FRÉCHET MEANS WITH RESPECT TO THE RIEMANNIAN DISTANCES: EVALUATIONS AND APPLICATIONS

FRÉCHET MEANS WITH RESPECT TO THE RIEMANNIAN DISTANCES: EVALUATIONS AND APPLICATIONS

BY

MEHDI RAZEGHI JAHROMI, M.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

Master of Applied Science (2014) McMaster University

(Electrical & Computer Engineering) Hamilton, Ontario, Canada

TITLE:	FRÉCHET MEANS WITH RESPECT TO THE RIEMANNIAN DISTANCES: EVALUATIONS AND APPLICATIONS
AUTHOR:	Mehdi Razeghi Jahromi M.Sc., (Control and System Engineering) University of Sheffield, Sheffield, UK
SUPERVISOR:	Prof. Kon Max Wong
NUMBER OF PAGES:	xii, 103

To my family

Abstract

The space of symmetric positive definite matrices forms a manifold with an ambient is Euclidean space. In order to measure the distances between the objects on this manifold several metrics have been proposed. In this work we study the concept of averaging over the elements of the manifold by using the notion of Fréchet mean. The main advantageous of this method is its connection to the metrics as a result of which we can utilize the Reimannian distances to obtain the mean of positive definite matrices. We consider three Reimannian metrics which have been developed on the manifold of symmetric positive definite matrices. The methods of obtaining the Fréchet mean in the case of each metric will be discussed. The performance of each estimator will be demonstrated by using models based on matrix Cholesky factor, matrix square root and matrix logarithm. The deviation from the nominal covariance in each case will be evaluated using loss function, Euclidean distance and root Euclidean distance. We will see that depending on the model under investigation, Fréchet mean of Reimannian distances performs better in most of the cases.

In terms of application, we analyse the performance of each Fréchet mean estimator in a classification task. For this purpose we evaluate the method of distance to the center of

mass using the simulated data. This method will also be applied on the high content cell image data set in order to classify the cells with respect to the type of treatment that has been used.

Acknowledgements

First I would like to thank my supervisor, Dr K.M.Wong for supporting me, encouraging me and giving me the freedom in research. I am deeply grateful to him for the long discussions that helped me sort out the technical details of my work, encouraging the use of correct grammar and consistent notation in my writings and for carefully reading and commenting on countless revisions of this manuscript. He taught me how good research and academic writing should look like. I have to also express my thanks to my co supervisor, Dr Alexandar Jeremic. He has been always available to listen and give valuable advice. He helped me regarding the processing of the real data and reminded me important things on writing my thesis. I need to acknowledge Dr McKenzie Wang for many grateful discussions that really helped me to have better understanding of differential geometry and the notion of metric. I am grateful of Dr David Andrews and members of his lab for providing me with the real data set and their helpful suggestions. I am indebted to my parents and my brother Kian because they were constantly there to help me when I did not think I could continue. Also I am thankful of the administrative members of the department of Electrical and Computer Engineering, McMaster university, especially Ms. Cheryl Gies for her countless help and support.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation and literature review

In statistical signal processing, the second order statistics of the the received signal from M-array of observation is more informative. For example, the covariance matrix is being used in classification to facilitate finding the type of object of interest; in estimation tasks it is used to find the location of the object, furthermore in detection we would like to know whether an object exists or not.

In bio-medical imaging a particular type of covariance matrix often arises which is known as diffusion tensor. The diffusion tensor is a 3×3 covariance matrix. It is obtained by fitting a model on the Fourier transform of the molecule displacement density function [4].

In field of bio-medical data processing such as EEG sleep state classification [5] the PSD matrix of the brain signal is studied for classification of sleep state (depending on the

frequency of each sleep state). In brain computer interface(BCI) [3] in order to facilitate the communication or movement of people with severe motor disorders the measured signals are used to form the covariance matrix and use it as a feature so called EEG signal descriptors.

In face recognition and human detection [6] for the sets of given images the covariance of specific regions in image is obtained with respect to the features of the region the covariance is then used for target tracking or classification of images.

As we can see in all of the above applications covariance matrices or power spectral density matrices play the key role in classification and detection. As far as the covariance matrix is concerned, they have specific structures namely symmetry and being positive definite. As a result the space of covariance matrices or equivalently positive definite matrices are form a subspace which is known as a "Riemannian manifold". This space lies in the space of all Hermitian matrices.

One important task in such a subspace is how to measure the distance in appropriate manner in which it reflects more accuracy and take to the account the curvature of the manifold in measuring the distance between symmetric positive definite matrices. To answer to this question it is needed to study different metrics on the manifold of symmetric positive definite matrices.

Intuitively, measuring the distance on the manifold is similar to measuring the distance between two cities. One can measure the distance between two cities as the straight line connecting them. However, in reality such measurement is not accurate since the road must traverse the mountains or the rivers.

Several metrics have been proposed to measure the distance between two positive definite matrices. The one which is widely used is based on projecting the points to the tangent space and measuring the distance in tangant space using metric which is known in literature as "Log-Riemannian" distance [7] [8].

Recently other metrics have been introduced to measure the distances on the manifold of symmetric positive definite matrices. They are based on finding the horizontal Euclidean subspace which is isometric with the tangent space of the space of symmetric positive definite matrices [5]. The advantage of this approach is that we no longer need to deal with a complicated formula to find the distance between the points on the manifold; rather we have a closed form solution in hand for the metrics that are developed on horizontal subspace.

1.2 Arithmetic and geometric mean

Signal processing often involves the evaluation of the mean of a collection of the signal features. Signal processing using the covariance matrix is no different; we often have to deal with a population of positive definite matrices resulting from several measurements and have to find the mean of them. The conventional mean that is widely used is the arithmetic mean or equivalently center of mass [7] of a group of symmetric positive definite matrices matrices. We will show that this definition is in fact related to the Euclidean distance or Frobenius norm . However, as we mentioned earlier, the space of symmetric positive definite matrices form a manifold so the natural question would be if there exists a systematic way to evaluate the mean of symmetric positive definite matrices with respect

to Riemannian distances.

The first step towards the definition of geometric mean was reflected in works of Maurice Fréchet who suggested that the mean of random variable in arbitrarily metric space is the point which minimizes the expected value of sum of squared distance. His work is followed by Hermann Karcher [9]. Recently it has been suggested that such definition can be used on the space of symmetric positive definite matrices [7, 10, 11].

The Fréchet mean of positive definite matrices recently finds its way in radar target tracking [8]. In image processing it has been used for pedestrian detection [6]. In this work our attention is to find the metric based mean of symmetric positive definite Hermitian matrices. Our approach is not only with respect to the metric "log-Riemannian" but also we will use the advantage of horizontal lift subspace and its resulting metrics in finding the mean of a population of covariance matrices.

1.3 Outline and contribution of the work

Our research is organized as follows. In Chapter 2 we will review the properties of the sample covariance matrix since it is being widely used in signal processing. We will employ the loss function to measure the closeness of it to symmetric positive definite matrices. It will be shown that the sample covariance matrix is the minimizer of the risk function. The arithmetic mean of symmetric positive definite matrices are also studied in this chapter. Finally we discuss the properties of the space of symmetric positive definite matrices.

In Chapter 3 we will discuss the concept of Fréchet mean in Reimanian manifold

of symmetric positive definite matrices, as we mentioned earlier, this method is metric based approach. As a result, it provides us with a reasonable connection between the metric and its corresponding mean. We will see that for the case of having two covariance matrices we have a closed form solution for the geometric mean of the covariance matrices with respect to the "log-Riemannian" metric. However, when we have more than two covariance matrices, the Fréchet mean will facilitate finding the mean of more than two covariance matrices.

On the other hand, we also consider the metrics which are obtained by mapping the points on the manifold to the horizontal subspace through the fibers. We will see that since the distance between two points on the horizontal subspace is straight line it enables us to find the mean on the horizontal subspace and project it back to the manifold.

Unlike the Fréchet mean using the log- Riemannian metric, for one of the metrics which is developed on horizontally subspace we have derived the closed form solution for the Fréchet mean which gives the remarkable result in terms of accuracy in comparison to the Fréchet means of other metrics.

Chapter 4 is devoted to the numerical results. In this chapter we validate our derivation of Fréchet mean of different metrics. Basically we measure the distance between actual and estimated mean of symmetric positive definite matrices via several simulation runs (Monte-Carlo) and measure the error of each estimator depending on the type of metric which has been used to estimate the covariance matrix.

We apply the techniques that we have developed in finding the mean of population of covariance matrices in classification of cell imaging data set using the technique which is based on the distance to the center of mass of each cluster. The method is also performed

on the simulated data set for validation purpose.

Chapter 2

Sample covariance and arithmetic mean

In array signal processing applications as well as data classification, the estimation of covariance matrix is important . The conventional (and also classical) method for covariance matrix estimation is the well-known sample covariance matrix (SCM). Consider a complex stationary random data vector $\mathbf{x}$ of dimension M. We define the mean and the covariance matrix of $\mathbf{x}$ as

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}] \tag{2.1a}$$

$$\boldsymbol{\Sigma} = \mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^H] \tag{2.1b}$$

where $\mathbb{E}[\cdot]$ denotes the expected value, and $(\cdot)^H$ denotes the Hermitian conjugate of a vector or matrix. We note that $\boldsymbol{\Sigma}$ is Hermitian symmetric and positive definite. The values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be evaluated if the data is stationary and there is an infinite number of data vectors.

For finite data, i.e., $\mathbf{X} = [\mathbf{x}(1), \mathbf{x}(2), ..., \mathbf{x}(N_0)]$, the mean vector and the covariance matrix are usually estimated by averaging over the N_0 data vectors available arriving at the sample mean and the sample covariance such that

$$\hat{\boldsymbol{\mu}} = \frac{1}{N_0}\sum_{k=1}^{N_0}\mathbf{x}(k) \tag{2.2a}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N_0}\sum_{k=1}^{N_0}(\mathbf{x}(k)-\hat{\boldsymbol{\mu}})(\mathbf{x}(k)-\hat{\boldsymbol{\mu}})^H = \frac{1}{N_0}\mathbf{S} \tag{2.2b}$$

Eq.(2.2) are the arithmetic average of the samples and the sample covariance matrices. It can be shown [12] that if $\mathbf{x}(1), \mathbf{x}(2), ..., \mathbf{x}(N_0)$ are independent normally distributed random vectors each has size $1 \times M$ (where M represents the number of sensors) with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, Eqs. (2.2) are respectively the maximum likelihood estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ [13].

To show this fact we can form the likelihood function for $\mathbf{x}(1), \mathbf{x}(2), ..., \mathbf{x}(N_0)$ as:[1]

$$\begin{aligned} L_{\boldsymbol{\mu},\boldsymbol{\Sigma}}\left(\mathbf{x}(1), \mathbf{x}(2), ..., \mathbf{x}(N_0)\right) &= \frac{1}{(\pi)^{N_0 M}\det\left(\boldsymbol{\Sigma}\right)^{N_0}} \\ &\times \exp\left(-\sum_{k=1}^{N_0}\left(\mathbf{x}(k)-\boldsymbol{\mu}\right)^H\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}(k)-\boldsymbol{\mu}\right)\right) \end{aligned} \tag{2.3}$$

[1] We note that observation $\mathbf{x}$ of size $1 \times M$ has the probability density function $f(\mathbf{x}) = \frac{1}{(\pi)^M \det(\Sigma)} \times \exp\left(-\left(\mathbf{x}-\boldsymbol{\mu}\right)^H\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right)$ [14].

Since we have

$$\begin{aligned}
\sum_{k=1}^{N_0} (\mathbf{x}(k)-\boldsymbol{\mu})^H \boldsymbol{\Sigma}^{-1} (\mathbf{x}(k)-\boldsymbol{\mu}) &= \sum_{k=1}^{N_0} (\mathbf{x}(k)-\hat{\boldsymbol{\mu}}+\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}^{-1} (\mathbf{x}(k)-\hat{\boldsymbol{\mu}}+\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}) \\
&= \sum_{k=1}^{N_0} ((\mathbf{x}(k)-\hat{\boldsymbol{\mu}})+(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}))^H \boldsymbol{\Sigma}^{-1} ((\mathbf{x}(k)-\hat{\boldsymbol{\mu}})+(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})) \\
&= \sum_{k=1}^{N_0} (\mathbf{x}(k)-\hat{\boldsymbol{\mu}})^H \boldsymbol{\Sigma}^{-1} (\mathbf{x}(k)-\hat{\boldsymbol{\mu}}) \\
&+ \sum_{k=1}^{N_0} (\mathbf{x}(k)-\hat{\boldsymbol{\mu}})^H \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}) \qquad (2.4) \\
&+ \sum_{k=1}^{N_0} (\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}^{-1} (\mathbf{x}(k)-\hat{\boldsymbol{\mu}}) \qquad (2.5) \\
&+ \sum_{k=1}^{N_0} (\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}) \\
& \qquad (2.6)
\end{aligned}$$

the terms(2.4) and (2.5) in above equation are equal to zero. As a result we have:

$$\begin{aligned}
&\sum_{k=1}^{N_0} (\mathbf{x}(k)-\boldsymbol{\mu})^H \boldsymbol{\Sigma}^{-1} (\mathbf{x}(k)-\boldsymbol{\mu}) = \\
&\qquad \sum_{k=1}^{N_0} (\mathbf{x}(k)-\hat{\boldsymbol{\mu}})^H \boldsymbol{\Sigma}^{-1} (\mathbf{x}(k)-\hat{\boldsymbol{\mu}}) + N_0 (\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}) \qquad (2.7)
\end{aligned}$$

We use the result of Eq.(2.7) to simplify equation(2.3)to:

$$\begin{aligned} L_{\boldsymbol{\mu},\boldsymbol{\Sigma}}\left(\mathbf{x}(1),\mathbf{x}(2),...,\mathbf{x}(N_0)\right) &= \frac{1}{(\pi)^{N_0 M}\det\left(\Sigma\right)^{N_0}} \\ &\times \exp\left(-\operatorname{Tr}\boldsymbol{\Sigma}^{-1}\mathbf{S}\right) \\ &\times \exp\left(-N_0\left(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\right)^H\boldsymbol{\Sigma}^{-1}\left(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\right)\right) \end{aligned} \tag{2.8}$$

by considering the logliklihood function of $L_{\boldsymbol{\mu},\boldsymbol{\Sigma}}\left(\mathbf{x}(1),\mathbf{x}(2),...,\mathbf{x}(N_0)\right)$ we will have:

$$\begin{aligned} l_{\boldsymbol{\mu},\boldsymbol{\Sigma}}\left(\mathbf{x}(1),\mathbf{x}(2),...,\mathbf{x}(N_0)\right) &= \log L_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \\ &= c - N_0\log\det\boldsymbol{\Sigma} - \operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right) \\ &- N_0\left(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\right)^H\boldsymbol{\Sigma}^{-1}\left(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\right) \end{aligned} \tag{2.9}$$

In Eq.(2.9) c is a constant. To maximize the log-likelihood equation we observe that the quadratic form in Eq.(2.9) is non negative because $\boldsymbol{\Sigma}$ is positive definite matirx and so is $\boldsymbol{\Sigma}^{-1}$. The quadratic form is equal to zero when $\boldsymbol{\mu}=\hat{\boldsymbol{\mu}}$. As a result $\hat{\boldsymbol{\mu}}$ is the ML estimator of $\boldsymbol{\mu}$. By substituting this value in Eq.(2.9) we need to maximize the following function with respect to $\boldsymbol{\Sigma}^{-1}$:

$$f\left(\boldsymbol{\Sigma}^{-1}\right) = c + N_0\log\det\boldsymbol{\Sigma}^{-1} - \operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right) \tag{2.10}$$

Using the rules for derivative of matrices [15] we will get :

$$\frac{\partial f\left(\boldsymbol{\Sigma}^{-1}\right)}{\partial\boldsymbol{\Sigma}^{-1}} = N_0\frac{1}{|\boldsymbol{\Sigma}^{-1}|}\frac{|\boldsymbol{\Sigma}^{-1}|}{\partial\boldsymbol{\Sigma}^{-1}} - \frac{\partial\operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right)}{\partial\boldsymbol{\Sigma}^{-1}} = N_0\boldsymbol{\Sigma}-\mathbf{S} \tag{2.11}$$

By setting the Eq.(2.11) equal to $\mathbf{0}$ we get the maximum likelihood estimator for covariance matrix $\mathbf{\Sigma}$ which is $\hat{\mathbf{\Sigma}} = \frac{1}{N_0}\mathbf{S}$.

2.1 Loss function

In order to have a bench mark to measure the discrepancy between two $M \times M$ symmetric positive definite matrices $\mathbf{A}$ and $\mathbf{B}$ there exist a well accepted non symmetric measure so called loss function [16]. It is defined as:

$$L\left(\mathbf{A}, \mathbf{B}\right) = \mathrm{Tr}\left(\mathbf{A}^{-1}\mathbf{B}\right) - \log\det\left(\mathbf{A}^{-1}\mathbf{B}\right) - M \tag{2.12}$$

It can be shown that under the loss function (2.12), $\mathbf{\Sigma}$ is estimated by $M \times M$ positive definite matrix $\phi(\mathbf{S})$ whose elements are functions of the elements of $\mathbf{S}$. To show this we consider the risk function $\bar{L}$ as the expected value of the loss function:

$$\begin{aligned}
\bar{L}(\mathbf{\Sigma}, \alpha S) &= \mathbb{E}\left[\alpha\,\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{S}\right) - \log\det\left(\alpha\mathbf{\Sigma}^{-1}\mathbf{S}\right) - M\right] \\
&= \alpha\,\mathrm{Tr}\,\mathbf{\Sigma^{-1}}\mathbb{E}(\mathbf{S}) - M\log\alpha - \mathbb{E}\left[\log\frac{\det\mathbf{S}}{\det\mathbf{\Sigma}}\right] - M \\
&= \alpha M(N_0 - 1) - M\log\alpha - \mathbb{E}\left[\log\prod_{k=1}^{M}\chi^2_{N_0-i}\right] - M \\
&= \alpha M(N_0 - 1) - M\log\alpha - \sum_{i=1}^{M}\mathbb{E}\left[\log\chi^2_{N_0-i}\right] - M
\end{aligned} \tag{2.13}$$

By taking the derivative of the Eq.(2.13) with respect to α and set it equal to zero we have $\alpha = \frac{1}{N_0-1}$.

Overall, we can find that $\hat{\mathbf{\Sigma}}$ is the optimum estimation of $\mathbf{\Sigma}$ when the estimation has form of $\alpha\mathbf{S}$ when $\alpha = \frac{1}{N_0-1}$. Furthermore, we have

$$\mathbb{E}\left(\frac{1}{N_0-1}\mathbf{S}\right) = \mathbf{\Sigma} \tag{2.14}$$

which means $\frac{1}{N_0-1}\mathbf{S}$ is an unbiased estimator for $\mathbf{\Sigma}$ [17].

2.2 Arithmetic Mean of $M \times M$ Hermitian matrices

From the least squares point of view, the arithmetic mean covariance matrix in Eq. (2.2b) can also be viewed as the "centroid" of all the observed sample covariance, i.e., suppose the set of IID M-dimensional observed vectors $\mathbf{x}(1), \cdots, \mathbf{x}(N_0)$ are divided into N groups each containing at least $M_0 \geq M$ observed vectors. For each group $\{\mathbf{x}_n(1), \cdots, \mathbf{x}_n(M_0), n = 1, \cdots, N\}$, we form the observed mean $\boldsymbol{\mu}_n = \frac{1}{M_0}\sum_{i=1}^{M_0}\mathbf{x}_n(i)$, and the observed covariance matrices $\mathbf{S}_n = \frac{1}{M_0}\sum_{i=1}^{M_0}[\mathbf{x}_n(i) - \boldsymbol{\mu}_n][\mathbf{x}_n(i) - \boldsymbol{\mu}_n]^H$.

Before we show that the arithmetic mean of $\{\mathbf{S}_n\}_{n=1}^N$ is the minimizer of the sum of the square distances we need to consider the directional derivative of a function (see Appendix A) in Euclidean space according to the natural inner product in such space.

