + (x_3 - \bar{x}))/3 = 0$. To remove the scale information in the centered triad, we simply divide it by its norm. The resultant vector $\tau \in \mathbb{C}^3$ is called the **pre-shape** of the landmarks :

$$\tau = \left(\frac{x_1 - \bar{x}}{\sqrt{\sum_{i=1}^3 \|x_i - \bar{x}\|^2}}, \frac{x_2 - \bar{x}}{\sqrt{\sum_{i=1}^3 \|x_i - \bar{x}\|^2}}, \frac{x_3 - \bar{x}}{\sqrt{\sum_{i=1}^3 \|x_i - \bar{x}\|^2}} \right).$$

In order for this representation to be well-defined, we must exclude the totally coincident triads $x_1 = x_2 = x_3$. Such triads are degenerate in shape and will be said to have an **indeterminate shape**. For the following, we shall assume that all triads are non totally coincident.
The pre-shape τ lies in a constrained subset of the original Euclidean space $\mathbb{C}^3$, intersection of the linear subspace :

$$F^4 = \{(x_1, x_2, x_3) \in \mathbb{C}^3 : \sum_{i=1}^3 x_i = 0\},$$

with the unit sphere :

$$S^5 = \{(x_1, x_2, x_3) \in \mathbb{C}^3 : \sum_{i=1}^3 \|x_i\|^2 = 0\}.$$

The intersection $S^3 = F^4 \cap S^5$, is a 1-dimensional sphere within the ambient Euclidean space $\mathbb{C}^3$ and a 3-dimensional sphere within the ambient Euclidean space $\mathbb{R}^6$. We shall refer to this sphere as the **pre-shape space**.
Finally, we remove the remaining information about orientation by defining an equivalence relationship onto the pre-shape space. Two pre-shapes $\alpha, \beta \in S^3$ are said to be equivalent if there exists $w \in \mathbb{C}, \|w\| = 1$ such that :

$$(\alpha_1, \alpha_2, \alpha_3) = (w\beta_1, w\beta_2, w\beta_3).$$

We'd like to say that two triads have the same shape if their pre-shape can be transformed one into the other through a rotation, or a multiplication by a complex w. Then, those two pre-shapes would be equivalent in the sense of the above equivalence relationship, which leads us to define the shape of a triad as the orbit of its pre-shape under this very same equivalence relationship :

$$\mathcal{O}(\tau) = \{(w\tau_1, w\tau_2, w\tau_3) : w \in \mathbb{C}, \|w\| = 1\} \subset S^3.$$

Exploiting the identification of $\mathbb{C}^3$ with $\mathbb{R}^6$, one can rewrite $w\tau$ as $cos(\theta)u + sin(\theta)v$, with $u = (\tau_{11}, \tau_{21}, \tau_{31}, \tau_{12}, \tau_{22}, \tau_{32})$, $v = (-\tau_{12}, -\tau_{22}, -\tau_{32}, \tau_{11}, \tau_{21}, \tau_{31})$ and $\tau_j = \tau_{j1} + i\tau_{j2}$, for $j = 1, \ldots, 3$ Therefore, those orbits are non intersecting great circles partitioning the 3-dimensional sphere S^3. The collection of those orbits is called the **shape space** Σ_2^3 :

$$\Sigma_2^3 = \{\mathcal{O}(\tau), \tau \in S^3\}.$$

One can show that this set of great circles can be naturally identified with the **complex projective plane** $\mathbb{CP}^1 = \{(w\tau_1, w\tau_2, w\tau_3) : w \in \mathbb{C}\}$, with $(\tau_1, \tau_2, \tau_3) \in \mathbb{C}^3$ distinct from the origin. Then, a famous result of topology states that this complex projective plane is topologically equivalent (homeomorphic) to the sphere S^2. This leads to the following theorem, introduced by Kendall in 1984 :

Theorem 1.1 (*D.G. Kendall, 1984*)

The shape space of triads of planar points is given by :

$$\Sigma_2^3 = \mathbb{CP}^1.$$

Moreover, we have the following identification :

$$\Sigma_2^3 = S^2.$$

Remark : This theorem can be extended for the shape space of k-ads of planar points : $\Sigma_2^k = \mathbb{CP}^{k-2}$ (see [2].

The above construction surely isn't the most straightforward one to establish this result, who could have been obtained much more easily. However, this construction has the advantage to provide us a natural metric on the shape space, that we shall discuss in the following section.

✧✧✧

1.2. The Procrustean metric onto the shape space Σ_2^3

The theorem 1.1 provides us a topological characterization of the shape space. However, for statistical purposes, topological considerations aren't enough, and we need some additional structure on the shape space. More precisely, we need to define a *metric* on the shape space, so we could apply the statistic tools previously designed for non linear situations. Rather that beginning at the local level by constructing a metric tensor, we will exploit the geodesic metric provided on the pre-shape space and the quotient structure of the shape space.

As we have seen previously, the pre-shape space is the 3-dimensional sphere S^3. Thus, it is a Riemannian manifold on which a natural metric is the geodesic distance $d(\cdot,\cdot)$. Then, exploiting the quotient structure of the shape space we can naturally define a metric on Σ_2^3 :

Definition 1.1 (*Procrustean distance*)

*Let $\tau_1, \tau_2 \in S^3$ be two pre-shapes of non-trivial planar triads. Then, the **Procrustean distance** $\tilde{d}(\cdot,\cdot)$ between the two shapes $\mathcal{O}(\tau_1), \mathcal{O}(\tau_2) \in \Sigma_2^3$ is :*

$$\tilde{d}\left(\mathcal{O}(\tau_1), \mathcal{O}(\tau_2)\right) = \inf\{d(x,y) : x \in \mathcal{O}(\tau_1), y \in \mathcal{O}(\tau_2)\}, \tag{3.1}$$

with $d(\cdot,\cdot)$ the geodesic distance defined on the pre-shape space S^3.