Lemma 2.2.1. *[18] : Let ϕ_1 and ϕ_2 be continuously differentiable real valued functions on the interval $(0, \infty)$ and let*

$$h(\mathbf{X}) = \langle \phi_1(\mathbf{X}), \phi_2(\mathbf{X}) \rangle = \operatorname{Tr} \phi_1(\mathbf{X})\phi_2(\mathbf{X}) \tag{2.15}$$

For all Hermitian matrix $\mathbf{X}$. Then the directional derivative of h is given by the formula:

$$D_Y h(\mathbf{X}) = \left\langle \phi_1^{'}(\mathbf{X})\phi_2(\mathbf{X}) + \phi_1(\mathbf{X})\phi_2^{'}(\mathbf{X}), \mathbf{Y} \right\rangle \tag{2.16}$$

If we define the square of the *Frobenius distance* between two matrices $\mathbf{A}$ and $\mathbf{B}$ as

$$d_{\mathrm{F}}^2(\mathbf{A}, \mathbf{B}) = \|\mathbf{A} - \mathbf{B}\|_2^2 = \mathrm{Tr}(\mathbf{A} - \mathbf{B})(\mathbf{A} - \mathbf{B})^H \tag{2.17}$$

then the sample mean covariance given by Eq.(2.2b) is the matrix which minimizes

$$\hat{\mathbf{\Sigma}}_{\mathrm{A}} = \arg\min_{\mathbf{\Sigma}} \sum_{n=1}^{N} d_{\mathrm{F}}^2(\mathbf{S}_n, \mathbf{\Sigma}) \tag{2.18}$$

That Eq.(2.18) is indeed true can be shown using the directional derivative expression:

Let

$$F(\mathbf{\Sigma}) = \sum_{n=1}^{N} \langle \mathbf{\Sigma} - \mathbf{S}_n, \mathbf{\Sigma} - \mathbf{S}_n \rangle = \sum_{n=1}^{N} \|\mathbf{\Sigma} - \mathbf{S}_n\|_2^2 \tag{2.19}$$

From Eq.(2.19) and using lemma (2.2.1) the directional derivative of the function F is given by

$$D_{\mathbf{R}} F(\mathbf{\Sigma}) = \sum_{n=1}^{N} \langle 2(\mathbf{\Sigma} - \mathbf{S}_n), \mathbf{R} \rangle \tag{2.20}$$

Letting $D_{\mathbf{R}} F(\mathbf{\Sigma}) = 0$ we have $\hat{\mathbf{\Sigma}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{S}_n$. As mentioned $\hat{\mathbf{\Sigma}}$ is the ML estimator of

$\boldsymbol{\Sigma}$. The expected value of this estimator is equal to

$$\begin{aligned}
\mathbb{E}\left(\hat{\boldsymbol{\Sigma}}\right) &= \mathbb{E}\left(\frac{1}{N}\sum_{n=1}^{N}\mathbf{S}_n\right) \\
&= \frac{1}{N}\mathbb{E}\left(\sum_{i=1}^{N}\mathbf{S}_n\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(\mathbf{S}_n\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{M_0-1}{M_0}\boldsymbol{\Sigma} = \frac{M_0-1}{M_0}\boldsymbol{\Sigma}
\end{aligned} \tag{2.21}$$

Eq.(2.21)shows that $\hat{\boldsymbol{\Sigma}}$ is the biased estimator of $\boldsymbol{\Sigma}$. However, as M_0 tends to infinity, $\hat{\boldsymbol{\Sigma}}$ converges to $\boldsymbol{\Sigma}$ [19].

From Eq.(2.21) we can conclude that the expected value of the estimation $\boldsymbol{\Sigma}$ only depends on the size of each group not number of groups.

2.3 Geometry of $M \times M$ Covariance Matrix

So far we have studied the properties of the sample covariance matrices; However we need to take to the account the nature of the space that such matrices exist on it. Since the covariance matrix of a vector is positive definite Hermitian, let us now denote the set of all the $M \times M$ complex matrices by $\mathcal{H}$, the set of all the $M \times M$ Hermitian matrices by $\mathcal{H}_H$, and the set of all positive definite Hermitian matrices by $\mathcal{M}$.

We first define the inner product of two $M \times M$ matrices $\mathbf{A}, \mathbf{B}$ as

$$\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{Tr}[\mathbf{A}\mathbf{B}^H] = \left(\sum_{m=1}^{M} \sum_{n=1}^{M} a_{mn} b_{mn}^* \right) \tag{2.22}$$

Where a_{mn} and b_{mn} are respectively the mnth elements of $\mathbf{A}$ and $\mathbf{B}$ respectively. We note that this is the sum of all the elements of the Hadamard (element by element) product $\mathbf{A} \bigodot \mathbf{B}^H$. We further note that for $\mathbf{A}$ and $\mathbf{B}$ being Hermitian, $\mathrm{Tr}[\mathbf{A}\mathbf{B}^H] = \mathrm{Tr}[\mathbf{A}\mathbf{B}]$ and $\mathbf{A} \bigodot \mathbf{B}^H = \mathbf{A} \bigodot \mathbf{B}$.

The following is an important property of matrices in $\mathcal{M}$ [5].

Lemma 2.3.1. *$\mathcal{M}$ is a manifold in the* real *linear vector space $\mathcal{H}_{\mathrm{H}}$.*

Proof. From the definitions of $\mathcal{H}_{\mathrm{H}}$ and $\mathcal{M}$, we can write $\mathcal{H}_{\mathrm{H}} = \{\mathbf{H} \in \mathcal{H} : \mathbf{H}^H = \mathbf{H}\}$ and $\mathcal{M} = \{\mathbf{S} \in \mathcal{H}_{\mathrm{H}} : \mathbf{S} \succ 0\}$. From which we can establish that $\mathcal{M} \subset \mathcal{H}_{\mathrm{H}} \subset \mathcal{H}$. Given $\mathbf{S} \in \mathcal{M}$, if a small perturbation is set on the eigenvalues of $\mathbf{S}$, the resulting $\mathbf{S} + \delta\mathbf{S}$ remains in $\mathcal{M}$. Hence, $\mathcal{M}$ is an open subset of $\mathcal{H}_{\mathrm{H}}$ and is thus a manifold in $\mathcal{H}_{\mathrm{H}}$. That $\mathcal{H}_{\mathrm{H}}$ is real can be shown such that, for $\mathbf{H} \in \mathcal{H}_{\mathrm{H}}$ and a complex scalar c, $c\mathbf{H} \notin \mathcal{H}_{\mathrm{H}}$ in general since $c\mathbf{H}$ may no longer be Hermitian. Therefore, $\mathcal{H}_{\mathrm{H}}$ is closed *only for real scalar field*.

Furthermore, we can establish a set of basis $\{\tilde{\mathbf{E}}_{mn};\ m, n = 1, \cdots, M\}$

such that $\tilde{\mathbf{E}}_{mn}, m > n$, has only two non-zero elements of 1 at the mnth and nmth positions, $\tilde{\mathbf{E}}_{nm}, m < n$, has only two non-zero elements of j and $-j$ at the mnth and the nmth positions, and $\tilde{\mathbf{E}}_{mm}$ has only one non-zero element of 1 at the mmth position. Thus, any $\mathbf{H} \in \mathcal{H}_{\mathrm{H}}$ can be represented as a linear combination of the Hermitian basis set

$\{\tilde{\mathbf{E}}_{mn};\ \ m,n=1,\cdots,M\}$ such that

$$\mathbf{H}=\sum_{m=1}^{M}\sum_{n=1}^{M}\alpha_{mn}\tilde{\mathbf{E}}_{mn},\qquad \text{with } \tilde{\mathbf{E}}_{mn}=\tilde{\mathbf{E}}_{mn}^{H} \tag{2.23}$$

The coefficients α_{mn} are real since $\mathbf{H}-\mathbf{H}^H=\sum_{m=1}^{M}\sum_{n=1}^{M}(\alpha_{mn}-\alpha_{mn}^*)\tilde{\mathbf{E}}_{mn}=\mathbf{0}$ from which we conclude $\alpha_{mn}=\alpha_{mn}^*$. Thus, $\mathbf{H}$ can be represented as an $(M\times M)$-tuple $\{\alpha_{mn};m,n=1,\cdots,M\}$ in a *real* $(M\times M)$-dimensional space $\mathbb{R}^{(M\times M)}$.

Since the basis matrices $\{\tilde{\mathbf{E}}_{mn}\}$ in Eq. (2.23) are all orthonormal the inner product $\langle\mathbf{H}_1,\mathbf{H}_2\rangle$ in $\mathcal{H}_H$ as given by Eq. (2.22) is reduced to

$$\langle\mathbf{H}_1,\mathbf{H}_2\rangle=\sum_{i=1}^{M}\sum_{j=1}^{M}\alpha_{1ij}\alpha_{2ij} \tag{2.24}$$

which is also real. Henceforth, we refer to $\mathcal{H}_H$ as a *real* vector space. □

In order to demonstrate the manifold of symmetric positive definite matrices we consider an example of the space of 2×2 symmetric positive definite matrices with real entries. Each covariance matrix can be depicted as the point in three dimensional Euclidean space . Two coordinates show the variance and the third coordinate shows the covariance. Now the covariance matrices in this space have the positive elements on the main diagonal which shows the variance. On the other hand the off diagonal elements satisfy the Cauchy Schwartz inequality [20]:

$$|cov(x_1,x_2)|\leq\sqrt{(var(x_1))}\sqrt{(var(x_2))} \tag{2.25}$$

In Eq.(2.25), $cov(x_1,x_2)$ denotes the covariance between two variables x_1 and x_2. Also

'var' represents the variance. By definition of the space of 2×2 covariance matrices, each matrix has a representation in three dimensional space $\mathbb{R}^3$ with coordinates:

$$\mathbf{p} = [var(x_1), var(x_2), cov(x_1, x_2)]^T \tag{2.26}$$

To preserve positive definiteness property of the covariance matrix the point in Eq.(2.26) must be interior point of the region. On the other hand, Once the representative point touches the boundary of the figure (2.1), i.e when the equality holds in equation (2.25), the matrix is no longer positive definite.

The example illustrates that treating the space of positive definite matrices as a linear space is not accurate. So we need to consider the natural geometry of space of such structured matrices for further study.

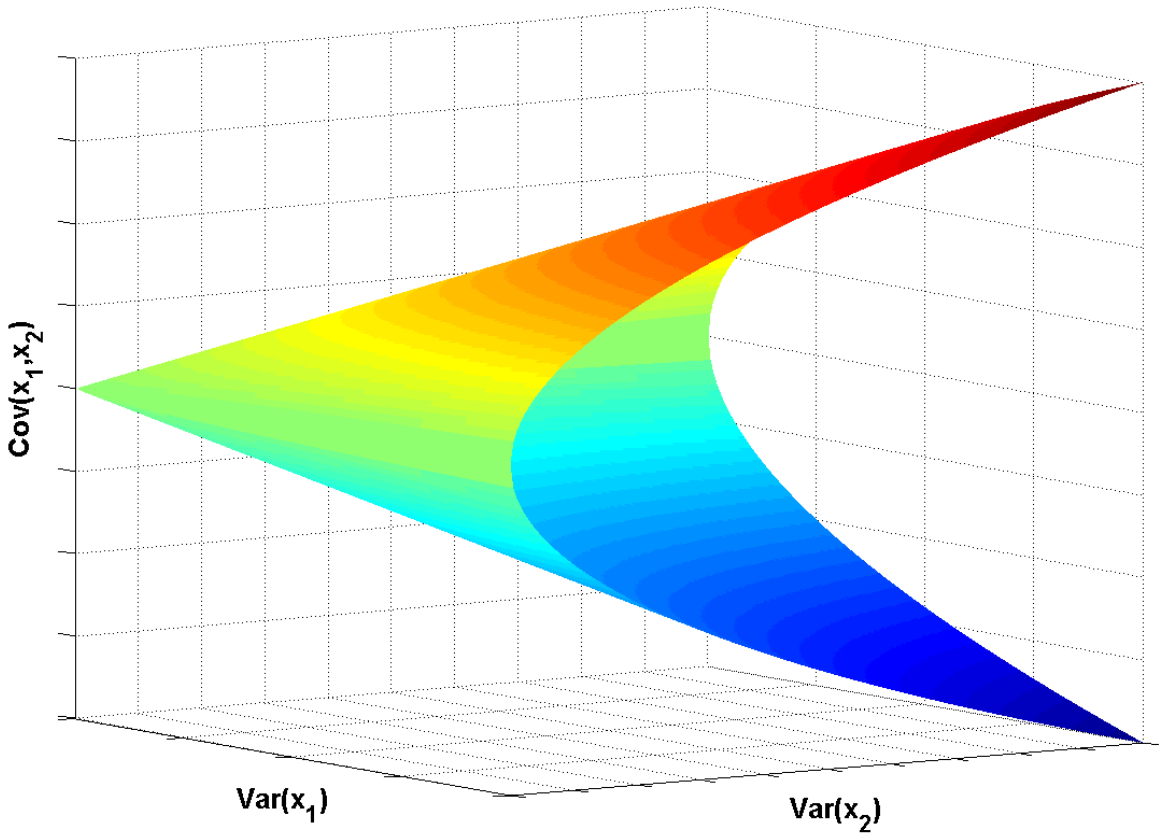


Figure 2.1: Visualization of space of 2×2 covariance matrices with real entries in three dimensional space.

Chapter 3

Fréchet mean of Hermitian positive definite matrices

3.1 Introduction

The covariance matrix, or equivalently, the power spectral density (PSD) matrix, of the signals from a multi-sensor system is a feature useful for many purposes in statistical signal processing including detection, estimation, classification, and signal design. In a recent paper [5],the importance of power spectral density matrix in classification of EEG signals was demonstrated.

In many applications of signal processing, the covariance matrix of the observed signal is utilized as a feature from which information is extracted. Often, for extraction of information, averaging and interpretation of these matrices are needed. To develop algorithms for such evaluations, one important fact has to be born in mind that the structural

constraints, i.e., Hermitian symmetry and positive definiteness, on such matrices must be maintained [5] [2].

It was suggested [5] that since such features form a *manifold* in the signal space due to their structural constraints, measurements of these features ought to be carried out *on* the manifold, and not as vectors in the signal space. It is further observed that if a curve on the covariance matrix feature manifold $\mathcal{M}$ is lifted along the *fibres* (special mapping connections) into the Euclidean signal space $\mathcal{H}$, then there exists a subspace $\mathcal{U} \subset \mathcal{H}$ isometric to the tangent space of the manifold to which the curve can be lifted. This implies that measurements on the feature manifold may be equivalently obtained from the measurement in the subspace $\mathcal{U}$. From this concept, using different Riemannian metrics and different mappings, various Riemannian distances on the feature manifold $\mathcal{M}$ can be derived.

In this chapter, the focus of our attention is on the estimation of the Frechet mean of the covariance matrices on the manifold $\mathcal{M}$ using the different measures of Riemannian distances.

The necessary frame work and algorithms to obtain the mean of covariance matrices from group of sample covariance matrices using Riemannian distances on manifold of positive definite matrices $\mathcal{M}$ will be developed and studied in this chapter.

First we will review the Riemannian distances which form the basis of measuring the Frechet mean. Then the Fréchet mean for each of the metrics will be obtained and discussed.

3.2 Fréchet mean

The history of defining mean goes back 2500 years when the ancient Greeks introduced ten types of different means. Among them only three of them survive and are still being used. These are the arithmetic, the geometric and the harmonic means.

We use the notion of Fréchet mean to unify the method of finding the mean of positive definite matrices. The Fréchet mean is given as the point which minimizes the sum of the squared distances [8]:

$$\hat{\mathbf{\Sigma}} = \underset{\Sigma \in \mathcal{M}}{\arg\min} \sum_{i=1}^{n} d^2(\mathbf{S}_i, \mathbf{\Sigma}) \tag{3.1}$$

where $\{\mathbf{S}_i\}_{i=1}^n$ represents the symmetric positive definite matrices and $d(.,.)$ denotes the metric being used respectively.

In fact if we have a closer look at the definition of arithmetic mean of positive measurement $\{x_i\}_{i=1}^n$, which is denoted as $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, and using the usual distance, we can see that it has the variational property. This means that it minimizes the sum of the squared distances to the points x_k :

$$\bar{x} = \underset{x \geq 0}{\arg\min} \sum_{i=1}^{n} |x - x_i|^2 \tag{3.2}$$

with respect to metric

$$d(x, y) = |x - y| \tag{3.3}$$

In fact if we form the quadratic cost function

$$f(x) = \sum_{i=1}^{n} (x - x_i)^2 \tag{3.4}$$

By taking the derivative of Eq.(3.4) with respect to the variable x and set it equal to zero one can obtain the $\bar{x}$ which is the arithmetic mean of positive scalars $\{x_i\}_{i=1}^{n}$.

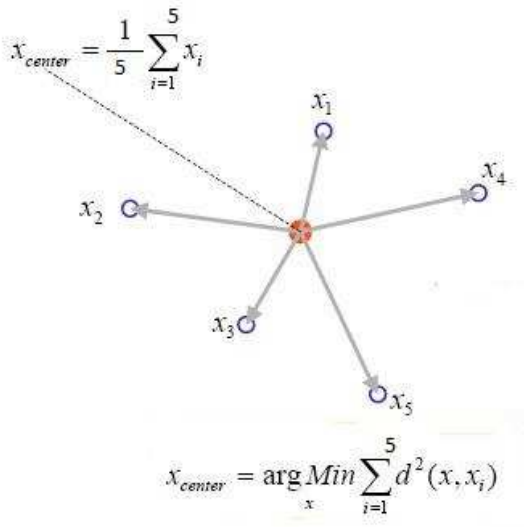


Figure 3.1: Fréchet mean of five points on Euclidian space [1].

3.3 Riemannian metrics

So far we only considered the Euclidean distance which is valid on the space with zero sectional curvature.

To measure the distance between two $M \times M$ covariance matrices $\mathbf{A}$ and $\mathbf{B}$ on manifold of positive definite matrices $\mathcal{M}$, we consider the metrics which have been developed to measure distance between two points on the manifold itself. The following metrics will be considered throughout the remaining chapters.

The first metric is obtained when we lift the points $\mathbf{A}, \mathbf{B}$ to the horizontal subspace $\mathcal{U} \subset \mathcal{H}$ using the fiber and measure the distance between them [5]:

$$d_{R_1}(\mathbf{A}, \mathbf{B}) = \underset{\tilde{\mathbf{U}}_1, \tilde{\mathbf{U}}_2 \in U(M)}{\arg\min} \left\| \mathbf{A}^{\frac{1}{2}} \tilde{\mathbf{U}}_1 - \mathbf{B}^{\frac{1}{2}} \tilde{\mathbf{U}}_2 \right\|_2 = \left\| \mathbf{A}^{\frac{1}{2}} \mathbf{U}_1 - \mathbf{B}^{\frac{1}{2}} \mathbf{U}_2 \right\|_2 \tag{3.5}$$

where $U(M)$ denotes the space of unitary matrices of size $M \times M$. Alternatively Eq.(3.5) can be rewritten as:

$$\sqrt{\mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B}) - 2\,\mathrm{Tr}(\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}} \tag{3.6}$$

In general for any positive definite matrix $\mathbf{A}$ its square root is defined as $\mathbf{A}^{\frac{1}{2}} = \mathbf{S}\sqrt{\mathbf{\Lambda}}\mathbf{D}^H$; where $\mathbf{A} = \mathbf{S}\mathbf{\Lambda}\mathbf{D}^H$ is the eigenvalue value decomposition of matrix $\mathbf{A}$ with diagonal matrix $\mathbf{\Lambda}$ consisting of eigenvalues of $\mathbf{A}$.

In Eq(3.5), $\mathbf{U}_1$ and $\mathbf{U}_2$ are the left and right multiplicative of singular value decomposition of $\mathbf{B}^{\frac{1}{2}} \mathbf{A}^{\frac{1}{2}}$ [21]. To see that the equation (3.5) and (3.6) are indeed equivalent we observe that if the singular value decomposition of matrix $\mathbf{B}^{\frac{1}{2}} \mathbf{A}^{\frac{1}{2}}$ is given by $\mathbf{U}_1 \mathbf{\Gamma} \mathbf{U}_2^H$ where $\mathbf{\Gamma}$ is the diagonal matrix consists of singular values of matrix $\mathbf{B}^{\frac{1}{2}} \mathbf{A}^{\frac{1}{2}}$ in descending

order and $\mathbf{U}_1$ and $\mathbf{U}_2$ are unitary matrices, then we can write:

$$\begin{aligned}
\underset{\tilde{\mathbf{U}}_1,\tilde{\mathbf{U}}_2\in U(M)}{\arg\min} \left\|\mathbf{A}^{\frac{1}{2}}\tilde{\mathbf{U}}_1 - \mathbf{B}^{\frac{1}{2}}\tilde{\mathbf{U}}_2\right\|_2^2 &= \\
&= \underset{\tilde{\mathbf{U}}_1,\tilde{\mathbf{U}}_2\in U(M)}{\arg\min} \operatorname{Tr}\left(\mathbf{A}^{\frac{1}{2}}\tilde{\mathbf{U}}_1 - \mathbf{B}^{\frac{1}{2}}\tilde{\mathbf{U}}_2\right)\left(\mathbf{A}^{\frac{1}{2}}\tilde{\mathbf{U}}_1 - \mathbf{B}^{\frac{1}{2}}\tilde{\mathbf{U}}_2\right)^H \\
&= \operatorname{Tr}(\mathbf{A}) + \operatorname{Tr}(\mathbf{B}) \\
&- 2\Re e \underset{\tilde{\mathbf{U}}_1,\tilde{\mathbf{U}}_2\in U(M)}{\arg\min} \operatorname{Tr}\left(\mathbf{A}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\tilde{\mathbf{U}}_1\tilde{\mathbf{U}}_2^H\right) \qquad (3.7)
\end{aligned}$$

In order to minimize the right side of Eq.(3.7), we need to maximize the last term. As a result the last term in Eq.(3.7) should be maximized with respect to variable $\mathbf{P}$, where $\mathbf{P} = \tilde{\mathbf{U}}_1\tilde{\mathbf{U}}_2^H$. We note that $\mathbf{P}$ is still unitary operator so it belongs to $U(M)$. So we have the following Lagrangian equation:

$$\operatorname{Tr}\left(\mathbf{A}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\mathbf{P} - \frac{1}{2}\mathbf{\Omega}\left(\mathbf{P}\mathbf{P}^H - \mathbf{I}\right)\right) \qquad (3.8)$$

where $\mathbf{\Omega}$ is diagonal Lagrangian multiplier matrix of size $M \times M$ and $\mathbf{I}$ is the identity matrix of size $M \times M$. By taking the derivative of the Lagrangian with respect to $\mathbf{P}$ and equating it to zero we have :

$$\mathbf{B}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}} = \mathbf{\Omega}\mathbf{P} \qquad (3.9)$$

Since $\mathbf{P}$ is a unitary matrix, we have $\mathbf{\Omega} = \mathbf{B}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}\mathbf{P}^H$. As a result we conclude that $\mathbf{\Omega}^2 = \mathbf{B}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}\mathbf{P}^H\mathbf{P}\mathbf{A}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}$. The latter equation can be simplified as:

$$\mathbf{\Omega}^2 = \mathbf{U}_1\mathbf{\Gamma}\mathbf{U}_2^H\mathbf{U}_2\mathbf{\Gamma}\mathbf{U}_1^H \qquad (3.10)$$

As a result we can set $\mathbf{\Omega} = \mathbf{U}_1\mathbf{\Gamma}\mathbf{U}_1^H$. Substituting this value in (3.9) we obtain the optimum value

$$\mathbf{P}_{opt} = \mathbf{U}_1\mathbf{U}_2^H. \tag{3.11}$$

We observe that according to the singular value decomposition of matrix $\mathbf{A}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}$ we have :

$$\mathbf{A}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\mathbf{P}_{opt} = \mathbf{U}_1\Gamma\mathbf{U}_2^H\mathbf{U}_2\mathbf{U}_1^H = \mathbf{U}_1\Gamma\mathbf{U}_1^H. \tag{3.12}$$

On the other hand one can observe that:

$$\mathbf{A}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}} = \mathbf{U}_2\mathbf{\Gamma}^2\mathbf{U}_2^H. \tag{3.13}$$

Thus we have:

$$\mathrm{Tr}\left(\mathbf{A}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}\right)^{\frac{1}{2}} = \mathrm{Tr}\left(\mathbf{A}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\mathbf{P}_{opt}\right). \tag{3.14}$$

According to the equations (3.12), (3.13) and (3.14) one can observe that the equations (3.5) and (3.6) hold.

If we put $\mathbf{U}_1$ and $\mathbf{U}_2$ to be equal to the identity matrix in Eq.(3.5), in other words lifting up the points along the fibers through the identity operator, we will have the expression for metric d_{R_2} as follows [5]:

$$d_{R_2}(\mathbf{A},\mathbf{B}) = \left\|\mathbf{A}^{\frac{1}{2}} - \mathbf{B}^{\frac{1}{2}}\right\|_2 = \sqrt{\mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B}) - 2\,\mathrm{Tr}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}})} \tag{3.15}$$

Let the points $\mathbf{A},\mathbf{B} \in \mathcal{M}$ and let $\mathbf{X}$ be a the point on the manifold at which we construct a tangent plane (it is usually denoted as $T_{\mathcal{M}}\mathbf{X}$). According to the inner-product $\langle \mathbf{A},\mathbf{B}\rangle_{\mathbf{X}} =$

$\mathrm{Tr}(\mathbf{X}^{-1}\mathbf{A}\mathbf{X}^{-1}\mathbf{B})$ the log- Riemannian metric is given as [7]:

$$d_{R_3}(\mathbf{A},\mathbf{B}) = \left\|\log(\mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}})\right\|_2 = \sqrt{\sum_{i=1}^{M}\log^2(\lambda_i)} \tag{3.16}$$

where the λ_i's are the eigenvalues of the matrix $\mathbf{A}^{-1}\mathbf{B}$ [22]. (Metric d_{R3} has been developed in various ways and has, for a long time, been used in theoretical physics).

So far we have introduced the matrices for which we would like to find the corresponding Fréchet mean. In the following sections we will develop the necessary framework to obtain the mean in each case.

3.4 Fréchet mean using metric d_{R1}

Obtaining the Fréchet mean of set of positive definite Hermitian matrices $\{\mathbf{S}_i\}_{i=1}^n$ with respect to metric d_{R1} results on the following optimization problem:

$$\underset{\mathbf{\Sigma}\in\mathcal{M}}{\arg\min}\sum_{i=1}^{n}\left\|\mathbf{S}_i^{\frac{1}{2}}\mathbf{U}_i - \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{U}\right\|_2^2 \tag{3.17}$$

To find the solution for Eq.(3.17) we need to use the next lemma. The result of it will be used to form an algorithm to obtain the solution $\hat{\mathbf{\Sigma}}$ in Eq.(3.17).

Lemma 3.4.1. *: Let $\{\mathbf{D}_i\}_{i=1}^n$ be set of n points in Hilbert space $(\mathcal{H}, \|.\|_2)$. Then we have:*

$$\sum_{i=1}^{n}\sum_{j\geq i}\|\mathbf{D}_i - \mathbf{D}_j\|_2^2 = n\sum_{i=1}^{n}\|\mathbf{D}_i - \mathbf{C}\|_2^2\,. \tag{3.18}$$

where $\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{D}_i$.

Proof.

$$\begin{aligned}\sum_{i=1}^{n} \|\mathbf{D}_i - \mathbf{X}\|_2^2 &= \\ &= \sum_{i=1}^{n} \|\mathbf{D}_i - \mathbf{C} + \mathbf{C} - \mathbf{X}\|_2^2 \\ &= \sum_{i=1}^{n} \|\mathbf{D}_i - \mathbf{C}\|_2^2 + 2(\mathbf{C} - \mathbf{X}).\sum_{i=1}^{n}(\mathbf{D}_i - \mathbf{C}) + n\,\|\mathbf{C} - \mathbf{X}\|_2^2\end{aligned} \tag{3.19}$$

Since $\mathbf{C}$ is center of mass we have $\sum_{i=1}^{n}(\mathbf{D}_i - \mathbf{C}) = 0$. As a result:

$$\sum_{i=1}^{n} \|\mathbf{D}_i - \mathbf{X}\|_2^2 = \sum_{i=1}^{n} \|\mathbf{D}_i - \mathbf{C}\|_2^2 + n\,\|\mathbf{C} - \mathbf{X}\|_2^2 \tag{3.20}$$

Now suppose $\mathbf{X}$ varies over the set $\{\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_n\}$. Using Eq.(3.20) and adding up n equations together we have:

$$\begin{aligned}\sum_{i,j} \|\mathbf{D}_i - \mathbf{D}_j\|_2^2 &= n\sum_{i=1}^{n} \|\mathbf{D}_i - \mathbf{C}\|_2^2 + n\sum_{j=1}^{n} \|\mathbf{C} - \mathbf{D}_j\|_2^2 \\ &= 2n\sum_{i=1}^{n} \|\mathbf{D}_i - \mathbf{C}\|_2^2\end{aligned} \tag{3.21}$$

We note that:

$$\sum_{i,j} \|\mathbf{D}_i - \mathbf{D}_j\|_2^2 = 2\sum_{i}\sum_{j\geq i} \|\mathbf{D}_i - \mathbf{D}_j\|_2^2 \tag{3.22}$$

So from the equation (3.21) and (3.22) and we get the desired result:

$$\sum_{i=1}^{n}\sum_{j\geq i}^{n}\|\mathbf{D}_i-\mathbf{D}_j\|_2^2=n\sum_{i=1}^{n}\|\mathbf{D}_i-\mathbf{C}\|_2^2 \tag{3.23}$$

□

Divide the sides of Eq.(3.23) by n and set $\mathbf{D}_i = \mathbf{A}_i\mathbf{U}_i$, where for the given set of positive definite matrices $\{\mathbf{S}_i\}_{i=1}^n$, $\mathbf{A}_i = \mathbf{S}_i^{\frac{1}{2}}$; define function g as [23]:

$$g\left(\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}_3,...,\mathbf{A}_n\right)=\frac{1}{n}\sum_{i=1}^{n}\sum_{j\geq i}\|\mathbf{A}_i\mathbf{U}_i-\mathbf{A}_j\mathbf{U}_j\|_2^2\,. \tag{3.24}$$

where in Eq.(3.24) $\mathbf{U}_i$'s are unitary operators. Lemma (3.4.1) facilitates the process of identifying the set of unitary matrices $\{\mathbf{U}_i\}_{i=1}^n$ such that minimize the function g.

From Eq.(3.23) we have $\mathbf{C}=\frac{1}{n}\sum_{j=1}^{n}\mathbf{A}_j\mathbf{U}_j$. Also $\mathbf{U}_j$'s are the optimum solution of function g.

The next algorithm simultaneously find the set of unitary matrices$\{\mathbf{U}_i\}_{i=1}^n$ in order to minimize function $g(\mathbf{A}_1,\mathbf{A}_2,...,\mathbf{A}_n)$ in (3.24) and as a consequent finding the Fréchet mean with respect to metric d_{R1}.

Algorithm for computing the Fréchet mean of metric d_{R_1}:

Algorithm 1 Fréchet mean for metric d_{R_1}

1. Initialize the positive threshold value ϵ. For the set $\{\mathbf{S}_i\}_{i=1}^n$ of positive definite matrices on manifold $\mathcal{M}$ find the square root of each element: $\mathbf{A}_i = \mathbf{S}_i^{\frac{1}{2}}$; $i = 1, 2, ..., n$.

2. For each $i = 1, 2, ..., n$ consider $\hat{\mathbf{A}}_i := \frac{1}{n-1}\sum_{j\neq i}^{n} \mathbf{A}_j$ and find $\hat{\mathbf{U}}_i$ which minimizes $\left\|\hat{\mathbf{A}}_i - \mathbf{A}_i\mathbf{U}_i\right\|_2$; then consider $\hat{\mathbf{A}}_{i\text{new}} := \mathbf{A}_i\hat{\mathbf{U}}_i$

3. At iteration $(k+1)$ set $\mathbf{A}_i = \hat{\mathbf{A}}_{i\text{new}}; i = 1, 2, ..., n$ and Evaluate g_{k+1} using Eq.(3.24)

4. Repeat step 2 until:

$$|g_k(\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_n) - g_{k+1}(\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_n)| \leq \epsilon \ .$$

5. Calculate $\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{A}}_i\mathbf{U}_i$.

6. The resulting Fréchet mean on manifold $\mathcal{M}$ is then obtained as $\hat{\mathbf{\Sigma}} = \hat{\mathbf{C}}\hat{\mathbf{C}}^H$

3.5 Fréchet mean corresponding to metric d_{R2}

So far we have seen how to obtain the Fréchet mean using metric d_{R1}. Unlike the Fréchet mean for metric d_{R1}, it will be shown that metric d_{R2} can be utalized in Eq.(3.1) to obtain a Fréchet mean which has a closed form expression.

The optimization problem in Eq.(3.1) with respect to the metric d_{R2} and given positive definite hermitian matrices $\{\mathbf{S}_i\}_{i=1}^n$, is expressed as:

$$\hat{\mathbf{\Sigma}} = \underset{\mathbf{\Sigma}\in\mathcal{M}}{\arg\min} \sum_{i=1}^{n} \left\| \mathbf{S}_i^{\frac{1}{2}} - \mathbf{\Sigma}^{\frac{1}{2}} \right\|_2^2 \tag{3.25}$$

The optimisation problem (3.25) has closed form solution on manifold of positive definite matrices. It can be obtained through the following lemma:

Lemma 3.5.1. *For the set of positive definite Hermitian matrices* $\{\mathbf{S}_i\}_{i=1}^n$ *on manifold* $\mathcal{M}$ *we consider* $\mathbf{L}_i = (\mathbf{S}_i)^{\frac{1}{2}}$ *;* $i = 1, 2, ..., n$ *. Then we have:*

$$\hat{\mathbf{L}} = \underset{\mathbf{L}}{\arg\min} \sum_{i=1}^{n} \left\| \mathbf{L}_i - \mathbf{L}^{\frac{1}{2}} \right\|_2^2 = \frac{1}{n} \sum_{i=1}^{n} \mathbf{L}_i \tag{3.26}$$

Proof.

$$\begin{aligned}
\underset{\mathbf{L}}{\arg\min} \sum_{i=1}^{n} \operatorname{Tr}\left(\mathbf{L}-\mathbf{L}_i\right)^H \left(\mathbf{L}-\mathbf{L}_i\right) &= \\
&= \underset{\mathbf{L}}{\arg\min} \sum_{i=1}^{n} \left(\|\mathbf{L}\|_2^2 + \|\mathbf{L}_i\|_2^2 - 2\operatorname{Tr}\left(\mathbf{L}^H\mathbf{L}_i\right)\right) \\
&\equiv \underset{\mathbf{L}}{\arg\min} \sum_{i=1}^{n} \|\mathbf{L}\|_2^2 - 2\sum_{i=1}^{n} \operatorname{Tr}\left(\mathbf{L}^H\mathbf{L}_i\right) \\
&= \underset{\mathbf{L}}{\arg\min} \sum_{i=1}^{n} \|\mathbf{L}\|_2^2 - 2\operatorname{Tr}\left(\mathbf{L}^H \sum_{i=1}^{n} \mathbf{L}_i\right)
\end{aligned} \tag{3.27}$$

Now in order to minimize the last expression in Eq.(3.27) we need to maximize $\operatorname{Tr}\left(\mathbf{L}^H \sum_{i=1}^n \mathbf{L}_i\right)$. In general for any two Hermitian matrices $\mathbf{A}$ and $\mathbf{B}$ we note that the Cauchy-Schwartz inequality for trace operators is given by [15]:

$$\mathrm{Tr}\left(\mathbf{A}^H\mathbf{B}\right) \leq \sqrt{\mathrm{Tr}\left(\mathbf{A}\mathbf{A}^H\right)}\sqrt{\mathrm{Tr}\left(\mathbf{B}\mathbf{B}^H\right)} \tag{3.28}$$

Equality in Eq.(3.28) holds when one of the matrices $\mathbf{A}$ and $\mathbf{B}$ are multiple of each other. So we have : $\mathbf{L} = \beta \sum_{i=1}^{n} \mathbf{L}_i$ where $\|\mathbf{L}\|_2 = \alpha$. Thus we can write:

$$\|\mathbf{L}\|_2 = |\beta| \left\| \sum_{i=1}^{n} \mathbf{L}_i \right\|_2 = \alpha$$

Since $\beta \geq 0$, we get $\beta = \frac{\alpha}{\left\|\sum_{i=1}^{n} \mathbf{L}_i\right\|_2}$. Substituting β in $\mathbf{L} = \beta \sum_{i=1}^{n} \mathbf{L}_i$ we have :

$$\mathbf{L}^H = \frac{\alpha \left(\sum_{i=1}^{n} \mathbf{L}_i\right)^H}{\left\|\sum_{i=1}^{n} \mathbf{L}_i\right\|_2} \tag{3.29}$$

As a result Eq.(3.29) can be written as:

$$\begin{aligned}
\arg\min_{\alpha} n\alpha^2 - 2\,\mathrm{Tr}\left(\frac{\alpha}{\left\|\sum_{i=1}^{n} \mathbf{L}_i\right\|_2}\left(\sum_{i=1}^{n} \mathbf{L}_i\right)^H\left(\sum_{i=1}^{n} \mathbf{L}_i\right)\right) &= \\
= \arg\min_{\alpha} n\alpha^2 - 2\alpha\,\mathrm{Tr}\left(\left(\sum_{i=1}^{n} \mathbf{L}_i\right)^H \sum_{i=1}^{n} \mathbf{L}_i\right) & \\
= \arg\min_{\alpha} n\alpha^2 - 2\frac{\alpha \left\|\sum_{i=1}^{n} \mathbf{L}_i\right\|_2^2}{\left\|\sum_{i=1}^{n} \mathbf{L}_i\right\|_2} &
\end{aligned} \tag{3.30}$$

Now we need solve the minimization problem in Eq.(3.30). Take the derivative of equation Eq.(3.30) with respect to α and equal it to zero we have:

$$\alpha = \frac{1}{n} \left\| \sum_{i=1}^{n} \mathbf{L}_i \right\|_2 \tag{3.31}$$

Thus we have:

$$\hat{\mathbf{L}} = \frac{\sum_{i=1}^{n} \mathbf{L}_i}{n} \tag{3.32}$$

□

This is the solution on the horizontal subspace $\mathcal{H}_{\mathcal{H}}$ the corresponding point on the manifold itself is given by:

$$\hat{\mathbf{\Sigma}} = \hat{\mathbf{L}}\hat{\mathbf{L}}^H; \text{ where}; \hat{\mathbf{L}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{S}_i^{\frac{1}{2}} \tag{3.33}$$

In similar manner to the above theorem we have seen in Chapter 2 that the solution of the following optimization problem :

$$\hat{\mathbf{\Sigma}} = \underset{\mathbf{\Sigma} \in \mathcal{M}}{\arg\min} \sum_{i=1}^{n} \|\mathbf{\Sigma} - \mathbf{S}_i\|_2^2 \tag{3.34}$$

is $\hat{\mathbf{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{S}_i$ which by definition is arithmetic mean of positive definite covariance matrices.

It is worth to note that in the case that the measurement are in form of non negative

scalars $\{x_i\}_{i=1}^n$ the analogy to the solution of optimization problem in (3.25) is :

$$\hat{\Sigma} = \left(\frac{1}{n} \sum_{i=1}^{n} \sqrt{x_i} \right)^2 \tag{3.35}$$

which is known as square root mean [24].

3.6 Fréchet mean using metric d_{R3}

The Fréchet mean with respect to metric d_{R3} for the set of positive definite matrices $\{\mathbf{S}_i\}_{i=1}^n$ in manifold $\mathcal{M}$ using Eq.(3.1) can be formulated as:

$$\hat{\mathbf{\Sigma}} = \underset{\mathbf{\Sigma} \in \mathcal{M}}{\arg\min} \sum_{i=1}^{n} d_{R3}^2(\mathbf{S}_i, \mathbf{\Sigma}) \tag{3.36}$$

3.6.1 Convexity of the problem

Before going through the algorithm for finding the optimum solution of Eq.(3.36) we show that this optimization problem has a unique solution[1]. The next theorem characterizes the geodesic in the case of metric d_{R3}. The result of it later on will be used in proof of convexity of Eq.(3.36).