The usual geodesic distance between two points of S^3 is the shorter of the two arcs of the great circle joining the two points. This is simply the angle made between the two vectors $\tau_1, \tau_2 \in S^3$. Thus, the geodesic distance from τ_1 to τ_2 is given by :

$$d(\tau_1, \tau_2) = \arccos(\langle \tau_1, \tau_2 \rangle),$$

with $\langle \cdot, \cdot \rangle$ the Hermitian inner product on $\mathbb{C}^3$: $\langle \tau_1, \tau_2 \rangle = \Re\left(\sum_{j=1}^3 \tau_{1j}\tau_{2j}^*\right)$. More generally, on the sphere $S^3(r)$, the geodesic distance from τ_1 to τ_2 is :

$$d(\tau_1, \tau_2) = r \times \arccos(r^{-2}\langle \tau_1, \tau_2 \rangle). \tag{3.2}$$

Then, if σ_1 and σ_2 are two shapes in Σ_2^3, and τ_1 and τ_2 are two representative pre-shapes so that $\sigma_j = \mathcal{O}(\tau_j)$ for $j = 1, 2$, we can show that the minimum in (3.1) is :

$$\tilde{d}(\sigma_1, \sigma_2) = \arccos\left(\| \sum_{j=1}^{3} \tau_{1j}\tau_{2j}^* \|\right). \tag{3.3}$$

An interesting thing to note about (3.3) is that, as the argument of $\arccos(\cdot)$ is always non-negative, the maximum Procrustean distance between two shapes in Σ_2^3 is $\frac{\pi}{2}$. Moreover, it is reassuring to see that the Procrustean distance do not depend on the orientation of the chosen representative pre-shapes τ_1 and τ_2. In fact, rotating τ_1 and τ_2 is equivalent to multiplying each τ_{kj} by a complex w of the unit sphere, which can be factored out of the summation and has modulus one, so the argument in the $\arccos$ function is unchanged.

✧✧✧

1.3. Shape coordinates

To complete our description of the shape space Σ_2^3, we now wish to define a system of coordinates on it. We could simply exploit its structure of Riemannian manifold and use a compatible system of charts on it, to provide local coordinate systems. However, such a parametrization isn't convenient for graphical representations and computations. Therefore, we wish to design a *single* coordinate system to parametrize the whole shape space. Unfortunatetely, being homeomorphic to the sphere S^2, the shape space cannot be fully describe through one single system of coordinates. Thus, the system of coordinate that we propose here is *degenerate*, in the sense that it fails to describe some shapes (namely the shapes of triads such that $x_1 = x_2 \neq x_3$).

► About the degeneracy

It's important to specify that this degeneracy forbids us to strictly speaking identify a shape with its coordinates, and we shall always remember that the appropriate setting to represent shapes is through their orbit under the symmetry group. However, mathematically elegant as this representation might be, it has some limitations, the main one being its lack of conveniency for graphical representations and computations, which is cured, somewhat, by the description we propose here.

Let consider a triad $(x_1, x_2, x_3) \in \mathbb{C}^3$ such that $x_1 \neq x_2$ and its associated shape $\mathcal{O}(x_1, x_2, x_3) = \{(wx_1 + z, wx_2 + z, wx_3 + z) : z \in \mathbb{C}, w \in \mathbb{C}^*\}$, orbit of the triads under the symmetry group. For any such triad, we will choose as a representant of the equivalence class the normalized triad, image of the original triad under the similarity transformation moving x_1 to -1, x_2 to 1 and x_3 to a complex z to determine (see fig. 3.2). Then, the shape of the triangle is entirely determined by the position of z, and we will call the real and imaginary coordinates of z the **Bookstein coordinates**.
To find the coordinates of z we just have to find $\alpha, \beta \in \mathbb{C}$ such that :

$$\begin{cases} \alpha x_1 + \beta &= -1, \\ \alpha x_2 + \beta &= 1. \end{cases}$$

Solving this linear system leads us to : $\alpha = 2/(x_2 - x_1)$, $\beta = -(x_1 + x_2)/(x_2 - x_1)$, and thus

$$z = \frac{2x_3 - (x_1 + x_2)}{x_2 - x_1}.$$

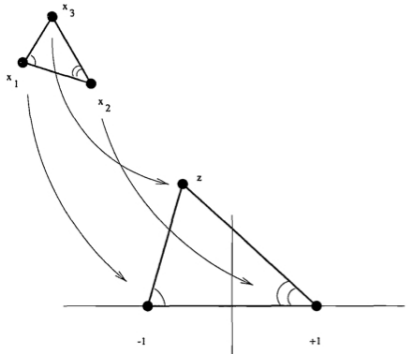


Figure 3.2.: ***Bookstein coordinates for three planar points.*** *The original triad is mapped by a similarity transformation to the standardized triad* $(-1, 1, z) \in \mathbb{C}^3$.
© ***G. Small***

Such a coordinatization is not without its deficiencies, as the representation breaks down if $x_1 = x_2 \neq x_3$, which are totally well-defined shapes whereas their Bookstein coordinates are not. In such case, the shape of these triads is most naturally interpreted as $z = \infty$. Then, one attempt of curing the degeneracy of the above representation, could be to add to it the point $z = \infty$. However, the extended

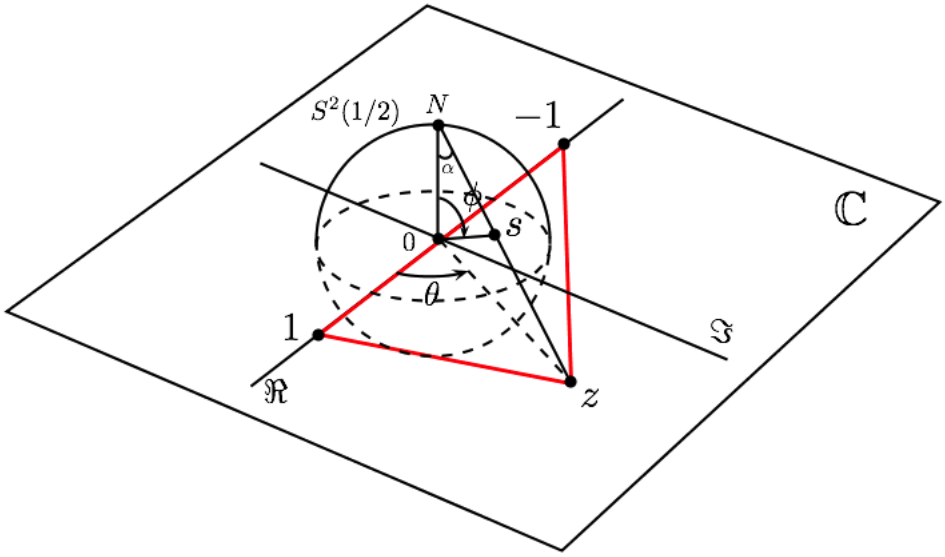


Figure 3.3.: ***The stereographic projection.*** *The stereographic projection is a 1-1 correspondence between $\mathbb{C} \cup \{\infty\}$ and the sphere $S^2(1/2)$. On the complex plane a sphere sits so that its equatorial plane merges with the complex plane. For any point $z \in \mathbb{C}$ we draw a line from the north pole of the sphere N to z. Then, the intersection point s of this line with the sphere $S^2(1/2)$ is the image of z under the stereographic projection.*

complex plane $\mathbb{C} \cup \{\infty\}$ obtained this way looses its Euclidean structure, as we can put it in 1-1 correspondence with the sphere S^2 through the **stereographic projection**.