Theorem 3.6.1. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *belong to the manifold of positive Hermitian matrices* $\mathcal{M}$.

[1] parts of the results of this section is the extension to the case of n covariance matrices from [10] .

Then there exists a unique geodesic [2] $\gamma(t)$ *joining* $\mathbf{A}$ *and* $\mathbf{B}$ *such that it can be parameterized as follows :*

$$\gamma(t) = \mathbf{A}^{\frac{1}{2}} \exp\left(t \log\left(\mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}}\right)\right)\mathbf{A}^{\frac{1}{2}}; 0 \leq t \leq 1 \tag{3.37}$$

Proof. We want to find the path with minimum length such that it connects the matrices $\mathbf{A}$ and $\mathbf{B}$. From [18], it has been shown that the congruence transformation $\Gamma_{\mathbf{X}}(\mathbf{A}) = \mathbf{X}^H\mathbf{A}\mathbf{X}$ and $\Gamma_{\mathbf{X}}(\mathbf{B}) = \mathbf{X}^H\mathbf{B}\mathbf{X}$ preserve the length between the two points A and B, i.e., $d_{R3}(\mathbf{A},\mathbf{B}) = d_{R3}(\Gamma_{\mathbf{X}}(\mathbf{A}), \Gamma_{\mathbf{X}}(\mathbf{B}))$ on manifold $\mathcal{M}$. On the other hand for any pair of commutative matrices $\mathbf{C}$ and $\mathbf{D}$, the shortest path (geodesic) is formulated as (see Appendix C):

$$\gamma_0(t) = \exp\left((1-t)\log\left(\mathbf{C}\right) + t\log\left(\mathbf{D}\right)\right) \tag{3.38}$$

Now we apply congruence operator $\Gamma_{\mathbf{A}^{-\frac{1}{2}}}$ on the path passing through the points $\mathbf{A}$ and $\mathbf{B}$. The result is the same length of the path passing through the points $\mathbf{I}$ and $\mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}}$; since they commute [18] the geodesic between them according to the Eq.(3.38) is given by :

$$\gamma_0(t) = \exp\left(t\log\left(\mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}}\right)\right) \tag{3.39}$$

Now we apply congruence $\Gamma_{\mathbf{A}^{\frac{1}{2}}}$ on Eq.(3.39) to get the geodesic passing through the points $\mathbf{A}$ and $\mathbf{B}$ as follows:

$$\gamma(t) = \mathbf{A}^{\frac{1}{2}} \exp\left(t \log\left(\mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}}\right)\right)\mathbf{A}^{\frac{1}{2}} \quad 0 \leq t \leq 1 \tag{3.40}$$

[2] The geodesic in our work can be assumed as the path which has the minimum length among the other paths which connect two points on the manifold $\mathcal{M}$. For more details on the notion of geodesic see [25] .

which is the geodesic connects $\mathbf{A}$ at $t = 0$ to $\mathbf{B}$ at $t = 1$. □

In Eq.(3.39) if $t = \frac{1}{2}$ the point $\gamma(\frac{1}{2})$ is the middle point on the geodesic connecting two points $\mathbf{A}$ and $\mathbf{B}$ in $\mathcal{M}$; it is called the geometric mean of $\mathbf{A}$ and $\mathbf{B}$ and represented by $\mathbf{A}\#\mathbf{B}$ [11] .

In Hilbert space $\mathcal{H}$ we know that for any two points $\mathbf{A}$ and $\mathbf{B}$ with the middle point $\mathbf{M} = \frac{\mathbf{A}+\mathbf{B}}{2}$ together with any point $\mathbf{C}$ we have *parallelogram* rule :

$$\mathbf{AB}^2 + \mathbf{CD}^2 = 2\left(\mathbf{AC}^2 + \mathbf{BC}^2\right) \tag{3.41}$$

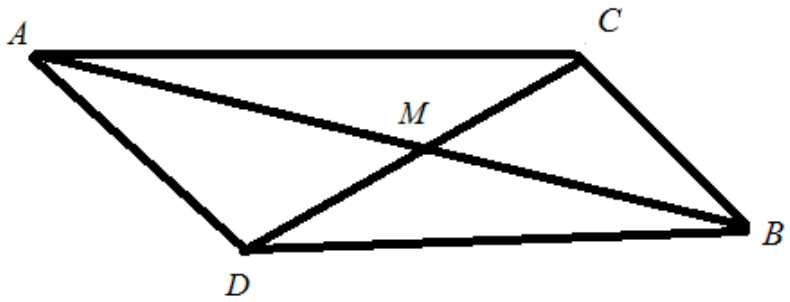


Figure 3.2: Parallelogram rule in the Hilbert space. $\mathbf{M}$ is the middle point in the sense that $\|\mathbf{A} - \mathbf{M}\| = \|\mathbf{M} - \mathbf{B}\|$. The forth vertex $\mathbf{D}$ is determined using Eq.(3.41).

Similar to the Eq.(3.41) we have *semi parallelogram* rule with respect to metric dR_3 [10]:

Theorem 3.6.2. *Suppose* $\mathbf{A}$ *and* $\mathbf{B}$ *are two points on the manifold of positive definite matrices* $\mathcal{M}$. *Let* $\mathbf{M} = \mathbf{A}\#\mathbf{B}$ *be the middle point of the geodesic passing through* $\mathbf{A}$ *and* $\mathbf{B}$. *Let* $\mathbf{C}$ *be any arbitrary point in* $\mathcal{M}$. *Then we have :*

$$d_{R_3}{}^2(\mathbf{M}, \mathbf{C}) \leq \frac{d_{R3}{}^2(\mathbf{A}, \mathbf{C}) + d_{R3}{}^2(\mathbf{B}, \mathbf{C})}{2} - \frac{d_{R3}{}^2(\mathbf{A}, \mathbf{B})}{4} \tag{3.42}$$

Lemma 3.6.3. *For the fixed positive definite Hermitian matrix* $\mathbf{A}$ *function* $f(\mathbf{X}) = {d_{R3}}^2(\mathbf{A}, \mathbf{X})$ *is convex.*

Proof. Since every continuous midpoint convex function [3] is convex [26] , it is sufficient to show that f is continuous and midpoint convex for any positive definite Hermitian matrices $\mathbf{X}_1$ and $\mathbf{X}_2$. Function $f(t) = \log(t)$ is continuous [27] on $(0, \infty)$. Let $\mathbf{X}_n$ be a sequence in $\mathcal{M}$ such that $\lim_{n\longrightarrow\infty} \mathbf{X}_n = \mathbf{X}$. For the fixed Hermitian positive matrix $\mathbf{A}$ we have $\lim_{n\longrightarrow\infty} \mathbf{A}^{\frac{1}{2}}\mathbf{X}_n\mathbf{A}^{\frac{1}{2}} = \mathbf{A}^{\frac{1}{2}}\mathbf{X}\mathbf{A}^{\frac{1}{2}}$. As a result: [4]

$$\lim_{n\longrightarrow\infty} \left\|\log\left(\mathbf{A}^{\frac{1}{2}}\mathbf{X}_n\mathbf{A}^{\frac{1}{2}}\right)\right\|_2^2 = \left\|\log\left(\mathbf{A}^{\frac{1}{2}}\mathbf{X}\mathbf{A}^{\frac{1}{2}}\right)\right\|_2^2 \tag{3.43}$$

So f is a countinious function. Using theorem (3.6.2) we can see that f is mid-point convex:

$$\begin{aligned} f(\mathbf{X}_1\#\mathbf{X}_2) = {d_{R3}}^2(\mathbf{X}_1\#\mathbf{X}_2, \mathbf{A}) &\leq \frac{{d_{R3}}^2(\mathbf{X}_1, \mathbf{A}) + {d_{R3}}^2(\mathbf{X}_2, \mathbf{A})}{2} - \frac{{d_{R3}}^2(\mathbf{X}_1, \mathbf{X}_2)}{4} \\ &\leq \frac{1}{2}\left({d_{R3}}^2(\mathbf{X}_1, \mathbf{A}) + {d_{R3}}^2(\mathbf{X}_2, \mathbf{A})\right) \end{aligned} \tag{3.44}$$

Equation (3.44) shows that f is mid point convex. □

Lemma 3.6.4. *The function* $f(\mathbf{X}) = \sum_{j=1}^{n} {d_{R3}}^2(\mathbf{A}_j, \mathbf{X})$ *has a unique minimizer solution in* $\mathcal{M}$.

[3] A function $f : \mathcal{M} \longrightarrow \mathbb{R}$ is called midpoint convex if for any arbitrary points $\mathbf{A}, \mathbf{B} \in \mathcal{M}$ one can conclude that: $f(\mathbf{A}\#\mathbf{B}) \leq \frac{1}{2}(f(\mathbf{A}) + f(\mathbf{B}))$.

[4] A function f is called continious if for any sequence $\{X_n\}$ in the given space, in here manifold $\mathcal{M}$,such that $\{\mathbf{X}_n\}$ converges to the $\mathbf{X}$ as n tends to the infinity, one can conclude that $f(\mathbf{X}_n)$ converges to $f(\mathbf{X})$. It is equivalent to the conventional definition of continuity of the function f and in some cases it facilitate the proof of continuity of a given function. For more information see [28].

Proof. Consider $m = \inf f(\mathbf{X})$[5] and let $\{\mathbf{X}_r\}$ be a sequence in $\mathcal{M}$ such that $\lim_{r\longrightarrow\infty} f(\mathbf{X}_r) = m$. Using semi parallelogram rule for $j = 1, 2, 3, ..., n$ we have :

$$d_{R3}{}^2(\mathbf{X}_r \# \mathbf{X}_s, \mathbf{A}_j) \leq \frac{d_{R3}{}^2(\mathbf{X}_r, \mathbf{A}_j) + d_{R3}{}^2(\mathbf{X}_s, \mathbf{A}_j)}{2} - \frac{d_{R3}{}^2(\mathbf{X}_r, \mathbf{X}_s)}{4} \tag{3.45}$$

Adding up all the equations in Eq.(3.45) we get:

$$\begin{aligned} f(\mathbf{X}_r \# \mathbf{X}_s) &= \sum_{j=1}^{n} d_{R3}{}^2(\mathbf{X}_r \# \mathbf{X}_s, \mathbf{A}_j) \\ &\leq \frac{1}{2}\left(\sum_{j=1}^{n} d_{R3}{}^2(\mathbf{X}_r, \mathbf{A}_j) + \sum_{j=1}^{n} d_{R3}{}^2(\mathbf{X}_s, \mathbf{A}_j)\right) - \sum_{j=1}^{n} \frac{d_{R3}{}^2(\mathbf{X}_r, \mathbf{X}_s)}{4} \\ &= \frac{1}{2}(f(\mathbf{X}_r) + f(\mathbf{X}_s)) - \frac{n}{4} d_{R3}{}^2(\mathbf{X}_r, \mathbf{X}_s) \end{aligned} \tag{3.46}$$

From Eq.(3.46) we have:

$$\begin{aligned} \frac{n}{4} d_{R3}{}^2(\mathbf{X}_r, \mathbf{X}_s) &\leq \frac{1}{2}(f(\mathbf{X}_r) + f(\mathbf{X}_s)) - f(\mathbf{X}_r \# \mathbf{X}_s) \\ &\leq \frac{1}{2}(f(\mathbf{X}_r) + f(\mathbf{X}_s)) - m \end{aligned} \tag{3.47}$$

Multiplying both sides of Eq.(3.47) by 2 and notice that the right hand side of it is positive

[5]Infimum of a function f which is usually denoted as $\inf f$ is the greatest lower bound of the function. For more information about infimum of a function see [28].

so we can consider its absolute value:

$$\begin{aligned}\frac{n}{2}{d_{R3}}^2\left(\mathbf{X}_r,\mathbf{X}_s\right) \leq \left|f\left(\mathbf{X}_r\right)-m+f\left(\mathbf{X}_s\right)-m\right| & \\ \leq \left|f\left(\mathbf{X}_r\right)-m\right|+\left|f\left(\mathbf{X}_s\right)-m\right| & \\ \leq 2\epsilon & \end{aligned} \tag{3.48}$$

Thus $\mathbf{X}_r$ is a Cauchy sequence in $\mathcal{M}$ and since $(\mathcal{M}, d_{R3})$ is a complete space it converges to its limit point $\mathbf{X}_0$; in other words $\lim_{r\longrightarrow\infty}\mathbf{X}_r = \mathbf{X}_0$. On the other hand since f is a continuous function we can write $\lim_{r\longrightarrow\infty} f(\mathbf{X}_r) = f(\mathbf{X}_0)$; since the limit is unique as a result we must have $f\left(\mathbf{X}_0\right) = m = \inf f\left(\mathbf{X}\right)$. Furthermore, f is the summation of convex functions as a result it is convex so it reaches its minimum at the unique point in $\mathcal{M}$. □

3.6.2 Steepest descent algorithm on Riemannian manifold

We have shown that Eq.(3.36) is a convex optimization problem and its solution, if it exists, must be unique. However, to our knowledge a closed form solution for Eq.(3.36) has not been found yet. To find the Fréchet mean we need to use numerical methods [29]. Among the numerical approaches that can be applied we will use gradient descent method [2] to find the optimum solution.

Recall that manifold of positive definite matrices $\mathcal{M}$ equipped with the inner product:

$$\langle \mathbf{A},\mathbf{B}\rangle_{\mathbf{X}} := \mathrm{Tr}\left(\mathbf{A}\mathbf{X}^{-1}\mathbf{B}\mathbf{X}^{-1}\right) \tag{3.49}$$

Which leads to the Riemannian distance d_{R3}.

The optimum solution of the smooth cost function can be obtained when its differential is equal to zero. On the manifold equipped with the inner product (3.49) we have the well-known relation [22] between vector $\mathbf{V}$ in $T_{\mathbf{X}}\mathcal{M}$ and the gradient of the real valued function $f : \mathcal{M} \longrightarrow \mathbb{R}^{+}$,$\nabla f$, at point $\mathbf{X} \in \mathcal{M}$:

$$D_{\mathbf{V}} f(\mathbf{X}) := \langle \nabla f, \mathbf{V} \rangle_{\mathbf{X}} \tag{3.50}$$

$D_{\mathbf{V}} f(\mathbf{x})$ is called directional derivative of function f at point $\mathbf{X}$ in direction of $\mathbf{V}$. For more details on the directional derivative, see Appendix A. In order to connect the definition of directional derivative on the manifold with its corresponding on the Euclidean space we observe that:

$$\left\langle \nabla f^{\Re}, \mathbf{V} \right\rangle_{\mathbf{X}} = \mathrm{Tr}\left(\nabla f^{\Re} \mathbf{X}^{-1} \mathbf{V} \mathbf{X}^{-1}\right) = \left\langle \mathbf{X}^{-1} \nabla f^{\Re} \mathbf{X}^{-1}, \mathbf{V} \right\rangle \tag{3.51}$$

where the notation $\nabla f^{\Re}$ emphasizes that the gradient vector is obtained on the Riemannian manifold $\mathcal{M}$. The first expression is given by definition of Riemannian metric inner product as mentioned in equation (3.49). The last expression by definition is equal to the inner product on the Euclidian space; in other words we have the following relation between the gradient vector on the Reimannian manifold and its counterpart on the Euclidean space:

$$\mathbf{X}^{-1} \nabla f^{\Re} \mathbf{X}^{-1} = \nabla f^{\mathcal{E}} \tag{3.52}$$

where $\nabla f^{\mathcal{E}}$ represents the corresponding gradient vector with respect to the inner product on the Euclidean space.

In view of Lemma (2.2.1), let $h(\mathbf{X}) = \|\log(\mathbf{X})\|_2^2$, $\mathbf{X} \in \mathcal{M}$ then we have :

$$D_{\mathbf{Y}} h(\mathbf{X}) = 2 \left\langle \mathbf{X}^{-1} \log \mathbf{X}, \mathbf{Y} \right\rangle \tag{3.53}$$

For all $\mathbf{Y} \in T_{\mathcal{M}}$.

The immediate consequence of this result shows that if we consider

$$h(\mathbf{X}) = \left\| \log(\mathbf{A}^{-\frac{1}{2}} \mathbf{X} \mathbf{A}^{-\frac{1}{2}}) \right\|_2^2 \tag{3.54}$$

then for every $\mathbf{Y} \in T_{\mathcal{M}}$ we have:

$$D_{\mathbf{Y}} h(\mathbf{X}) = 2 \left\langle \mathbf{A}^{-\frac{1}{2}} \mathbf{X} \mathbf{A}^{-\frac{1}{2}} \log(\mathbf{A}^{-\frac{1}{2}} \mathbf{X} \mathbf{A}^{-\frac{1}{2}}), \mathbf{A}^{-\frac{1}{2}} \mathbf{Y} \mathbf{A}^{-\frac{1}{2}} \right\rangle \tag{3.55}$$

Theorem 3.6.5. *The directional derivative of the function* $f(\mathbf{X}) = \sum_{i=1}^{n} \left\| \log(\mathbf{A}_i^{-\frac{1}{2}} \mathbf{X} \mathbf{A}_i^{-\frac{1}{2}}) \right\|_2^2$*, where* $\{\mathbf{A}_i\}_{i=1}^n \in \mathcal{M}$*, is given by:*

$$D_{\mathbf{Y}} f(\mathbf{X}) = 2 \sum_{i=1}^{n} \left\langle \mathbf{X}^{-1} \log(\mathbf{X} \mathbf{A}_i^{-1}), \mathbf{Y} \right\rangle \tag{3.56}$$

Proof. The function f is given to have n terms and according to Eq.(3.55) each term has

the following form:

$$
\begin{aligned}
2\left\langle \mathbf{A}_i^{-\frac{1}{2}}\mathbf{X}\mathbf{A}_i^{-\frac{1}{2}}\log(\mathbf{A}_i^{-\frac{1}{2}}\mathbf{X}\mathbf{A}_i^{-\frac{1}{2}}), \mathbf{A}_i^{-\frac{1}{2}}\mathbf{Y}\mathbf{A}_i^{-\frac{1}{2}}\right\rangle & \\
&= 2\,\mathrm{Tr}\left(\mathbf{A}_i^{\frac{1}{2}}\mathbf{X}^{-1}\mathbf{A}_i^{\frac{1}{2}}\log\left(\mathbf{A}_i^{-\frac{1}{2}}\mathbf{X}\mathbf{A}_i^{-\frac{1}{2}}\right)\mathbf{A}_i^{-\frac{1}{2}}\mathbf{Y}\mathbf{A}_i^{-\frac{1}{2}}\right) \\
&= 2\,\mathrm{Tr}\left(\mathbf{X}^{-1}\mathbf{A}_i^{\frac{1}{2}}\log\left(\mathbf{A}_i^{-\frac{1}{2}}\mathbf{X}\mathbf{A}_i^{-\frac{1}{2}}\right)\mathbf{A}_i^{-\frac{1}{2}}\mathbf{Y}\right) \\
&= 2\,\mathrm{Tr}\left(\mathbf{X}^{-1}\log\left(\mathbf{X}\mathbf{A}_i^{-1}\right)\mathbf{Y}\right) \\
&= \left\langle 2\mathbf{X}^{-1}\log\left(\mathbf{X}\mathbf{A}_i^{-1}\right), \mathbf{Y}\right\rangle \qquad (3.57)
\end{aligned}
$$

Using the linearity of gradient and the definition of gradient on the Euclidean space (see appendix A) we have:

$$\nabla f^{\mathcal{E}}(\mathbf{X}) = 2\sum_{i=1}^{n}\mathbf{X}^{-1}\log\left(\mathbf{X}\mathbf{A}_i^{-1}\right) \qquad (3.58)$$

□

We observe that the corresponding gradient with respect to the Riemannian inner product of $\langle \mathbf{A}, \mathbf{B}\rangle_{\mathbf{X}} := \mathrm{Tr}\left(\mathbf{A}\mathbf{X}^{-1}\mathbf{B}\mathbf{X}^{-1}\right)$ is obtained as:

$$\nabla f^{\Re}(\mathbf{X}) = \mathbf{X}\nabla f^{\mathcal{E}}\mathbf{X} = 2\sum_{i=1}^{n}\log\left(\mathbf{X}\mathbf{A}_i^{-1}\right)\mathbf{X} \qquad (3.59)$$

The last term in above equation can be rewritten [29] as:[6]

$$\begin{aligned}\nabla^{\Re} f(\mathbf{X}) = \mathbf{X}\nabla^{\mathcal{E}} f\mathbf{X} &= \\ &= 2\sum_{i=1}^{n} \log\left(\mathbf{X}\mathbf{A}_i^{-1}\right)\mathbf{X} \\ &= 2\sum_{i=1}^{n} \mathbf{X}^{\frac{1}{2}}\mathbf{X}^{-\frac{1}{2}} \log\left(\mathbf{X}\mathbf{A}_i^{-1}\right)\mathbf{X}^{\frac{1}{2}}\mathbf{X}^{\frac{1}{2}} \\ &= 2\sum_{i=1}^{n} \mathbf{X}^{\frac{1}{2}} \log\left(\mathbf{X}^{\frac{1}{2}}\mathbf{A}_i^{-1}\mathbf{X}^{\frac{1}{2}}\right)\mathbf{X}^{\frac{1}{2}} \end{aligned} \tag{3.60}$$

So far we have shown that the Eq.(3.36) is a convex problem. Furthermore, we have obtained the gradient vector according to Eq.(3.60). At this point we can use gradient descent algorithm to find the minimizer of $f(\mathbf{X})$. This algorithm is illustrated in figure (3.3).

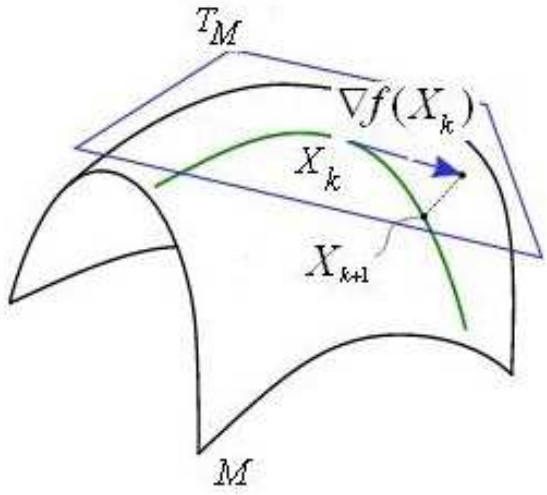


Figure 3.3: Gradient descent algorithm at step k [2].

[6]For any invertable matrix $\mathbf{X}$ and positive matrix $\mathbf{A}$ we have: $\mathbf{X}^{-1}\log(\mathbf{A})\mathbf{X} = \log\left(\mathbf{X}^{-1}\mathbf{A}\mathbf{X}\right)$;see [27], [30].

The geodesic $\gamma(t)$,where $t \in [0,1]$, such that $\gamma(0) = \mathbf{P} \in \mathcal{M}$ and $\frac{d\gamma(t)}{dt}|_{t=0} = \mathbf{S} \in T_{\mathcal{M}}$ is represented by [7] :

$$\gamma(t) = \mathbf{P}^{\frac{1}{2}} \exp\left(t\mathbf{P}^{-\frac{1}{2}}\mathbf{S}\mathbf{P}^{-\frac{1}{2}}\right)\mathbf{P}^{\frac{1}{2}} \tag{3.61}$$

At point $\mathbf{X}_k \in \mathcal{M}$ together with the gradient vector $\nabla f^{\Re}(\mathbf{X}_k) \in T_{\mathcal{M}}\mathbf{X}_k$, the point $\mathbf{X}_{k+1}$ at iteration $k+1$ is obtained by considering the descending direction of gradient vector at point $\mathbf{X}_k \in \mathcal{M}$. The resulting point according to the Eq.(3.61)and Eq.(3.60) is given by :

$$\mathbf{X}_{k+1} = \mathbf{X}_k^{\frac{1}{2}} \exp\left(-t\mathbf{X}_k^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{X}_k^{\frac{1}{2}} \log\left(\mathbf{X}_k^{\frac{1}{2}}\mathbf{A}_i^{-1}\mathbf{X}_k^{\frac{1}{2}}\right)\mathbf{X}_k^{\frac{1}{2}}\mathbf{X}_k^{-\frac{1}{2}}\right)\mathbf{X}_k^{\frac{1}{2}} \tag{3.62}$$

The following algorithm summarizes the process of gradient descent method with respect to metric d_{R3} on manifold $\mathcal{M}$:

Descent algorithm for finding the Fréchet mean using metric dR_3:

Algorithm 2 Fréchet mean for metric d_{R_3}

1. For the given set of symmetric positive definite matrices $\{\mathbf{S}_i\}_{i=1}^{n}$ Initialize the starting point $\mathbf{X}_0$ as the arithmetic mean of the positive definite matrices.

2. Compute the gradient of function f according to the equation Eq.(3.60).

3. Consider the negative direction of gradient vector. It shows the descent search direction.

4. At step $k+1$ calculate the geodesic on manifold $\mathcal{M}$ starting at point $\mathbf{X}_k$ in direction of $-\nabla f^{\Re}(\mathbf{X}_k)$ with step size [7] $t \in (0,1)$ using Eq.(3.62):

$$\mathbf{X}_{k+1} = (\mathbf{X}_k)^{\frac{1}{2}} \exp\left(-t \sum_{i=1}^{n} \log\left(\mathbf{X}_k^{\frac{1}{2}} \mathbf{S}_i^{-1} \mathbf{X}_k^{\frac{1}{2}}\right)\right) (\mathbf{X}_k)^{\frac{1}{2}} \tag{3.63}$$

5. For the given threshold ϵ; repeat step 4 until $\|\nabla f(\mathbf{X}_k)\|_2 \leq \epsilon$.

[7] Step size t can be determined by using the Armijo rule [2].

3.6.3 Alternative representation for metric d_{R3}

We note that it has been shown that in general $\left\|\log\left(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}}\right)\right\|_2$ is not equal to

$$\|\log(\mathbf{A}) - \log(\mathbf{B})\|_2 \tag{3.64}$$

Unless for any two positive definite matrices $\mathbf{A}$ and $\mathbf{B}$ of size $M \times M$ in $\mathcal{M}$ they commute with each other [10]; i.e $\mathbf{AB} = \mathbf{BA}$. However, the following relation between metric in Eq.(3.64) and metric d_{R3} exists:

$$\|\log(\mathbf{A}) - \log(\mathbf{B}))\|_2 \leq d_{R3}(\mathbf{A}, \mathbf{B}) \tag{3.65}$$

The inequality(3.65) is called [31] "Exponential Metric Increasing property" or (EMI).

Since (3.64) satisfies on the all axioms of metric it can be used for obtaining the Fréchet mean.

If we define

$$d(\mathbf{A}, \mathbf{B}) = \|\log(\mathbf{A}) - \log(\mathbf{B})\|_2 \tag{3.66}$$

It can be seen that :

(1) For any symmetric positive definite matrices $\mathbf{A}$ and $\mathbf{B}$ in $\mathcal{M}$, $d(\mathbf{A}, \mathbf{B})$ is non negative.

(2) $d(\mathbf{A}, \mathbf{B}) = 0$ if and only if we have

$$\log(\mathbf{A}) = \log(\mathbf{B}) \tag{3.67}$$

It can be seen [30] that by taking the exponential from both side of Eq.(3.67) and since $\mathbf{A}$ and $\mathbf{B}$ are positive we conclude that $\mathbf{A} = \mathbf{B}$

(3) For any triple $(\mathbf{A}, \mathbf{B}, \mathbf{C})$, we have :

$$\begin{aligned} d(\mathbf{A}, \mathbf{B}) = \|\log(\mathbf{A}) - \log(\mathbf{B})\|_2 &= \|\log(\mathbf{A}) - \log(\mathbf{B}) - \log(\mathbf{C}) + \log(\mathbf{C})\|_2 \\ &\leq \|\log(\mathbf{A}) - \log(\mathbf{C})\|_2 + \|\log(\mathbf{C}) - \log(\mathbf{B})\|_2 \\ &= d(\mathbf{A}, \mathbf{C}) + d(\mathbf{B}, \mathbf{C}) \end{aligned} \tag{3.68}$$

Equation (3.68) confirms that Eq.(3.66) satisfies in triangle inequality. As a result Eq.(3.66) meets all axioms of the metric.[8]

As a consequence of Eq.(3.66) if we have the set of positive scaler observation $\{x_i\}_{i=1}^n$ then its Fréchet mean with respect to metric d_{R3} is given as :

$$\hat{\Sigma} = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log(x_i)\right) = \Pi_{i=1}^{n} x_i{}^{\frac{1}{n}} \tag{3.69}$$

Equation(3.69) is known as the geometric mean of for the set $\{x_i\}_{i=1}^n$.

In this chapter we have developed the frame work to obtain the Fréchet mean of Hermitian positive definite matrices depending on the metrics which have been developed on the Riemannian manifold $\mathcal{M}$. As far as the analysis of the performance of each estimator is concerned, next chapter is devoted to evaluate the methods in estimation of mean of positive semi definite matrices together with the application in classification of real data sets.

[8] If we form the definition of Fréchet mean for this metric it can be seen that for the given positive definite matrices $\{\mathbf{S}\}_{i=1}^n$ it has the closed form solution of form $\hat{\mathbf{\Sigma}} = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log(\mathbf{S}_i)\right)$.

Chapter 4

Simulation results and implementation

In Chapter 3 we focused on the manifold of positive definite matrices $\mathcal{M}$. It has been shown that we could not only work with the Euclidean metric but also could consider the Riemannian distances. Consequently we considered the notion of geometric means. Depending on the different metrics, we developed a metric based method to approach to the problem of estimating the mean of collection of positive definite matrices.

In this chapter the performance of each metric in estimation of Fréchet mean of group of covariance matrices will be studied. Basically, through sets of simulations it will be shown that how close is the estimated symmetric positive definite matrix to the nominal value.

To utilize the concept of Fréchet mean on a real data application,the high content cell image date set (HCI) have been chosen for classification task. For this purpose the method of distance to the center of mass will be performed on the data set.

4.1 Evaluation of Fréchet mean of symmetric positive definite matrices

We have introduced different estimators corresponding to the different distance measures to find the Fréchet mean of set of symmetric positive definite matrices $\{\mathbf{S}_i\}_{i=1}^n$ on manifold $\mathcal{M}$. In order to compare the performance of each estimator we consider a population of $M \times M$ covariance matrices and find the mean of them using each estimator. We will consider different models having the same "true" means so that a comparison of the closeness of the different estimates to this "true" mean is possible.

4.1.1 Model description

To come up with the first model we consider the known symmetric positive definite matrix $\mathbf{\Sigma}$ as the nominal value. Then we apply the Cholesky decomposition to it. By definition the Cholesky factor of a symmetric positive definite matrix $\mathbf{\Sigma}$ is a lower triangular matrix $\mathbf{\Psi}$ with positive diagonal elements such that $\mathbf{\Sigma} = \mathbf{\Psi}\mathbf{\Psi}^H$.

We denote the Cholesky factor of $\mathbf{\Sigma}$ in the model with $\mathbf{\Psi}$ and set $\mathbf{\Psi}$=Chol($\mathbf{\Sigma}$); where "Chol" represents the Cholesky factor of $\mathbf{\Sigma}$. We also consider set of matrices $\{\mathbf{X}_i\}_{i=1}^n$ with the entries $\left\{x_{jk}^i\right\}_{j,k}$ drawn from a normal distribution with zero mean and prescribed variance σ^2. Now to form the new population of covariance matrices $\{\mathbf{S}_i\}_{i=1}^n$ with respect to the nominal covariance matrix $\mathbf{\Sigma}$ we consider the following model:

$$\mathbf{S}_i = (\mathbf{\Psi} + \mathbf{X}_i)(\mathbf{\Psi} + \mathbf{X}_i)^H \tag{4.1}$$

The next model that we consider for the simulation purpose is similar to the first model; However, the noise is added to the upper triangular entries of the Cholesky factor $\mathbf{\Psi}$.

$$\mathbf{S}_i = (\mathbf{\Psi} + \mathbf{X}_i)(\mathbf{\Psi} + \mathbf{X}_i)^H; \quad i = 1, 2, ..., n \tag{4.2}$$

where the upper triangular entries of $\mathbf{X}_i$ are non zero.

In the third model we perturb the covariance matrix $\mathbf{\Sigma}$ by taking its square root : $\mathbf{\Sigma}^{\frac{1}{2}} = \mathbf{\Delta}$. The new population $\mathbf{S}_i$ is defined as :

$$\mathbf{S}_i = (\mathbf{\Delta} + \mathbf{X}_i)(\mathbf{\Delta} + \mathbf{X}_i)^H; \quad i = 1, 2, ..., n \tag{4.3}$$

where $\mathbf{X}_i$ is a random matrix such that the entries are withdrawn from Gaussian distribution with zero mean and variance σ^2.

The model (4.3) can be viewed as the Gaussian noise is added to the $\mathbf{\Delta}$ on the horizontal subspace $\mathcal{H}_H$ and the resulting $\mathbf{S}_i$'s are their corresponding covariance matrices on the manifold $\mathcal{M}$.

The last model that will be considered in here uses the natural connection between manifold $\mathcal{M}$ and its tangent space $T_{\mathcal{M}}$ by using logarithmic and exponential maps.

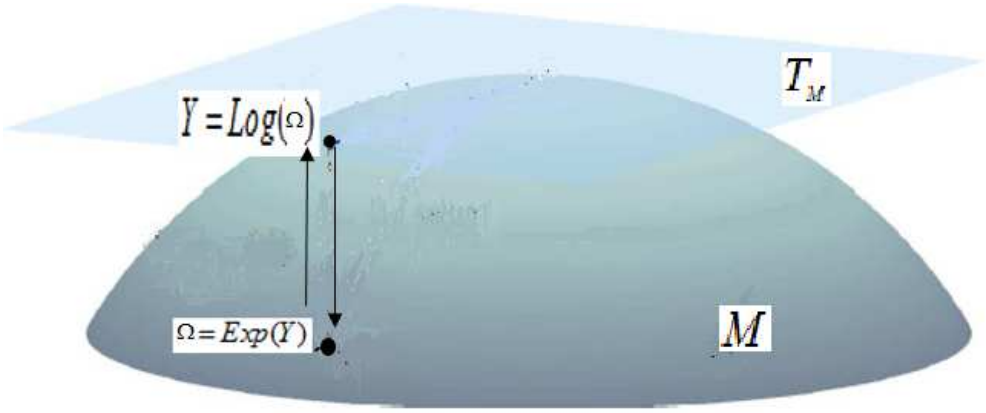


Figure 4.1: connectivity between manifold $\mathcal{M}$ and its tangent space $T_{\mathcal{M}}$ using Log-Exponential map [3].

In this model for the given covariance matrix $\mathbf{\Sigma}$, the logarithm of its Cholesky factor is considered. The random Gaussian noise matrix $\{\mathbf{X}_i\}_{i=1}^n$ is added to the logarithm of the Cholesky factor of true covarience matrix $\mathbf{\Sigma}$ to form the matrix $\{\mathbf{Y}_i\}_{i=1}^n$. The entries of the additive Gaussian random matrix have zero mean and certain amount of variance σ^2; The amount of variance in additive noise is chosen to be a fraction of the trace of nominal covariance matrix. The population $\{\mathbf{S}_i\}_{i=1}^n$ where n represents the population size of symmetric positive definite matrices is defined as:

$$\mathbf{S}_i = \exp\left(\mathbf{Y}_i \mathbf{Y}_i^H\right) \quad i = 1, 2, .., n. \tag{4.4}$$

In order to evaluate the performance of each estimator, we measure the closeness of the estimated covariance matrix, $\hat{\mathbf{\Sigma}}$, to the nominal covariance matrix, $\mathbf{\Sigma}$, using different metrics. In this work three metrics will be considered to evaluate the distance between $\mathbf{\Sigma}$ and the resulting estimation $\hat{\mathbf{\Sigma}}$ according to each model.

First, the discrepancy is measured by using metric d_F:

$$d_F\left(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}\right) = \left\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\right\|_2 \tag{4.5}$$

where for matrix $\mathbf{A}$, $\|\mathbf{A}\|_2 = \sqrt{\mathrm{Tr}\,\mathbf{A}\mathbf{A}^H}$.

In order to take to account the signature of randomness in producing samples using model (4.1) we use Eq.(4.6) to measure the discrepancy in several simulation runs; This approach is known as Monte Carlo simulation [32]. In this method for the fixed covariance matrix $\mathbf{\Sigma}$ one can generate the population set $\{\mathbf{S}_i\}_{i=1}^n$ for N times. Each time the Fréchet mean of the population will be evaluated, $\tilde{n} = 1, 2, ..., N$. Finally, the criterion which is known as "Root Mean Square Error" or (RMSE) is formed as follows :

$$RMSE_{d_F} = \sqrt{\frac{1}{N}\sum_{\tilde{n}=1}^{N} {d_F}^2\left(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}_{\tilde{n}}\right)} \tag{4.6}$$

Next method for measuring the deviation can be defined as the 'Root Mean Square Error' using the metric:

$$d_{R2}\left(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}\right) = \|\mathbf{\Sigma}^{\frac{1}{2}} - \hat{\mathbf{\Sigma}}^{\frac{1}{2}}\|_2 \tag{4.7}$$

Likewise to the Eq.(4.6) we define the Root mean square error as follows:

$$RMSE_{d_{R2}} = \sqrt{\frac{1}{N}\sum_{\tilde{n}=1}^{N} {d_{R2}}^2 \left(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}_{\tilde{n}}\right)} \tag{4.8}$$

The other measure of deviation that will be used in this research is based on the loss function L. It was mentioned in Chapter 2 that the expected value of loss function is called risk function,$\bar{L}$. We follow very similar approach as it was described for $RMSE_{d_F}$; The exception is unlike root mean square error we consider the ensemble average of loss function between true covariance matrix and its estimation as follows:

$$\bar{L} = \frac{1}{N}\sum_{\tilde{n}=1}^{N} L \left(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}_{\tilde{n}}\right) \tag{4.9}$$

In the following section either of the explained models will be analyzed and evaluated through sets of numerical simulations.

4.1.2 Simulation results

In previous chapter the mathematical tools to find the Fréchet mean of covariance matrices was developed. According to the models that we explained and considered in last section, we examine different Fréchet means to obtain the corresponding estimation of mean for the group of symmetric positive definite matrices.

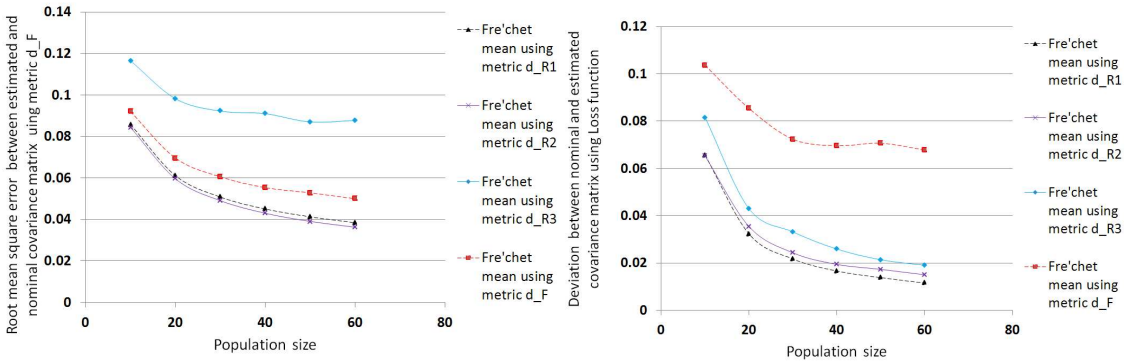
Our objective is to demonstrate that when we consider the inherent structure of manifold of symmetric positive definite matrices it enables us to achieve more accurate estimation by utilizing geometric based means in comparison to the arithmetic mean of covariance matrices.