Proposition 1.1 (*Stereographic projection*)

The 1-1 correspondence between the extended complex plane $\mathbb{C} \cup \{\infty\}$ and the sphere $S^2(1/2)$ of radius $r = 1/2$ is called the stereographic projection and is given by :

$$\xi : \mathbb{C} \cup \{\infty\} \to S^2(1/2); \quad z \mapsto s = \begin{cases} \left(\frac{1}{2}, \arg(z), \pi - 2\arctan(2|z|)\right) & \text{if } z \neq \infty \\ \left(\frac{1}{2}, 0, 0\right) & \text{if } z = \infty \end{cases},$$

where the result is expressed in spherical coordinates (r, θ, ϕ) (see fig. 3.3).

Remark : In the above proposition, we imposed the radius of the sphere to be equal to $1/2$ so the maximum geodesic distance between two points on this sphere is $\pi/2$ (according to (3.2)). This way, we can show that the geodesic distance on this sphere is equivalent with the natural Procrustean distance on the shape space Σ_2^3 defined in (3.3).

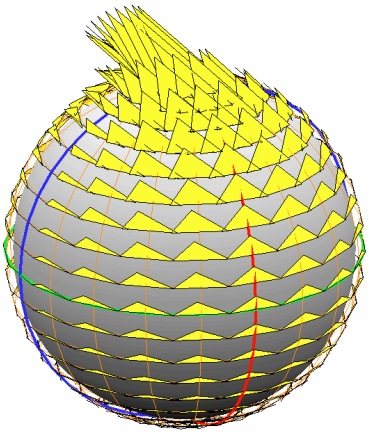
This proposition allow the representation of the shape of a triad as a point of the sphere $S^2(1/2)$ by the **spherical coordinates** :

$$\xi(z) = \left(\frac{1}{2}, \theta, \phi\right), \quad \text{where } \theta \in [0, 2\pi] \text{ and } \phi \in [0, \pi].$$

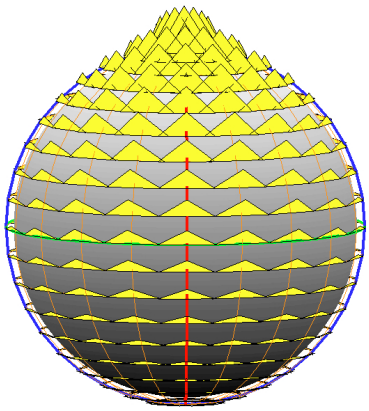
Moreover, we note the following useful relation :

$$z = \xi^{-1}(s) = \frac{1}{2}\tan\left(\frac{\pi - \phi}{2}\right)e^{i\theta}.$$

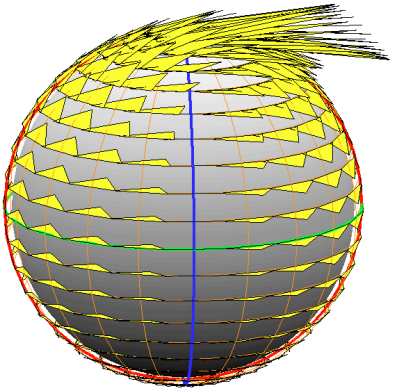
In figure we exploited this representation to find the location on the sphere of various classes of triads.



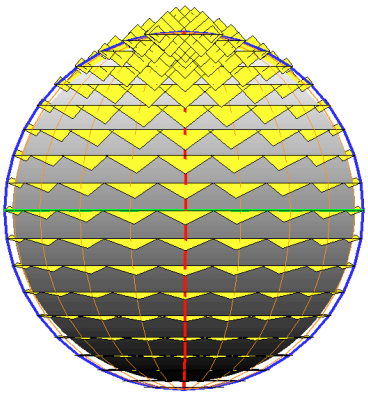
(a) *On this perspective, we can clearly see the degeneracy of the representation on the North pole.*



(b) *Isosceles triads remains isocele while moving on the meridian from north to pole, only their height is dilated or shrunk.*



(c) *The effect of crossing the collinear triads great circle while moving on a parallel line.*



(d) *Rear view.*

Figure 3.4.: ***Different views of the shape space of planar triads.*** *For 400 points on the sphere, we computed the associated Bookstein shape coordinates through the inverse of the stereographic projection and reconstructed the associated shape. We also drawn some important classes of shapes. The blue great circle corresponds to the great circle of collinear triads, while the red one corresponds to one of the great circles of isosceles triads. The green line is the equatorial line. On each view, we can see the effect of moving along a meridian line or a parallel line. See appendix 6 for the Matlab code used to produce these figures.*

2. Nonlinear statistics on the shape space of triangles

We finish this chapter by applying the principal geodesics analysis procedure to a sample of shapes chosen in Σ_2^3 (see fig. 3.5).

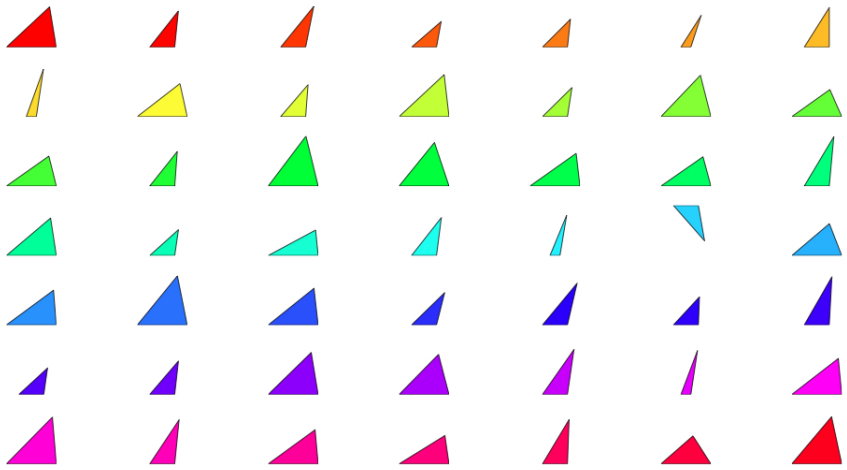
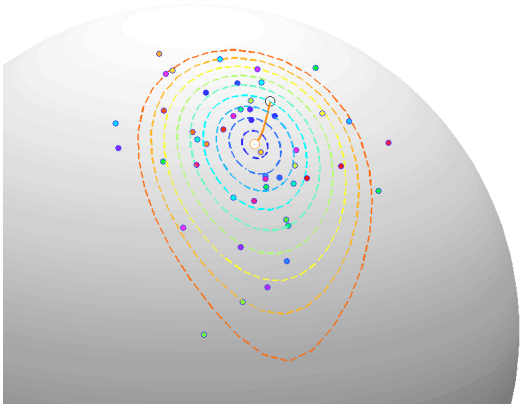