Example 4.1.1. First we consider the model (4.1) to demonstrate the performance of the Fréchet mean of Riemannian distances. For this reason we consider a Covariance matrix $\mathbf{\Sigma}_{3\times 3}$. The eigenvalues of the covariance matrix is $\lambda = \text{diag}\,[1, 0.3573, 0.065]$. As far as the model (4.1) is concerned the Cholesky factor of the covariance matrix is considered. The additive random noise matrix $\{\mathbf{X}_i\}_{i=1}^{n}$ has independent and identically distributed (i.i.d) entries come from Gaussian distribution with zero mean:

$$\mathbb{E}\left(x_{j,k}^{i}\right) = 0 \quad j,k = 1,2,3 \;\; , i = 1,2,..,n \tag{4.10}$$

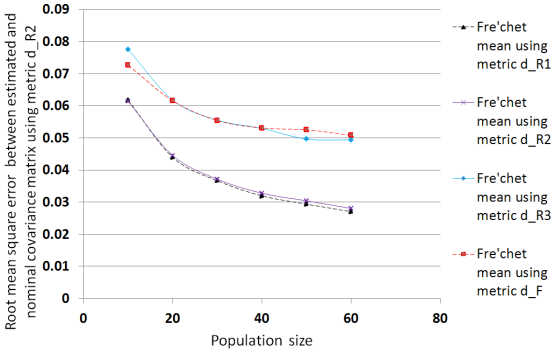
where $\mathbb{E}$ denotes the expected value of the random variable. The standard deviation of the entries of random noise is 0.09 in this experiment.

The population size of the covariance matrices, $\{\mathbf{S}_i\}_{i=1}^{n}$, varies between 10 to 60 in step size 10. In order to take to account the signature of randomness of the additive Gaussian noise matrix $\mathbf{X}_i$'s in model (4.1), for each population we perform the Monte-Carlo simulation 2000 times and obtain the resulting error between $\hat{\mathbf{\Sigma}}$ and the nominal covariance matrix using loss function, $RMSE_{d_F}$ and $RMSE_{d_{R2}}$. The results are shown in Figure 4.2.



(a)

(b)



(c)

Figure 4.2: Error evaluation corresponding to the different Fréchet means using Eqs.(4.6),(4.8) and (4.9). Model (4.1) is used.
(a): Error is measured using metric d_F.
(b): Error is measured using Loss function.
(c): Error is measured using metric d_{R2}.

We observe that the amount of discrepancy between the Fréchet mean using metric d_F, which was shown in Chapter 2 that it is indeed the arithmetic mean of the samples for the given population, and the Fréchet mean of the developed metrics on the manifold of symmetric positive definite matrices using the algorithms that have been studied in Chapter

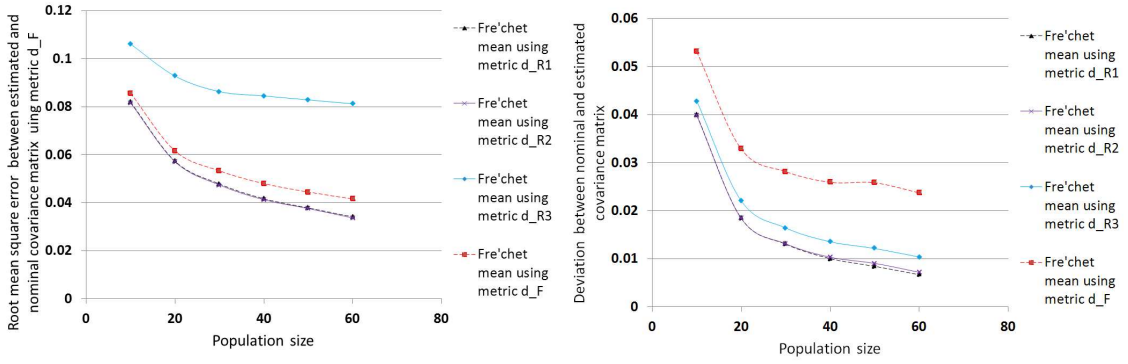
3 are in average quite significant. It can be seen that the Fréchet means of the metrics d_{R1}, d_{R2} and d_{R3} are very close to each other when the bench mark for error is loss function.

As far as the error measurement with respect to $RMSE_{d_{R2}}$ is concerned, estimators based on metrics d_{R1} and d_{R2} demonstrate less RMSE error in comparison to the metrics d_{R3} and d_F. Meanwhile, the RMSE error between Fréchet mean using Riemannian metrics d_{R1} and d_{R2} is very close to each other.

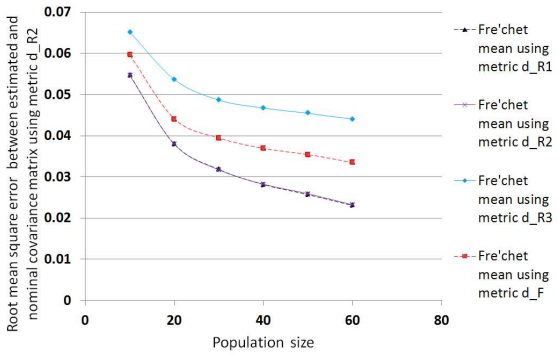
To illustrate the Fréchet mean based on the second model (4.2), we consider the same covariance matrix $\mathbf{\Sigma}_{3\times 3}$ as used for model one. The result of this experiment are depicted in figure(4.3). It can be observed that the behavior of model one and model two are quite similar and metrics d_{R1} and d_{R2} perform in general better than metric d_F.

In the third model the Gaussian noise is added to the square root of $\mathbf{\Sigma}_{3\times 3}$. We observe that when the error is measured using $RMSE_{dF}$, Fréchet mean of metric d_{R2} performs slightly better than the other estimators. In overall metrics d_{R1} and d_{R2} provides smaller error across the different methods of measuring discrepancy.

Regarding to the last model, with the same seeded covariance matrix $\mathbf{\Sigma}_{3\times 3}$, we observe that estimator $\hat{\mathbf{\Sigma}}$ based on metric d_{R3} offers best performance mainly due to the model structure which inherently consider the connection between the manifold and tangent space. The results are provided in Figure 4.5.

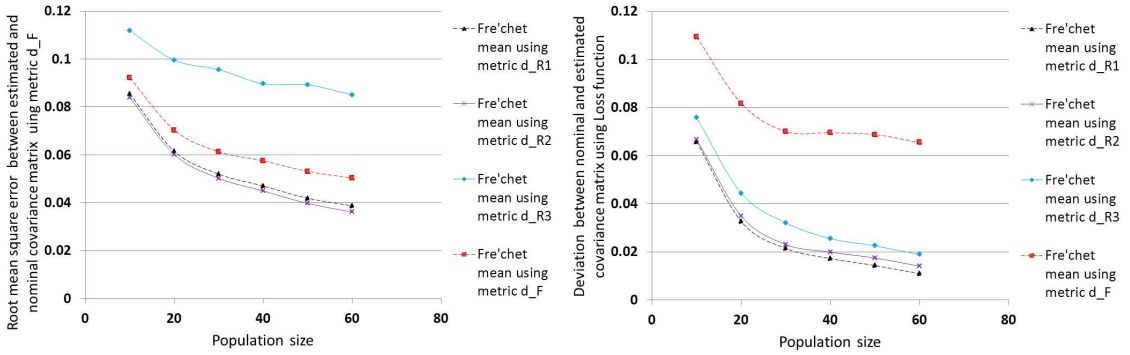


(a) (b)

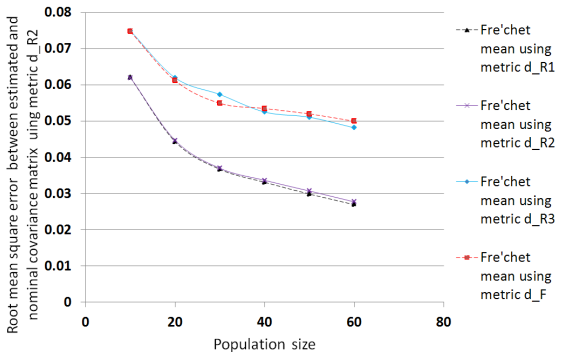


(c)

Figure 4.3: Error evaluation corresponding to the different Fréchet means using Eqs.(4.6),(4.8) and (4.9). Model (4.2) is used.
(a): Error is measured using metric d_F.
(b): Error is measured using Loss function.
(c): Error is measured using metric d_{R2}.

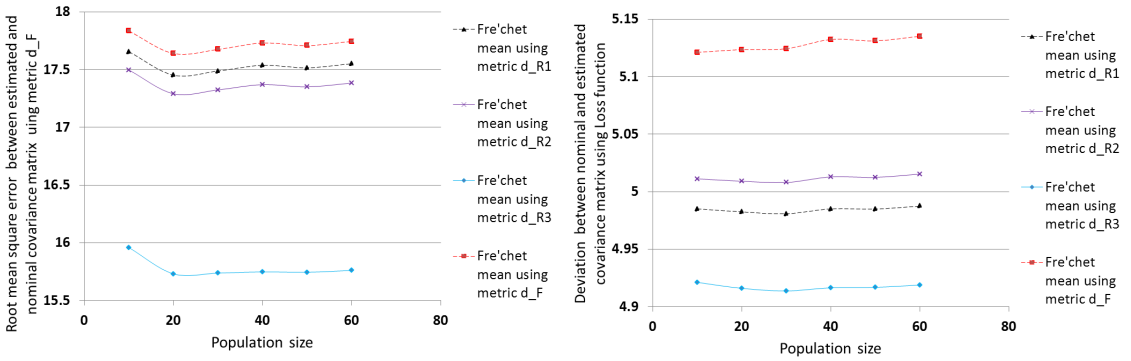


(a) (b)

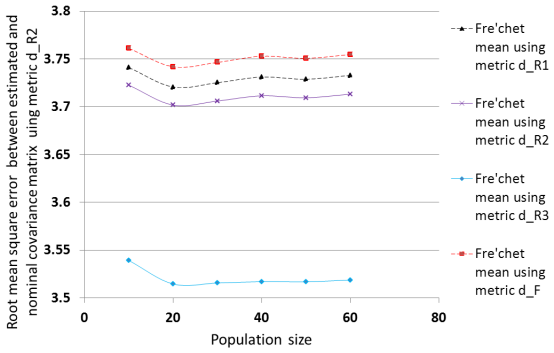


(c)

Figure 4.4: Error evaluation corresponding to the different Fréchet means using Eqs.(4.6),(4.8) and (4.9). Model(4.3) is used.
(a): Error is measured using metric d_F.
(b): Error is measured using Loss function.
(c): Error is measured using metric d_{R2}.



(a) (b)



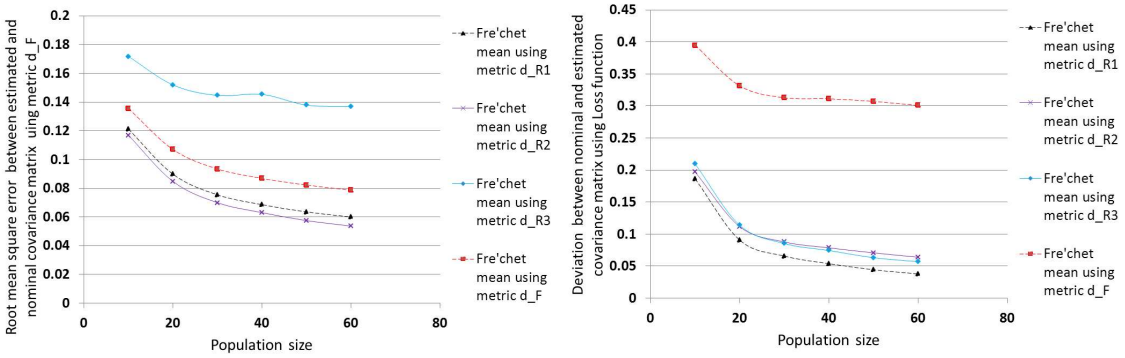
(c)

Figure 4.5: Error evaluation corresponding to the different Fréchet means using Eqs.(4.6),(4.8) and (4.9). Model (4.4) is used.
(a): Error is measured using metric d_F.
(b): Error is measured using Loss function.
(c): Error is measured using metric d_{R2}.

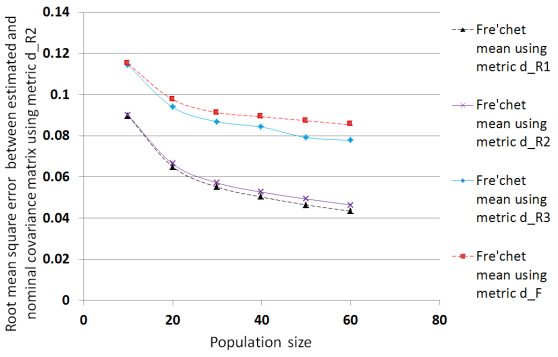
Example 4.1.2. In this example we seed a covariance matrix $\mathbf{\Sigma}_{5\times 5}$, which is obtained from five sources of measurement . The eigenvalues of this covariance matrix is

$$\lambda = \text{diag}\left[1, 0.5860, 0.3602, 0.1427, 0.0466\right].$$

The standard deviation of additive noise to the models in this experiment is the same as example one. We observe that when the models (4.1), (4.2) and model(4.3) are used, as the population size is increased, the error is deceasing by using different measurements. In the last model it can be seen that Fréchet mean of metric d_{R3} provides smaller error among other metrics; We expect that due to the structure of this model which generates the population by projecting back the logarithm of Cholesky factor of $\mathbf{\Sigma}$ using exponential map . Based on the examples we expect that the Fréchet means of the metrics which have arisen from Riemannian manifold reflect better performance as oppose to the Fréchet mean of metric d_F.

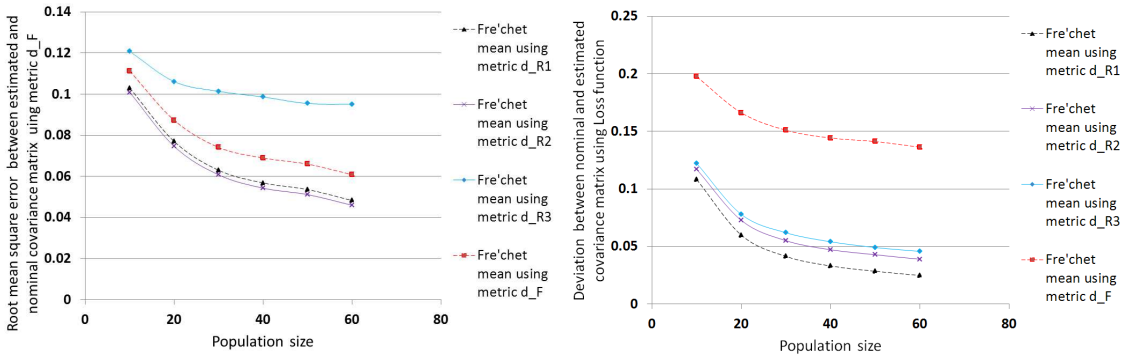


(a) (b)

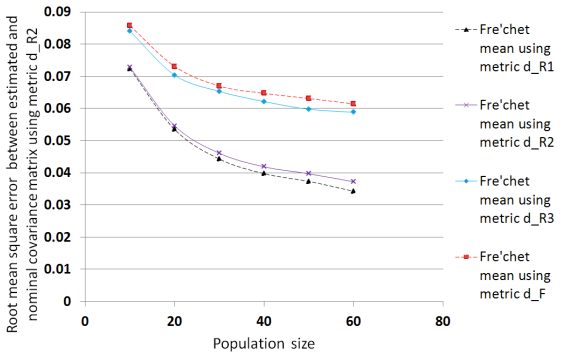


(c)

Figure 4.6: Error evaluation corresponding to the different Fréchet means using Eqs.(4.6),(4.8) and (4.9).Model(4.1) is used.
(a): Error is measured using metric d_F.
(b): Error is measured using Loss function.
(c): Error is measured using metric d_{R2}.

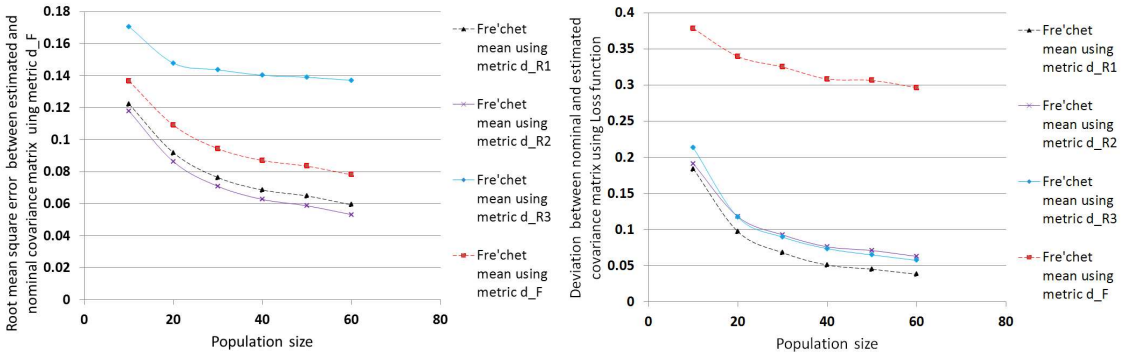


(a) (b)

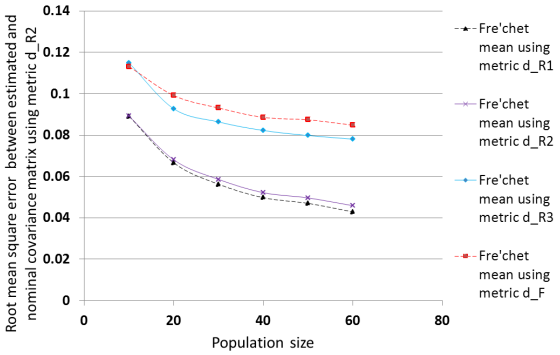


(c)

Figure 4.7: Error evaluation corresponding to the different Fréchet means using Eqs.(4.6),(4.8) and (4.9). Model (4.2) is used.
(a): Error is measured using metric d_F.
(b): Error is measured using Loss function.
(c): Error is measured using metric d_{R2}.

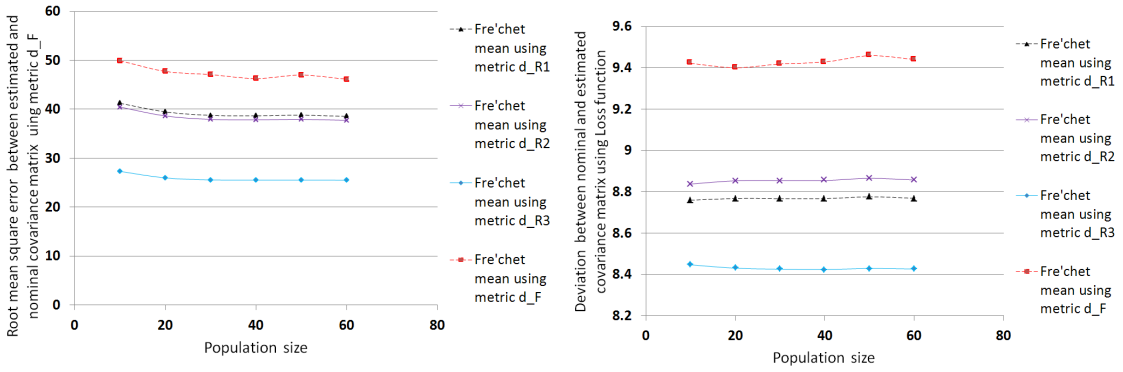


(a) (b)

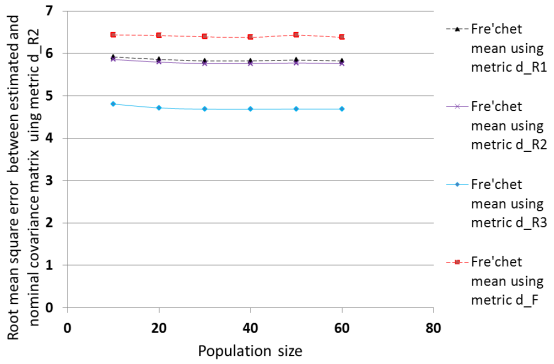


(c)

Figure 4.8: Error evaluation corresponding to the different Fréchet means using Eqs.(4.6),(4.8) and (4.9). Model(4.3) is used.
(a): Error is measured using metric d_F.
(b): Error is measured using Loss function.
(c): Error is measured using metric d_{R2}.



(a) (b)



(c)

Figure 4.9: Error evaluation corresponding to the different Fréchet means using Eqs.(4.6),(4.8) and (4.9).Model(4.4) is used.
(a): Error is measured using metric d_F.
(b): Error is measured using Loss function.
(c): Error is measured using metric d_{R2}.

Tables (4.1) to (4.8) show the distance between Fréchet means of studied metrics ,according to the different models, to the corresponding nominal value using three methods of measurements . In the tables n represents the population size. Q indicates the size of nominal covariance matrix. The estimators $\hat{\Sigma}_{d_{R1}}$, $\hat{\Sigma}_{d_{R2}}$, $\hat{\Sigma}_{d_{R3}}$ and $\hat{\Sigma}_{d_F}$ are corresponding to

the Fréchet mean of metrics d_{R1}, d_{R2}, d_{R3} and d_F respectively. The discrepancy has been evaluated using the benchmarks $RMSE_{d_F}$, $RMSE_{d_{R2}}$ and 'Average of loss function' respectively through out the tables. The eigenvalues of each seeded covariance matrix are within range $[0, 1]$.

As far as the closeness of each estimator of population mean is concerned, one can observe that when the estimation $\hat{\mathbf{\Sigma}}$ is obtained by Fréchet mean corresponding to the metric d_{R2} or d_{R1} and the error is measured by using criterion (4.6) or (4.8) the error is smaller than the estimations $\hat{\mathbf{\Sigma}}_{dR3}$ and $\hat{\mathbf{\Sigma}}_{d_F}$. Regarding to estimation based on metric d_{R3}, we observe that it provides higher root mean square error in compare to the estimators using metrics d_F, d_{R1} and d_{R2} when the model is based on the matrix Choleskey factor or the matrix square root and the error is measured by using metric d_F; On the other hand when we use the loss function to measure the distance we observe that Fréchet means of metrics $\hat{\mathbf{\Sigma}}_{dR1}$,$\hat{\mathbf{\Sigma}}_{dR2}$ and $\hat{\mathbf{\Sigma}}_{dR3}$ perform better than the metric d_F on the same model.

The distance between 'true' convenience matrix and estimated one, depending on the type of metric, has also been provided for the exponential model (4.4) in tables (4.7) and (4.8) respectively. The results illustrate that the mean of the population resulting from metric d_{R3} has smaller error in comparison to the other estimators with respect to the different metrics of measuring errors.

		$RMSE_{d_F}$				$RMSE_{d_{R2}}$				Average of Loss Function			
n	Q	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$
10	3	0.086	0.084	0.117	0.092	0.062	0.061	0.078	0.073	0.066	0.065	0.082	0.104
20	3	0.061	0.060	0.098	0.070	0.044	0.045	0.062	0.062	0.032	0.035	0.043	0.085
30	3	0.051	0.049	0.092	0.061	0.037	0.037	0.056	0.055	0.022	0.024	0.033	0.072
40	3	0.045	0.043	0.091	0.055	0.032	0.033	0.053	0.053	0.017	0.019	0.026	0.070
50	3	0.041	0.039	0.087	0.053	0.029	0.030	0.050	0.053	0.014	0.017	0.021	0.071
60	3	0.038	0.036	0.088	0.050	0.027	0.028	0.049	0.051	0.012	0.015	0.019	0.068
10	4	0.100	0.097	0.162	0.109	0.078	0.077	0.106	0.097	0.148	0.149	0.179	0.276
20	4	0.072	0.069	0.146	0.085	0.058	0.058	0.088	0.085	0.080	0.089	0.099	0.240
30	4	0.062	0.058	0.142	0.076	0.049	0.050	0.082	0.079	0.056	0.069	0.075	0.227
40	4	0.053	0.049	0.139	0.068	0.043	0.044	0.078	0.075	0.042	0.056	0.059	0.216
50	4	0.052	0.046	0.138	0.066	0.040	0.042	0.076	0.073	0.036	0.050	0.051	0.214
60	4	0.047	0.042	0.137	0.062	0.037	0.039	0.075	0.072	0.031	0.046	0.047	0.211
10	5	0.121	0.117	0.172	0.135	0.090	0.090	0.114	0.115	0.186	0.197	0.209	0.395
20	5	0.090	0.085	0.152	0.107	0.065	0.066	0.094	0.098	0.091	0.112	0.114	0.331
30	5	0.075	0.070	0.145	0.093	0.055	0.057	0.087	0.091	0.066	0.088	0.085	0.313
40	5	0.069	0.063	0.145	0.087	0.050	0.053	0.084	0.089	0.054	0.079	0.074	0.311
50	5	0.064	0.058	0.138	0.082	0.046	0.049	0.079	0.087	0.044	0.071	0.063	0.307
60	5	0.060	0.054	0.137	0.079	0.043	0.046	0.078	0.085	0.038	0.064	0.057	0.301

Table 4.1: Error analysis of different estimators using model (4.1) for $Q = 3, 4, 5$. The dimension of seeded covariance matrices varies depending on the choice of population.

		$RMSE_{d_F}$				$RMSE_{d_{R2}}$				Average of Loss Function			
n	Q	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$
10	6	0.138	0.132	0.152	0.156	0.102	0.101	0.106	0.124	0.189	0.187	0.180	0.304
20	6	0.105	0.098	0.131	0.125	0.077	0.077	0.084	0.104	0.102	0.110	0.096	0.249
30	6	0.088	0.081	0.121	0.109	0.064	0.065	0.074	0.094	0.069	0.079	0.066	0.221
40	6	0.078	0.072	0.117	0.100	0.058	0.059	0.070	0.090	0.058	0.070	0.055	0.216
50	6	0.074	0.067	0.114	0.096	0.054	0.055	0.066	0.087	0.049	0.063	0.046	0.212
60	6	0.069	0.062	0.112	0.091	0.050	0.052	0.064	0.084	0.043	0.056	0.040	0.204
10	7	0.166	0.158	0.166	0.187	0.111	0.110	0.112	0.136	0.206	0.211	0.196	0.360
20	7	0.122	0.114	0.138	0.146	0.082	0.083	0.087	0.113	0.108	0.121	0.103	0.289
30	7	0.106	0.098	0.129	0.131	0.071	0.073	0.078	0.105	0.080	0.098	0.077	0.276
40	7	0.094	0.086	0.123	0.119	0.064	0.066	0.073	0.100	0.065	0.083	0.063	0.262
50	7	0.088	0.081	0.117	0.114	0.060	0.062	0.068	0.097	0.056	0.075	0.053	0.258
60	7	0.082	0.074	0.115	0.108	0.056	0.058	0.066	0.095	0.050	0.070	0.049	0.253
10	8	0.182	0.173	0.166	0.207	0.122	0.120	0.115	0.147	0.242	0.242	0.216	0.394
20	8	0.135	0.126	0.142	0.163	0.092	0.092	0.091	0.123	0.136	0.147	0.118	0.325
30	8	0.117	0.108	0.129	0.145	0.079	0.080	0.079	0.113	0.101	0.116	0.084	0.303
40	8	0.103	0.095	0.122	0.132	0.071	0.073	0.074	0.107	0.081	0.098	0.068	0.286
50	8	0.098	0.090	0.116	0.127	0.067	0.068	0.069	0.104	0.071	0.090	0.057	0.284
60	8	0.092	0.083	0.113	0.120	0.062	0.064	0.066	0.100	0.063	0.082	0.051	0.276

Table 4.2: Error analysis of different estimators using model (4.1) for $Q = 6, 7, 8$.

		$RMSE_{d_F}$				$RMSE_{d_{R2}}$				Average of Loss Function			
n	Q	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$
10	3	0.082	0.082	0.106	0.086	0.055	0.055	0.065	0.060	0.040	0.040	0.043	0.053
20	3	0.057	0.057	0.093	0.062	0.038	0.038	0.054	0.044	0.018	0.018	0.022	0.033
30	3	0.048	0.047	0.086	0.053	0.032	0.032	0.049	0.039	0.013	0.013	0.016	0.028
40	3	0.042	0.041	0.084	0.048	0.028	0.028	0.047	0.037	0.010	0.010	0.014	0.026
50	3	0.038	0.038	0.083	0.044	0.026	0.026	0.045	0.035	0.008	0.009	0.012	0.026
60	3	0.034	0.034	0.081	0.042	0.023	0.023	0.044	0.034	0.007	0.007	0.010	0.024
10	4	0.083	0.082	0.091	0.089	0.061	0.061	0.068	0.070	0.081	0.079	0.124	0.100
20	4	0.062	0.061	0.076	0.070	0.047	0.048	0.054	0.060	0.047	0.047	0.080	0.078
30	4	0.051	0.050	0.070	0.061	0.040	0.040	0.049	0.054	0.033	0.033	0.064	0.065
40	4	0.043	0.042	0.067	0.053	0.035	0.035	0.045	0.051	0.025	0.025	0.056	0.057
50	4	0.039	0.038	0.064	0.051	0.033	0.034	0.043	0.049	0.023	0.023	0.054	0.055
60	4	0.037	0.037	0.063	0.049	0.031	0.032	0.042	0.048	0.021	0.021	0.051	0.052
10	5	0.103	0.101	0.121	0.111	0.072	0.073	0.084	0.086	0.108	0.117	0.122	0.197
20	5	0.077	0.075	0.106	0.087	0.053	0.055	0.070	0.073	0.060	0.073	0.078	0.166
30	5	0.063	0.061	0.101	0.074	0.044	0.046	0.065	0.067	0.041	0.055	0.062	0.151
40	5	0.057	0.054	0.099	0.069	0.040	0.042	0.062	0.065	0.033	0.047	0.054	0.144
50	5	0.054	0.051	0.095	0.066	0.037	0.040	0.060	0.063	0.028	0.043	0.049	0.141
60	5	0.048	0.046	0.095	0.061	0.034	0.037	0.059	0.061	0.025	0.039	0.046	0.136

Table 4.3: Error analysis of different estimators using model(4.2). $Q = 3, 4, 5$.

		$RMSE_{d_F}$				$RMSE_{d_{R2}}$				Average of Loss Function			
n	Q	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$
10	6	0.112	0.110	0.145	0.123	0.084	0.084	0.095	0.100	0.130	0.133	0.130	0.202
20	6	0.083	0.082	0.126	0.097	0.063	0.065	0.078	0.085	0.074	0.082	0.075	0.163
30	6	0.068	0.068	0.122	0.083	0.054	0.055	0.072	0.078	0.054	0.062	0.059	0.144
40	6	0.061	0.060	0.116	0.076	0.048	0.050	0.068	0.074	0.044	0.053	0.050	0.137
50	6	0.055	0.054	0.112	0.072	0.045	0.047	0.064	0.072	0.038	0.048	0.043	0.133
60	6	0.051	0.051	0.111	0.069	0.042	0.045	0.063	0.071	0.035	0.045	0.041	0.132
10	7	0.133	0.129	0.140	0.145	0.087	0.086	0.091	0.098	0.110	0.111	0.111	0.153
20	7	0.097	0.094	0.118	0.112	0.064	0.064	0.073	0.080	0.061	0.064	0.064	0.114
30	7	0.084	0.080	0.107	0.100	0.055	0.055	0.066	0.072	0.045	0.049	0.050	0.100
40	7	0.075	0.071	0.101	0.092	0.049	0.049	0.061	0.067	0.035	0.039	0.040	0.091
50	7	0.069	0.066	0.098	0.087	0.046	0.046	0.059	0.064	0.030	0.035	0.036	0.088
60	7	0.065	0.061	0.095	0.082	0.043	0.043	0.056	0.061	0.026	0.031	0.032	0.083
10	8	0.137	0.134	0.151	0.153	0.096	0.095	0.100	0.113	0.153	0.157	0.151	0.227
20	8	0.102	0.100	0.130	0.121	0.073	0.074	0.080	0.095	0.090	0.099	0.091	0.181
30	8	0.086	0.084	0.124	0.107	0.063	0.065	0.073	0.088	0.067	0.077	0.069	0.163
40	8	0.078	0.075	0.118	0.099	0.057	0.059	0.068	0.083	0.056	0.066	0.059	0.153
50	8	0.072	0.070	0.114	0.094	0.054	0.056	0.065	0.081	0.048	0.059	0.052	0.147
60	8	0.068	0.066	0.111	0.090	0.051	0.054	0.063	0.079	0.045	0.056	0.049	0.146

Table 4.4: Error analysis of different estimators using model (4.2). $Q = 6, 7, 8$.

		$RMSE_{d_F}$				$RMSE_{d_{R2}}$				Average of Loss Function			
n	Q	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$
10	3	0.085	0.084	0.112	0.092	0.066	0.067	0.076	0.109	0.062	0.062	0.075	0.075
20	3	0.062	0.060	0.100	0.070	0.033	0.035	0.044	0.082	0.044	0.045	0.062	0.061
30	3	0.052	0.050	0.096	0.061	0.021	0.023	0.032	0.070	0.037	0.037	0.057	0.055
40	3	0.047	0.045	0.090	0.058	0.017	0.020	0.026	0.070	0.033	0.034	0.052	0.053
50	3	0.042	0.040	0.089	0.053	0.014	0.017	0.023	0.069	0.030	0.031	0.051	0.052
60	3	0.039	0.036	0.085	0.050	0.011	0.014	0.019	0.065	0.027	0.028	0.048	0.050
10	4	0.096	0.093	0.156	0.106	0.143	0.146	0.174	0.277	0.076	0.076	0.103	0.097
20	4	0.073	0.069	0.144	0.085	0.083	0.092	0.100	0.244	0.058	0.059	0.086	0.085
30	4	0.061	0.057	0.144	0.075	0.056	0.069	0.074	0.230	0.049	0.050	0.083	0.079
40	4	0.053	0.049	0.139	0.068	0.041	0.055	0.058	0.215	0.042	0.044	0.078	0.074
50	4	0.049	0.044	0.137	0.064	0.036	0.050	0.053	0.210	0.040	0.042	0.076	0.073
60	4	0.047	0.042	0.137	0.062	0.031	0.045	0.047	0.206	0.037	0.039	0.075	0.071
10	5	0.122	0.118	0.171	0.137	0.184	0.191	0.213	0.378	0.089	0.089	0.115	0.113
20	5	0.092	0.086	0.148	0.109	0.097	0.118	0.117	0.339	0.067	0.068	0.093	0.099
30	5	0.076	0.071	0.144	0.094	0.068	0.092	0.089	0.325	0.056	0.059	0.086	0.093
40	5	0.069	0.063	0.140	0.087	0.051	0.076	0.073	0.308	0.050	0.052	0.082	0.089
50	5	0.065	0.059	0.139	0.083	0.045	0.071	0.065	0.306	0.047	0.050	0.080	0.087
60	5	0.059	0.053	0.137	0.078	0.038	0.063	0.057	0.296	0.043	0.046	0.078	0.085

Table 4.5: Error analysis of different estimators using model (4.3). $Q = 3, 4, 5$.

		$RMSE_{d_F}$				$RMSE_{d_{R2}}$				Average of Loss Function			
n	Q	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$
10	6	0.136	0.130	0.151	0.154	0.182	0.181	0.175	0.298	0.100	0.099	0.105	0.122
20	6	0.105	0.098	0.128	0.125	0.100	0.108	0.092	0.248	0.076	0.076	0.082	0.104
30	6	0.087	0.080	0.123	0.109	0.070	0.080	0.066	0.221	0.064	0.064	0.075	0.094
40	6	0.079	0.072	0.116	0.101	0.057	0.069	0.053	0.216	0.058	0.059	0.069	0.090
50	6	0.074	0.067	0.111	0.096	0.050	0.063	0.045	0.210	0.054	0.056	0.065	0.087
60	6	0.070	0.063	0.111	0.092	0.045	0.059	0.041	0.209	0.051	0.053	0.064	0.085
10	7	0.163	0.156	0.165	0.185	0.203	0.207	0.193	0.354	0.110	0.110	0.111	0.135
20	7	0.123	0.116	0.141	0.148	0.112	0.126	0.104	0.296	0.084	0.084	0.088	0.114
30	7	0.104	0.097	0.129	0.129	0.080	0.096	0.076	0.272	0.071	0.072	0.078	0.104
40	7	0.094	0.086	0.121	0.119	0.065	0.083	0.062	0.262	0.064	0.066	0.072	0.100
50	7	0.088	0.080	0.117	0.113	0.056	0.076	0.054	0.259	0.059	0.061	0.068	0.097
60	7	0.082	0.074	0.114	0.108	0.049	0.070	0.047	0.253	0.056	0.058	0.066	0.095
10	8	0.177	0.169	0.170	0.203	0.239	0.237	0.218	0.389	0.120	0.118	0.118	0.145
20	8	0.134	0.126	0.141	0.162	0.137	0.148	0.118	0.325	0.092	0.092	0.090	0.123
30	8	0.115	0.107	0.130	0.144	0.099	0.114	0.084	0.300	0.079	0.080	0.080	0.113
40	8	0.103	0.095	0.121	0.132	0.080	0.097	0.067	0.285	0.071	0.072	0.073	0.106
50	8	0.098	0.089	0.116	0.126	0.071	0.089	0.057	0.280	0.066	0.068	0.068	0.103
60	8	0.093	0.084	0.113	0.122	0.064	0.083	0.050	0.277	0.063	0.065	0.066	0.101

Table 4.6: Error analysis of different estimators using model (4.3). $Q = 6, 7, 8$.

		$RMSE_{d_F}$				$RMSE_{d_{R2}}$				Average of Loss Function			
n	Q	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$
10	3	17.654	17.496	15.958	17.833	3.741	3.723	3.539	3.761	4.985	5.011	4.921	5.121
20	3	17.451	17.290	15.731	17.639	3.721	3.702	3.514	3.742	4.982	5.009	4.916	5.123
30	3	17.487	17.324	15.737	17.675	3.725	3.706	3.516	3.746	4.981	5.008	4.914	5.124
40	3	17.537	17.369	15.747	17.730	3.731	3.711	3.517	3.753	4.985	5.013	4.916	5.132
50	3	17.515	17.350	15.744	17.708	3.729	3.709	3.517	3.751	4.985	5.012	4.917	5.131
60	3	17.551	17.383	15.760	17.744	3.733	3.713	3.519	3.755	4.987	5.015	4.919	5.135
10	4	12.431	12.260	11.126	12.848	3.292	3.268	3.105	3.346	7.780	7.794	7.612	7.976
20	4	12.283	12.100	10.923	12.684	3.276	3.250	3.079	3.331	7.787	7.802	7.606	8.001
30	4	12.226	12.046	10.880	12.628	3.269	3.244	3.074	3.324	7.782	7.797	7.599	7.999
40	4	12.250	12.067	10.891	12.648	3.275	3.249	3.077	3.329	7.791	7.807	7.606	8.012
50	4	12.251	12.066	10.881	12.657	3.275	3.248	3.075	3.330	7.787	7.802	7.600	8.012
60	4	12.243	12.058	10.872	12.649	3.274	3.248	3.074	3.330	7.795	7.811	7.607	8.021
10	5	41.228	40.438	27.286	49.914	5.916	5.856	4.805	6.437	8.760	8.838	8.447	9.422
20	5	39.467	38.630	25.968	47.737	5.859	5.794	4.715	6.422	8.768	8.853	8.431	9.400
30	5	38.759	37.927	25.563	47.075	5.825	5.759	4.684	6.393	8.767	8.854	8.426	9.419
40	5	38.674	37.843	25.513	46.282	5.826	5.760	4.683	6.377	8.767	8.855	8.424	9.428
50	5	38.786	37.941	25.534	46.984	5.842	5.775	4.688	6.428	8.776	8.865	8.429	9.459

Table 4.7: Error analysis of different estimators using model (4.4). $Q = 3, 4, 5$.