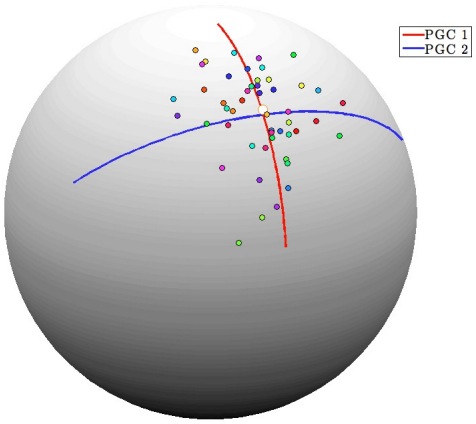
Figure 3.5.: *A sample of 49 shapes randomly chosen in* Σ_2^3.

We wish to compute a principal geodesic analysis to efficiently describe the variability of the above sample of shapes. To do so, we apply the algorithm 2, which assumes that we already computed the intrinsic mean μ of the sample. In this case, we approximate the intrinsic mean μ by the extrinsic mean μ_ϕ, computed by the gradient descent algorithm (see fig. 3.6a). This approximation is good enough (at μ_ϕ, we have $\nabla f(\mu_\phi) = (0.03, 0.01) \simeq 0$, thus $\mu_\phi \simeq \mu$), and allow us to save computation time.
The results of the PGA can be seen on fig. 3.6b, where we drawn the two principal geodesic components (PGC) V_1 and V_2. The first principal geodesic V_1, which is responsible for approximately 63.2% of the total variability[1] , helps us to reduce the dimensionality of the problem : fig. 3.6d shows how we can efficiently describe the variability of the sample by focusing only on the first principal geodesic.
Finally, we drawn on fig. 3.6c the Bookstein coordinates of the shapes in $\mathbb{C} \cup \{\infty\}$ and the image of the PGCs by the inverse of the stereographic projection. On this figure, we can clearly see that the principal geodesic analysis performed on the shapes in Σ_2^3 is not at all equivalent with a classic PCA performed on the Bookstein coordinates. Moreover, we computed the mean of the 49 Bookstein coordinates of the shapes and we can see that this mean slightly differs with the Bookstein coordinates of the intrinsic mean. Those two remarks emphasizes the fact that classic statistical tools do not stand anymore in nonlinear situations, forcing us to redesign them for such cases.

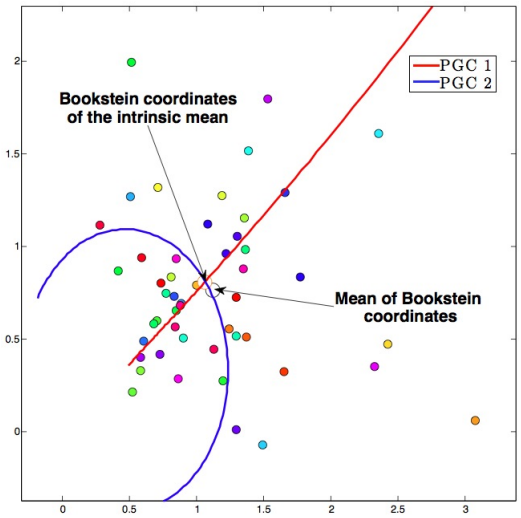
[1] the way we defined the principal geodesic components legitimates such a decomposition of the variability, see eq. (2.15)



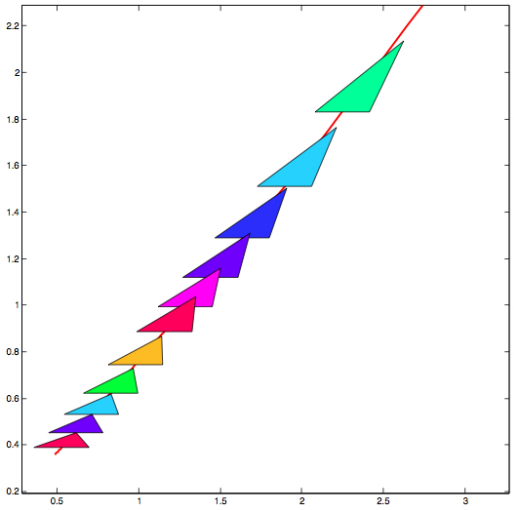
(a) *Computing the extrinsic mean of the shapes in Σ_2^3 with the gradient descent algorithm. This provides a good approximation of the intrinsic mean and saves computation time.*



(b) *Computing a principal geodesic analysis on the sample of shapes. PGC1 and PGC2 respectively correspond to V_1 and V_2, the principal geodesic components.*



(c) *Bookstein coordinates of the shapes and the PGCs in $\mathbb{C} \cup \{\infty\}$ obtained by the inverse of the stereographic projection.*



(d) *Variation of the shape along the first geodesic component.*

Figure 3.6.: ***Principal geodesic analysis of 49 shapes randomly chosen in Σ_2^3.*** *See appendix 7 for the Matlab code used to produce these figures.*

Conclusion

In conclusion, our knowledge of Riemannian manifolds from chapter 1 allowed us to successfully extend principal components analysis into manifold setting. While existing works on the subject mostly focus on the principal geodesics submanifolds, best n-dimensional submanifolds to describe the data, we additionally tried to provide a generalization of principal components through principal geodesics components, allowing us to re-express the data as a sequence of "independent" components. However, our construction is based on the approximation (2.13), which might not be good enough in some cases. Thus, an interesting future work would be to look for a construction independent from such an approximation.
Moreover, the construction of the shape space of triads Σ_2^3 could be pursued in greater generality for any k-ads of planar landmarks. This would allow us to handle more complex shape representations.
Finally, an interesting goal could be to extend new statistical tools onto manifold setting, to allow a better analysis of objects lying in nonlinear spaces.

✧✧✧

Acknowledgments

I wish to acknowledge the Pr. Victor Panaretos for having proposed me this project and followed me all along its realization. His advices and support have always been very helpful, and I keep a very good memory of our discussions, scientific or not.

Bibliography

[1] C. G. Small *The Statistical Theory of Shape*, Springer series in statistics, 1996.

[2] Dietrich Stoyan, Wilfrid S. Kendall, Joseph Mecke, *Stochastic geometry and its applications*, Wiley series in probability and mathematical statistics, Second edition, 1995.

[3] I.L. Dryden and K.V. Mardia, *Statistical shape analysis*, Chichester : John Wiley & Sons, 2002.

[4] I. T. Joliffe, *Principal Component Analysis*, Second edition, Springer series in statistics, 2002.

[5] P. T. Fletcher, C. Lu, S. M. Pizer, S. Joshi, *Principal Geodesic Analysis for the Study of Nonlinear Statistics of Shape*, MIDAG.

[6] H.Karcher, *Riemannian center of mass and mollifier smoothing*, Communications on pure and applied mathematics, vol.30, no. 5, 1977.

[7] S. R. Buss and J. P. Fillmore, *Spherical averages and applications to spherical splines and interpolations*, ACM transactions on Graphics, vol 20, 2001.

A Matlab and R code

1. Figure 2.1

```
FTSE<-read.table("FTSE.csv",header=FALSE,sep=";",dec=",")
FTSE.ts<-ts(FTSE$V2,frequency=12,start=c(1986,1))
#Partitioning graphic
layout(matrix(1:4, 2, 2))
#Variance stabilization :
FTSE.ts<-log(FTSE.ts)
#Removing trend :
#Using linear regression (detrended):
FTSE.lm<-lm(FTSE.ts~time(FTSE.ts))
abline(FTSE.lm)
FTSE.notrend<-FTSE.ts-fitted(FTSE.lm)
plot.ts(FTSE.notrend,main="Detrended (linear reg)")
abline(0,0)
#Cyclic behavior (seasonality)
#Periodogram :
I=abs(fft(FTSE.notrend))^2/length(FTSE.notrend)
P=(4/length(FTSE.notrend))*I[1:(length(FTSE.notrend)/2)]
f=0:(length(FTSE.notrend)/2-1)/length(FTSE.notrend)
plot(f,P,type="l",xlab="Frequency",
+ ylab="Scaled periodogram")
P.freq<-data.frame(f,P)
P.freq<-P.freq[ order(-P.freq[,2]), ]
#n linear regression
plot(1:length(FTSE.notrend),FTSE.notrend,
+main="Detrended FTSE",type="l")
cosine<-cos(2*pi*P.freq[1,1]*1:length(FTSE.notrend))
sine<-sin(2*pi*P.freq[1,1]*1:length(FTSE.notrend))
phase.lm<-lm(FTSE.notrend~0+cosine+sine)
summary.phase<-summary(phase.lm)
phi=atan(-summary.phase$coefficients[2,1]
+/summary.phase$coefficients[1,1])+pi
```

```
lines(1:length(FTSE.notrend),
sqrt(P.freq[1,2])*cos(2*pi*P.freq[1,1]*1:length(FTSE.notrend)
+phi),col="red")
FTSE.notrend.lm<-FTSE.notrend
sum=sqrt(P.freq[1,2])*
+ cos(2*pi*P.freq[1,1]*1:length(FTSE.notrend)+phi)
color<-c("blue","green","purple","orange","red","blue",
"green","purple")

for (i in 2:((length(FTSE.notrend)-1)/2)){
  FTSE.notrend.lm<-FTSE.notrend.lm-fitted(phase.lm)
  cosine<-cos(2*pi*P.freq[i,1]*1:length(FTSE.notrend))
  sine<-sin(2*pi*P.freq[i,1]*1:length(FTSE.notrend))

  phase.lm<-lm(FTSE.notrend.lm~0+cosine+sine)
  summary.phase<-summary(phase.lm)
  phi=2*atan(summary.phase$coefficients[2,1]/
  (summary.phase$coefficients[1,1]+
  sqrt(summary.phase$coefficients[1,1]^2+
  summary.phase$coefficients[2,1]^2)))
  x=sqrt(summary.phase$coefficients[1,1]^2+
  summary.phase$coefficients[2,1]^2)*
  cos(2*pi*P.freq[i,1]*(1:length(FTSE.notrend))+phi)
  sum=sum+x
  if (i<9){lines(length(FTSE.notrend):1,sum,col=color[i-1])}
  if (i==9){lines(length(FTSE.notrend):1,sum,col="red",lwd=2)
  }
  }
  }
}
```

2. Figures 2.2 and 2.4

```
mu<-c(0,0)
sigma<-matrix(c(7,3,3,7),2,2)
draws<-mvrnorm(n=300,mu,sigma)
plot(draws,asp=1,pch=20,xlab="x",ylab="y")
ellipse(mu,sigma,3,col="blue")
ellipse(mu,sigma,2,col="green")
ellipse(mu,sigma,1,col="red")
rm(spectral)
spectral<-eigen(sigma,symmetric=TRUE)
pente1=spectral$vectors[2,1]/spectral$vectors[1,1]
pente2=spectral$vectors[2,2]/spectral$vectors[1,2]
```

```
abline(0,pente1,lwd=2,lty="dashed",col="red")
abline(0,pente2,lwd=2,lty="dashed",col="blue")
legend("topright" ,c("c=1","c=2","c=3","1st PC","2d PC"),
lty=c(1,1,1,2,2),col=c("red","green","blue","red","blue"),
lwd=c(2,2,2,2,2),bty="n")
```

3. Example 1.4

```
universities_scores<-read.table("ranking.csv",header=TRUE,sep=";")
universities.names<-as.vector(universities_scores$X)
universities_scores<-read.table("ranking2.csv",header=TRUE,sep=";")
row.names(universities_scores)<-universities.names
universities_scores=t(universities_scores)
#universities_scores=log(universities_scores)
#Centering the data
for (i in 1:19){universities_scores[i,]<-universities_scores[i,]
-mean(universities_scores[i,])}
#Re-weighting the data
for (i in 1:19){universities_scores[i,]<-universities_scores[i,]/
sqrt(var(universities_scores[i,]))}
#Building sample cov matrix :
#sample_cov<-(1/(27))*(as.matrix(universities_scores)
sample_cov<-cor(t(universities_scores))
spectral_decomp<-eigen(sample_cov,symmetric=TRUE)
spectral_decomp
spectral_decomp$values[1]/sum(spectral_decomp$values)
PC_scores<-c()
spectral_decomp$vectors<--spectral_decomp$vectors
for (i in 1:29) {
sum=0
for (j in 1:19) { sum=sum + universities_scores[j,i]*
spectral_decomp$vectors[j,1] }
PC_scores[i]<-sum/sum(spectral_decomp$vectors[,1])
}
Ranking<-data.frame(Universities=universities.names,Score=PC_scores)
#row.names(Ranking)<-universities.names
Ranking<-Ranking[ order(-Ranking[,2]), ]

#2d PC :
PC_scores_2<-c()
for (i in 1:28) {
  sum=0
  for (j in 1:19) { sum=sum + universities_scores[j,i]*
  spectral_decomp$vectors[j,2] }
  PC_scores_2[i]<-sum/sum(spectral_decomp$vectors[,2])
```

```
}
plot(PC_scores,PC_scores_2,pch=16)
abline(h=0)
abline(v=0)
Partial_Sigma=matrix(c(spectral_decomp$values[1],0,0,
spectral_decomp$values[2]),2,2)
ellipse(c(0,0),Partial_Sigma,1,col="red")
ellipse(c(0,0),Partial_Sigma,2,col="green")
ellipse(c(0,0),Partial_Sigma,3,col="blue")


(spectral_decomp$values[1]+spectral_decomp$values[2])/
sum(spectral_decomp$values)

Cumulative_PC_scores<-(spectral_decomp$values[1]/
(spectral_decomp$values[1]+spectral_decomp$values[2]))*
PC_scores+(spectral_decomp$values[2]/(spectral_decomp$values[1]+
spectral_decomp$values[2]))*PC_scores_2
rm(Ranking)
Ranking<-data.frame(Universities=universities.names,
Score=Cumulative_PC_scores)
Ranking<-Ranking[ order(-Ranking[,2]), ]
```

4. Figure 2.5

```
[x,y] = meshgrid([-2:.2:2]);
Z = x.*exp(-x.^2-y.^2);
syms u v
NORM=(u-x).^2+(v-y).^2+(u.*exp(-u.^2-v.^2)-Z).^2;
g=(1/(2*length(NORM)))*sum(sum(NORM,1),2);
grad_g=gradient(g,[u,v]);
colormap(gray)
surf(x,y,Z,'EdgeColor','none')
hold on
plot3(x,y,Z,'o','MarkerEdgeColor','k','MarkerFaceColor','w',
'MarkerSize',5);
mu=[x(340),y(340)];
MU_1(1)=mu(1);
MU_2(1)=mu(2);
tau=0.002;
for i=1:400
    evaluate=subs(grad_g,[u,v],[mu(1),mu(2)]);
    r=-evaluate;
    mu=mu+tau*r';
    MU_1(i+1)=mu(1);
```

```
    MU_2(i+1)=mu(2);
end
MU_Z= MU_1.*exp(-MU_1.^2-MU_2.^2);
plot3(MU_1,MU_2,MU_Z,'Color',[1 .5 0],'LineWidth',3);
plot3(MU_1(1),MU_2(1),MU_Z(1),'o','MarkerEdgeColor','k',
'MarkerFaceColor','w','MarkerSize',10)
plot3(MU_1(length(MU_1)),MU_2(length(MU_1)),MU_Z(length(MU_Z)),
'o','MarkerEdgeColor',[1 .5 0],'MarkerFaceColor','w','MarkerSize',10)

Color=jet(11);
t=linspace(0,2*pi,30);
j=0;
for i=0:0.2:2
    X=i*cos(t);
Y=i*sin(t);
Z_cont=X.*exp(-X.^2-Y.^2);
j=j+1;
plot3(X,Y,Z_cont,'--','Color',Color(j,1:3),'LineWidth',1);
end


axis vis3d
axis auto
axis off
camlight('headlight');
lighting gouraud;
set(gcf,'Color','w')
```

5. Figure 2.7

```
[x,y] = meshgrid([-3:.1:3]);
Z = sin(2*x).*sin(y).*sqrt(x.^2+y.^2);
surf(x,y,Z,'EdgeColor','none')
hold on

%shading flat
colormap(gray)


mu = [0;0];
Sigma = [0.3 .1; .2 0.3]; R = chol(Sigma);

for i=1:100
```

```
r = mu + R*randn(2,1);
M(:,i)=r;
plot3(r(1),r(2),sin(2*r(1)).*sin(r(2)).*sqrt(r(1).^2+r(2).^2),'o',
'MarkerEdgeColor','k','MarkerFaceColor','w','MarkerSize',8);
end


for i=1:2
M(i,:)=M(i,:)- mean(M(i,:));
end

SIGMA=(1/99)*M*M';

[V,D]=eig(SIGMA);
t1=linspace(-3,3,100);
t2=linspace(-1.5,1.5,100);

plot3(t1,(V(2,1)/V(1,1)).*t1,sin(2*t1).*sin((V(2,1)/V(1,1)).*t1)
.*sqrt(t1.^2+((V(2,1)/V(1,1)).*t1).^2),'-r','LineWidth',3);
plot3(t2,(V(2,2)/V(1,2)).*t2,sin(2*t2).*sin((V(2,2)/V(1,2)).*t2)
.*sqrt(t2.^2+((V(2,2)/V(1,2)).*t2).^2),'-b','LineWidth',3);

%axis vis3d
axis off
camlight('headlight');
lighting phong;
set(gcf,'Color','w')
```

6. Figure 3.4

```
 [x,y,z]=sphere(100);
 colormap(gray)
 surf((1/2)*x,(1/2)*y,(1/2)*z,'EdgeColor','none')
hold on
axis vis3d
axis auto
axis off
%camlight('headlight');
%lighting gouraud;
set(gcf,'Color','w')

[x1,y1,z1]=sphere(20);
surf((1/2+0.023)*x1,(1/2+0.023)*y1,(1/2+0.023)*z1,'EdgeColor',[1 .5 0],
'FaceColor','none','LineWidth',1,'Linestyle','-')
```

```
k=0;
ratio=1/12;
 plot3((1/2+0.023)*x1(:,1),(1/2+0.023)*y1(:,1),
 (1/2+0.023)*z1(:,1),'-b','LineWidth',3)
 plot3((1/2+0.023)*x1(:,11),(1/2+0.023)*y1(:,11),
 (1/2+0.023)*z1(:,11),'-b','LineWidth',3)
 plot3((1/2+0.023)*x1(:,6),(1/2+0.023)*y1(:,6),
 (1/2+0.023)*z1(:,6),'-r','LineWidth',3)
 plot3((1/2+0.023)*x1(:,16),(1/2+0.023)*y1(:,16),
 (1/2+0.023)*z1(:,16),'-r','LineWidth',3)
 t=linspace(0,2*pi,100);
 e1=(1/2+0.023)*cos(t);
 e2=(1/2+0.023)*sin(t);
 e3=zeros(length(t));
 plot3(e1,e2,e3,'-g','LineWidth',3)



 for j=1:size(x1,2)
     for i=1:size(x1,1)
         k=k+1;
         r=1;
         phi= acos(z1(i,j)/r);
         if y1(i,j)>=0
         theta=acos(x1(i,j)/sqrt(x1(i,j)^2+y1(i,j)^2));
         else
         theta=2*pi-acos(x1(i,j)/sqrt(x1(i,j)^2+y1(i,j)^2));
         end
         zx=(1/2)*tan((pi-phi)/2)*cos(theta);
         zy=(1/2)*tan((pi-phi)/2)*sin(theta);

         %new basis
         X=-sin(theta)*[1;0;0]+cos(theta)*[0;1;0];
         Z=[x1(i,j);y1(i,j);z1(i,j)];
         Z=Z/norm(Z);
         Y=cross(X,Z);

         triad_x1=(1/2+0.023)*[x1(i,j);y1(i,j);z1(i,j)]-ratio*X;
         triad_x2=(1/2+0.023)*[x1(i,j);y1(i,j);z1(i,j)]+ratio*X;
         triad_x3=(1/2+0.023)*[x1(i,j);y1(i,j);z1(i,j)]+
         ratio*zx*X+ratio*zy*Y;


        %subplot(11,11,k)
        % fill(triad_new(1,:),triad_new(2,:),'green');
         fill3([triad_x1(1),triad_x2(1),triad_x3(1)],[triad_x1(2),
```

```
        triad_x2(2),triad_x3(2)],[triad_x1(3),triad_x2(3),
        triad_x3(3)],'yellow');


    end
 end
```

7. Figures 3.5 and 3.6

```
figure(1)
[x,y,z]=sphere(100);
 colormap(gray)
 surf((1/2)*x,(1/2)*y,(1/2)*z,'EdgeColor','none')
hold on
axis vis3d
axis auto
axis off
%camlight('headlight');
%lighting gouraud;
set(gcf,'Color','w')

mu = [1/4;1/6];
Sigma = [0.009 .0003; .0006 0.009]; R = chol(Sigma);
Color=hsv(49);

for i=1:49
r = mu + R*randn(2,1);

if r(1)^2+r(2)^2<1/4

M(:,i)=[r;sqrt(1/4-r(1)^2-r(2)^2)];

plot3(r(1),r(2),sqrt(1/4-r(1)^2-r(2)^2+0.01),'o','MarkerEdgeColor',
'k','MarkerFaceColor',Color(i,:),'MarkerSize',8);
end
end
%Visualization
figure(2)
set(gcf,'Color','w')


 for j=1:49
        %k=mod(j,5)+1;
        r=1/2;
```

```
          phi= acos(M(3,j)/r);
          if M(2,j)>=0
          theta=acos(M(1,j)/sqrt(M(1,j)^2+M(2,j)^2));
          else
          theta=2*pi-acos(M(1,j)/sqrt(M(1,j)^2+M(2,j)^2));
          end
          zx=(1/2)*tan((pi-phi)/2)*cos(theta);
          zy=(1/2)*tan((pi-phi)/2)*sin(theta);


          subplot(7,7,j)

          fill([-1,1,zx],[0,0,zy],Color(j,:));
          axis square
          axis off


 end
 figure(4)
 set(gcf,'Color','w')
sumx=0;
sumy=0;
Max_x=0;
Min_x=0;
Max_y=0;
Min_y=0;
 for j=1:49
          %k=mod(j,5)+1;
          r=1/2;
          phi= acos(M(3,j)/r);
          if M(2,j)>=0
          theta=acos(M(1,j)/sqrt(M(1,j)^2+M(2,j)^2));
          else
          theta=2*pi-acos(M(1,j)/sqrt(M(1,j)^2+M(2,j)^2));
          end
          zx=(1/2)*tan((pi-phi)/2)*cos(theta);
          zy=(1/2)*tan((pi-phi)/2)*sin(theta);
          sumx=sumx+zx;
          sumy=sumy+zy;
          plot(zx,zy,'o','MarkerEdgeColor',
          'k','MarkerFaceColor',Color(j,:),'MarkerSize',8);
          hold on
          axis square
          Max_x=max(Max_x,zx);
          Min_x=min(Min_x,zx);
```

```
          Max_y=max(Max_y,zy);
          Min_y=min(Min_y,zy);



 end

 sumx=(1/49)*sumx;
 sumy=(1/49)*sumy;
 plot(sumx,sumy,'o','MarkerEdgeColor',
 'k','MarkerFaceColor','w','MarkerSize',15);

 %Extrinsic mean :
 syms u v
NORM=(u-M(1,:)).^2+(v-M(2,:)).^2+(sqrt(1/4-u.^2-v.^2)-M(3,:)).^2;
g=(1/(2*length(NORM)))*sum(sum(NORM,1),2);
grad_g=gradient(g,[u,v]);
figure(6)
colormap(gray)
 surf((1/2)*x,(1/2)*y,(1/2)*z,'EdgeColor','none')
hold on
axis vis3d
axis auto
axis off
set(gcf,'Color','w')
for i=1:49
plot3(M(1,i),M(2,i),sqrt(1/4-M(1,i)^2-M(2,i)^2+0.02),'o',
'MarkerEdgeColor','b','MarkerFaceColor',Color(i,:),'MarkerSize',8);
end
[C,I]=max(-M(2,:));

MU_1(1)=M(1,10);
MU_2(1)=M(2,10);
mu=[M(1,10);M(2,10)];
tau=0.02;
for i=1:100
    evaluate=subs(grad_g,[u,v],[mu(1),mu(2)]);
    r=-evaluate;
    mu=mu+tau*r;
    MU_1(i+1)=mu(1);
    MU_2(i+1)=mu(2);
end
MU_Z= sqrt(1/4-MU_1.^2-MU_2.^2+0.02);
plot3(MU_1,MU_2,MU_Z,'Color',[1 .5 0],'LineWidth',3);
plot3(MU_1(1),MU_2(1),MU_Z(1),'o','MarkerEdgeColor','k',
'MarkerFaceColor','w','MarkerSize',15)
```

```
plot3(MU_1(length(MU_1)),MU_2(length(MU_1)),MU_Z(length(MU_Z)),'o',
'MarkerEdgeColor',[1 .5 0],'MarkerFaceColor','w','MarkerSize',15)
Color2=jet(11);
t=linspace(0,2*pi,30);
j=0;
for i=0:0.02:0.2
    X=i*cos(t)+mu(1);
Y=i*sin(t)+mu(2);
Z_cont=sqrt(1/4-X.^2-Y.^2+0.02);
j=j+1;
plot3(X,Y,Z_cont,'--','Color',Color2(j,:),'LineWidth',2);
end

figure(1)
plot3(mu(1),mu(2),sqrt(1/4-mu(1)^2-mu(2)^2+0.01),'o',
'MarkerEdgeColor',[1 .5 0],'MarkerFaceColor','w','MarkerSize',15)
```

```
 %New basis
 r=1/2;
 phi= acos(sqrt(1/4-mu(1)^2-mu(2)^2)/r);
   if mu(2)>=0
         theta=acos(mu(1)/sqrt(mu(1)^2+mu(2)^2));
   else
         theta=2*pi-acos(mu(1)/sqrt(mu(1)^2+mu(2)^2));
         end
 figure(4)
  zx=(1/2)*tan((pi-phi)/2)*cos(theta);
  zy=(1/2)*tan((pi-phi)/2)*sin(theta);
 plot(zx,zy,'o','MarkerEdgeColor',[1 .5 0],'MarkerFaceColor',
 'w','MarkerSize',15);


 er=cos(theta)*[1;0;0]+sin(theta)*[0;1;0];
 etheta=-sin(theta)*[1;0;0]+cos(theta)*[0;1;0];
 P=[er etheta [0;0;1]];
 Rot=[cos(-phi) 0 sin(-phi); 0 1 0; -sin(-phi) 0 cos(-phi)];
 figure(3)
 [x,y,z]=sphere(100);
 colormap(gray)
 surf((1/2)*x,(1/2)*y,(1/2)*z,'EdgeColor','none')
hold on
axis vis3d
axis auto
```

```
axis off
%camlight('headlight');
%lighting gouraud;
set(gcf,'Color','w')
 for i=1:49
     N(:,i)=P*Rot*P'*M(:,i);
     plot3(N(1,i),N(2,i),N(3,i),'o','MarkerEdgeColor','k',
     'MarkerFaceColor',Color(i,:),'MarkerSize',8);
 end

 sum(N(1:2,:),2)

for i=1:2
N(i,:)=N(i,:)- mean(N(i,:));
end

SIGMA=(1/48)*N(1:2,:)*N(1:2,:)';

[V,D]=eig(SIGMA);
[V,D]=sortem(V,D);

t1=linspace(-1/4,1/4,100);
t2=linspace(-1/4,1/4,100);

plot3(t1,(V(2,1)/V(1,1)).*t1,
sqrt(1/4-t1.^2-((V(2,1)/V(1,1)).*t1).^2),'-r','LineWidth',3);
plot3(t2,(V(2,2)/V(1,2)).*t2,
sqrt(1/4-t1.^2-((V(2,2)/V(1,2)).*t2).^2),'-b','LineWidth',3);

figure(1)
PGA1=[t1',(V(2,1)/V(1,1)).*t1',
sqrt(1/4-t1'.^2-((V(2,1)/V(1,1)+0.01).*t1').^2)];
PGA2=[t2',(V(2,2)/V(1,2)).*t2',
sqrt(1/4-t1'.^2-((V(2,2)/V(1,2)+0.01).*t2').^2)];

Rot2=[cos(phi) 0 sin(phi); 0 1 0; -sin(phi) 0 cos(phi)];
 for i=1:length(t1)

     PGA1(i,:)=P*Rot2*P'*PGA1(i,:)';
     PGA2(i,:)=P*Rot2*P'*PGA2(i,:)';
 end

plot3(PGA1(:,1),PGA1(:,2),PGA1(:,3),'-r','LineWidth',3);
plot3(PGA2(:,1),PGA2(:,2),PGA2(:,3),'-b','LineWidth',3);
```

```
figure(4)
 for j=1:size(PGA1,1)

         r=1/2;
         phi= acos(PGA1(j,3)/r);
         if PGA1(j,2)>=0
         theta=acos(PGA1(j,1)/sqrt(PGA1(j,1)^2+PGA1(j,2)^2));
         else
         theta=2*pi-acos(PGA1(j,1)/sqrt(PGA1(j,1)^2+PGA1(j,2)^2));
         end
         Zx(j)=(1/2)*tan((pi-phi)/2)*cos(theta);
         Zy(j)=(1/2)*tan((pi-phi)/2)*sin(theta);




 end
  plot(Zx,Zy,'-r','LineWidth',2);
  xlim([Min_x-0.3;Max_x+0.3]);
  ylim([Min_y-0.3;Max_y+0.3]);

         hold on
         axis square
 for j=1:size(PGA2,1)

         r=1/2;
         phi= acos(PGA2(j,3)/r);
         if PGA2(j,2)>=0
         theta=acos(PGA2(j,1)/sqrt(PGA2(j,1)^2+PGA2(j,2)^2));
         else
         theta=2*pi-acos(PGA2(j,1)/sqrt(PGA2(j,1)^2+PGA2(j,2)^2));
         end
         Zx1(j)=(1/2)*tan((pi-phi)/2)*cos(theta);
         Zy1(j)=(1/2)*tan((pi-phi)/2)*sin(theta);




 end
 plot(Zx1,Zy1,'-b','LineWidth',2);

 figure(5)
 plot(Zx,Zy,'-r','LineWidth',2);
```

```
xlim([Min_x-0.3;Max_x+0.3]);
 ylim([Min_y-0.3;Max_y+0.3]);
        hold on
        axis square
        axis on
        set(gcf,'Color','w')

for j=1:5:length(Zx)/2

        phi= acos(PGA1(j,3)/r);
        if PGA1(j,2)>=0
        theta=acos(PGA1(j,1)/sqrt(PGA1(j,1)^2+PGA1(j,2)^2));
        else
        theta=2*pi-acos(PGA1(j,1)/sqrt(PGA1(j,1)^2+PGA1(j,2)^2));
        end
        zx=(1/2)*tan((pi-phi)/2)*cos(theta);
        zy=(1/2)*tan((pi-phi)/2)*sin(theta);

      fill([-1/6+Zx(j),1/6+Zx(j),(1/6)*zx+Zx(j)],
      [+Zy(j),+Zy(j),(1/6)*zy+Zy(j)],Color(mod(j,49)+1,:));
        axis square




end
for j=length(Zx)/2+5:10:length(Zx)

        phi= acos(PGA1(j,3)/r);
        if PGA1(j,2)>=0
        theta=acos(PGA1(j,1)/sqrt(PGA1(j,1)^2+PGA1(j,2)^2));
        else
        theta=2*pi-acos(PGA1(j,1)/sqrt(PGA1(j,1)^2+PGA1(j,2)^2));
        end
        zx=(1/2)*tan((pi-phi)/2)*cos(theta);
        zy=(1/2)*tan((pi-phi)/2)*sin(theta);

      fill([-1/6+Zx(j),1/6+Zx(j),(1/6)*zx+Zx(j)],
      [+Zy(j),+Zy(j),(1/6)*zy+Zy(j)],Color(mod(j,49)+1,:));
        axis square

end
```