		$RMSE_{d_F}$				$RMSE_{d_{R2}}$				Average of Loss Function			
n	Q	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$	$\hat{\mathbf{\Sigma}}_{dR1}$	$\hat{\mathbf{\Sigma}}_{dR2}$	$\hat{\mathbf{\Sigma}}_{dR3}$	$\hat{\mathbf{\Sigma}}_{d_F}$
60	5	38.550	37.722	25.509	46.144	5.825	5.758	4.686	6.380	8.769	8.857	8.427	9.441
10	6	8.832	8.623	6.282	12.643	2.500	2.471	2.129	2.883	8.229	8.251	7.853	8.666
20	6	7.942	7.754	5.863	10.694	2.403	2.374	2.067	2.778	8.255	8.277	7.845	8.753
30	6	7.840	7.655	5.775	11.136	2.389	2.360	2.053	2.792	8.273	8.295	7.853	8.792
40	6	7.690	7.508	5.690	10.370	2.365	2.336	2.038	2.753	8.278	8.299	7.852	8.807
50	6	7.700	7.515	5.679	10.475	2.368	2.338	2.037	2.775	8.283	8.305	7.852	8.829
60	6	7.614	7.433	5.637	10.440	2.355	2.325	2.030	2.761	8.280	8.302	7.848	8.828
10	7	81.707	80.008	23.231	250.783	6.936	6.843	4.120	9.909	8.726	8.873	8.106	9.845
20	7	57.025	55.469	21.026	140.791	6.567	6.466	4.005	9.359	8.780	8.939	8.109	10.110
30	7	57.485	55.824	20.609	195.313	6.654	6.548	3.982	9.998	8.831	9.001	8.118	10.321
40	7	54.905	53.270	20.424	181.802	6.631	6.522	3.981	10.151	8.854	9.029	8.132	10.435
50	7	54.235	52.648	20.285	182.139	6.597	6.490	3.966	10.093	8.845	9.019	8.125	10.436
60	7	52.785	51.202	19.935	172.054	6.555	6.447	3.938	10.087	8.848	9.024	8.116	10.500
10	8	54.783	52.886	14.826	168.105	5.675	5.544	3.197	8.318	9.428	9.623	8.698	10.775
20	8	50.473	48.696	13.319	382.265	5.508	5.369	3.073	9.434	9.469	9.683	8.674	11.081
30	8	36.592	34.803	12.880	113.666	5.220	5.076	3.028	7.965	9.464	9.682	8.671	11.173
40	8	35.301	33.538	12.833	90.327	5.182	5.036	3.024	7.648	9.472	9.692	8.676	11.250
50	8	35.638	33.848	12.655	127.582	5.206	5.058	3.006	8.223	9.482	9.705	8.666	11.345
60	8	35.661	33.840	12.579	207.488	5.193	5.042	2.999	8.313	9.483	9.710	8.665	11.380

Table 4.8: Error analysis of different estimators using model (4.4). $Q = 6, 7, 8$.

4.2 Classification based on the distance to the center of mass

So far we have mathematically developed the concept of mean for group of positive definite Hermitian matrices on manifold $\mathcal{M}$ from the distance point of view. Moreover, we have seen that depending on model and the criterion of measuring the closeness of each estimator to the nominal covariance matrix, Fréchet means of Riemannian distances are better estimators.

The concept of Fréchet mean can be utilized in distance based detection and classification on manifold $\mathcal{M}$ [33], [34]. For this purpose suppose that we have a set of covariance matrices $\{\mathbf{S}_{ik}\}_{i=1}^{n_k}$ where k represents the label of each class and n_k denotes the number of covariance matrices within k^{th} class. For each class k the Fréchet mean of the class, depending on type of metric, can be obtained as representative of each class. For the unknown observation its covariance matrix is formed and considered as the unknown feature. The observation is assigned to the class which has minimum distance to the Fréchet mean of the class. This method can be recapitulated in form of the following algorithm.

Algorithm 3 Distance to the center of mass algorithm

1. Input: the given known classes $1, 2, 3, , ..., k$ and set of covariance matrices $\{\mathbf{S}_{ik}\}_{i=1}^{n_k}$ within each class.

2. For each class k compute $\hat{\mathbf{\Sigma}}_{ik}$ as the Fréchet mean of $\{\mathbf{S}_{ik}\}_{i=1}^{n_k}$.

3. For the covariance matrix $\mathbf{S}$ of unknown observation compute

$$\hat{k} = \arg\min_{k} d\left(\mathbf{S}, \mathbf{\Sigma}_k\right). \tag{4.11}$$

4. The covariance matrix $\mathbf{S}$ corresponding to the unknown observation in step 3 will be assigned to class $\hat{k}$.

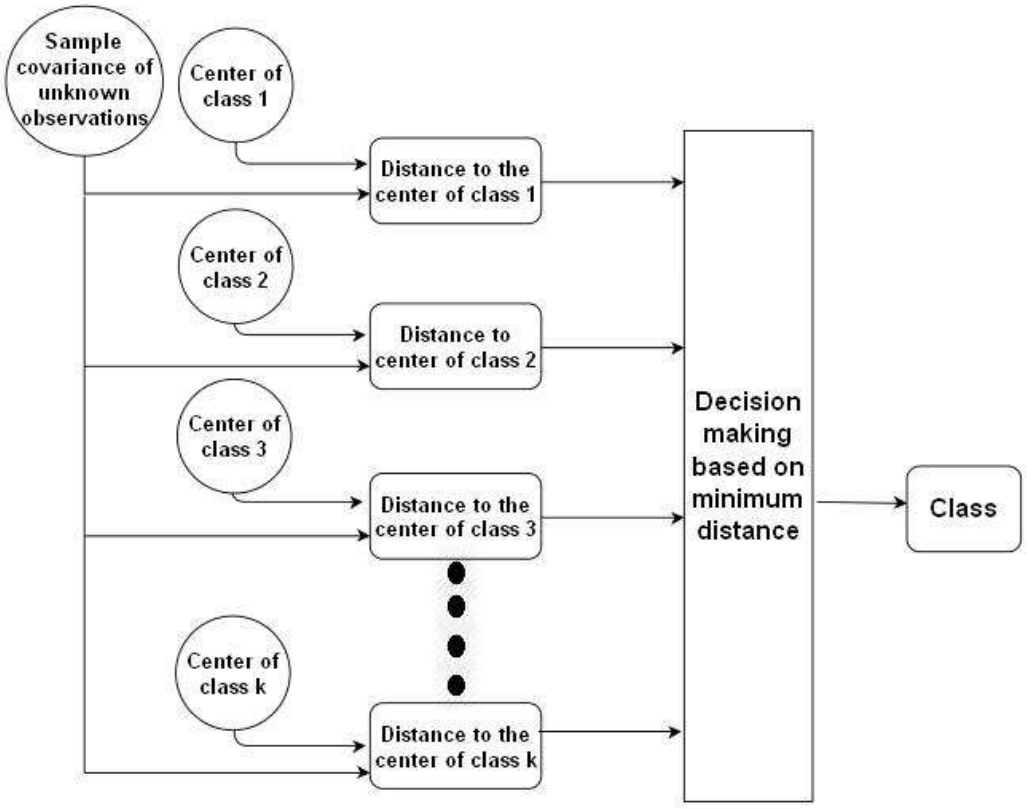


Figure 4.10: Process of decision making according to the concept of center of each class

In order to inspect and evaluate the Algorithm 3 we perform it on the simulated data set. For this purpose we consider three classes $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ consisting of samples $\{\mathbf{x}_1(i)\}_{i=1}^{10000}$, $\{\mathbf{x}_2(i)\}_{i=1}^{10000}$ and $\{\mathbf{x}_3(i)\}_{i=1}^{10000}$ drawn from the normal distribution with zero mean and covariance matrices $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$ and $\mathbf{\Sigma}_3$ respectively. At the same time Gaussian random noise with mean zero and standard deviation σ is added to the samples of both classes. Then we split each class to the half for train and test purpose and perform two fold cross validation .

At training step we consider training sets $\mathcal{C}_{1train}$, $\mathcal{C}_{2train}$ and $\mathcal{C}_{3train}$. From each training set $\mathcal{C}_{j_{train}}$, $j = 1, 2, 3$; we form a sequence of $\{\mathbf{X}_{k,j}\}$ of observations $k = 1, 2, .., 20$. Each observation $\{\mathbf{X}_{k,j}\}$ has 40 samples which can be shown as $[\mathbf{x}_{jk}(1), \mathbf{x}_{jk}(2), ..., \mathbf{x}_{jk}(40)]^T$.

The Frechét mean of the covariance matrices $\{\mathbf{S}_{k,j}\}$ of the observation $\{\mathbf{X}_{k,j}\}$ are obtained using metrics d_{R_1},d_{R_2}, d_{R_3} and d_F respectively. The method of distance to the center of mass is performed to classify the new observation $\mathbf{X}_{test}$ according to its observed covariance matrix $\mathbf{S}_{test}$.

From [35] it has been known that when we have sample of observations from a p-variate normal distribution with zero mean and covariance matrix $\mathbf{\Sigma}$ then $N\bar{\mathbf{x}}\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}^T$ has chi-square distribution with p degrees of freedom (see Appendix B); where $\bar{\mathbf{x}}$ is the sample mean vector of size $1 \times p$ for the observation $\mathbf{X}$ of size $N \times p$; which is taken over columns of $\mathbf{X}$ and N is the sample size. When the sample size is fairly large we can replace $\mathbf{\Sigma}$ with $\hat{\mathbf{\Sigma}}$ [17]. As a result, a new observation $\mathbf{X}_{test}$ is classified to class j whenever:

$$\chi^2_p(1 - \alpha/2) \leq N\bar{\mathbf{x}}_{test}\hat{\mathbf{\Sigma}}_j^{-1}\bar{\mathbf{x}}^T_{test} \leq \chi^2_p(\alpha/2) \tag{4.12}$$

where $\chi^2_p(\alpha)$ is given by:

$$P\left(\chi^2_p > \chi^2_p(\alpha)\right) = \alpha$$

In Eq.(4.12) the significant level is set to be $\alpha = 0.05$. At the same time we also compare the result of distance to the center of mass in classification with the result of Eq.(4.12).

To illustrate the results of our classifiers using the simulated data as explained earlier, we consider two experiments: the first one consists of two classes $\mathcal{C}_1$ and $\mathcal{C}_2$ of observations

with prescribed covariance matrices[1] $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ of size 4×4 with parameters $\rho_1 = 0.75$, $\rho_2 = 0.8$ respectively. The second model consists of three classes $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ such that the classes $\mathcal{C}_1$ and $\mathcal{C}_2$ have the same covariance matrices as the first experiment. The third class $\mathcal{C}_3$, on the other hand, has the covariance matrix $\mathbf{\Sigma}_3$ of size $p \times p$ with parameter $\rho_3 = 0.85$. The standard deviation of additive Gaussian noise to the observations of each classes in both experiment is $\sigma = 0.1$.

The results are shown in Figures 4.11 and 4.12 for the first model of classification. As far as the probability of correct classification is concerned, the Reimannian classifiers based on metrics d_{R1}, d_{R2} and d_{R3} on average have smaller probability of missclassification in comparison to the classifier based on Euclidean distance d_F. Furthermore, we compare our result with the classifier using Eq.(4.12) we note that in this method rather than forming the covariance matrix of the observations we classify the observations based on the mean. The accuracy of this approach compared to the classifiers using Fréchet mean is not satisfactory; see Tables 4.9 and 4.10.

[1] In this work by prescribed covariance matrix we mean that the covariance matrix $\mathbf{\Sigma}$ of size $M \times M$ has the form of $\mathbf{\Sigma} = (\sigma_{ll'})_{M \times M}$ where $1 \leq l, l' \leq M$ and $\sigma_{ll'} = \rho^{|l-l'|}$. ρ is a constant between 0 and 1. This type of covariance matrix is also known as the structured covariance matrix.

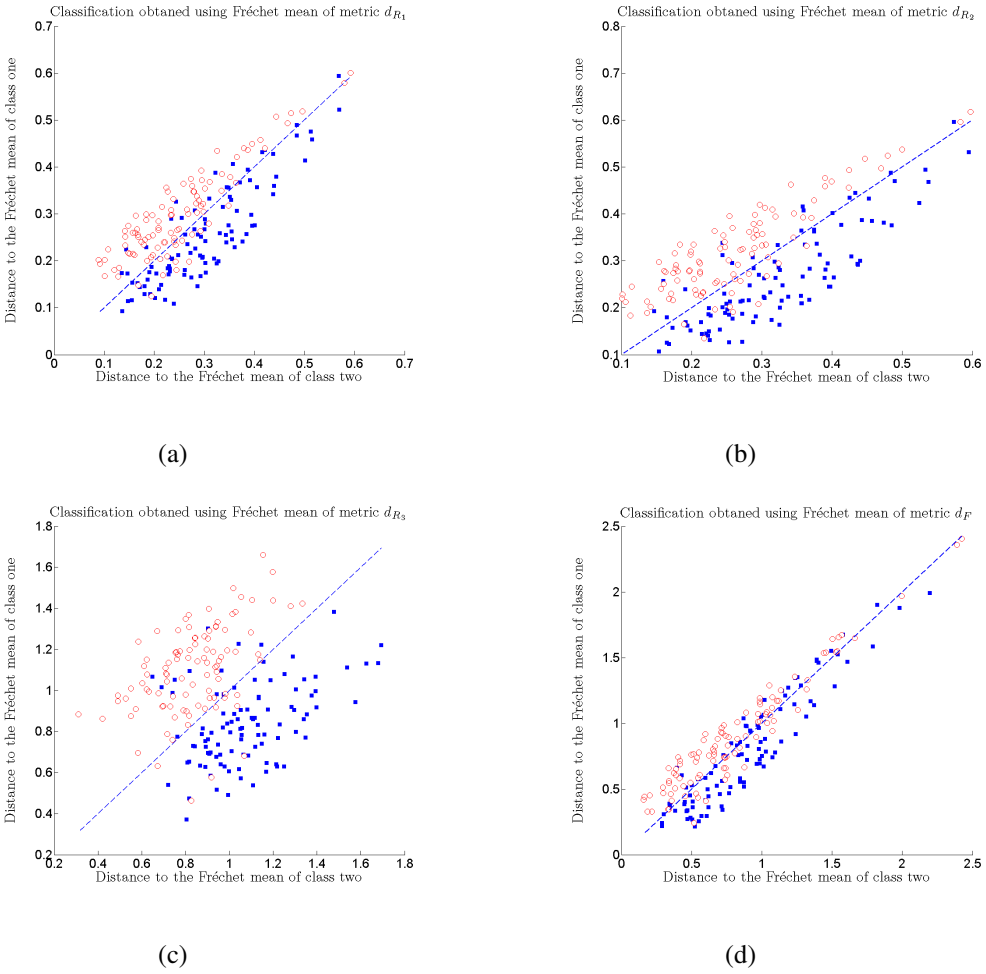
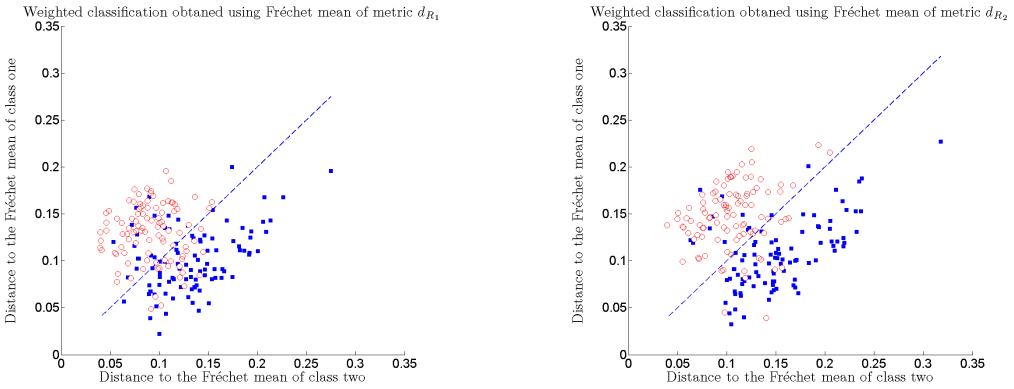
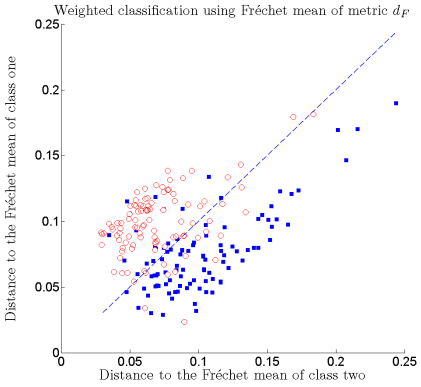


Figure 4.11: Classification based on the distance to the Fréchet mean of each class. The dash line shows the border such that the distance between the center of either class one or two are the same and decision can be made based on it. The solid square belongs to the class one; the circles represent class two. The distance of unknown covariance matrix has been measured to the Fréchet mean of each class. The covariance is assigned to the class to which it has closer distance.



(a)

(b)



(c)

Figure 4.12: Classification based on the distance to the Fréchet mean of each class when weight has been applied.

Fréchet means	Accuracy of classification for Class 1 (%)	Accuracy of classification for Class 2 (%)
Metric d_{R_3}	0.88	0.92
Metric d_F	0.63	0.77
Metric d_{R_1}	0.82	0.85
Metric d_{R_2}	0.83	0.86
Metric d_F (Weighted)	0.81	0.82
Metric d_{R_1} (Weighted)	0.85	0.86
Metric d_{R_2} (Weighted)	0.86	0.88
Classification using Eq.(4.12)	0.68	0.67

Table 4.9: Probability of correct classification within two classes $\mathcal{C}_1$ and $\mathcal{C}_2$. The classifiers have been compared with the result of Eq.(4.12).

Fréchet means	Accuracy of classification for Class 1 (%)	Accuracy of classification for Class 2 (%)	Accuracy of classification for Class 3 (%)
Metric d_{R3}	0.92	0.86	0.95
Metric d_F	0.83	0.39	0.62
Metric d_{R1}	0.91	0.51	0.78
Metric d_{R2}	0.92	0.51	0.82
Metric d_F (Weighted)	0.87	0.63	0.90
Metric d_{R1} (Weighted)	0.91	0.78	0.94
Metric d_{R2} (Weighted)	0.95	0.80	0.98
Classification using Eq.(4.12)	0.80	0.60	0.72

Table 4.10: Probability of correct classification within three classes $\mathcal{C}_1, \mathcal{C}_2$ and $\mathcal{C}_3$ in comparison to the resulting classifier using Eq.(4.12).

In detection and classification process we can improve the performance of a classifier in distinguishing between the features with similar properties resulting from same class by keeping them as close as possible and same time keep the dissimilar features as far as possible using the a priori knowledge of the data during the training step. This process can be performed using the concept of weighted distances [5]. This approach has also been

used in beam forming [36].

To keep the similar training convenience matrices (features) close together and at the same time keep those features which have been obtained from different classes as far as possible we can maximise the follwing optimisation problem with respect to the metric under inspection [5]:

$$\arg\max_{\mathbf{W}} \frac{\sum d^2_{\mathbf{W}}\left(\mathbf{S}_{ik}, \mathbf{S}_{jk}\right)}{\sum d^2_{\mathbf{W}}\left(\mathbf{S}_{ik}, \mathbf{S}_{jk'}\right)} \quad \mathbf{W} > \mathbf{0} \tag{4.13}$$

where in Eq.(4.13) $\mathbf{W} = \mathbf{\Omega}\mathbf{\Omega}^H$ is positive semi definite Hermitian matrix. It is known as weighting factor where $\mathbf{\Omega}$ is a matrix of size $M \times K$, $K \leq M$; The summation in nominator of Eq.(4.13) is over all covariance matrices in similar classes , On the other hand, the denominator in same equation is summation over the all possibilities of covariance matrices in the dissimilar classes.

The metric d_{R3} is weight invariant; in other words for any weighting factor $\mathbf{W}$ and positive definite matrices $\mathbf{A}$ and $\mathbf{B}$ we have [1] :

$$d_{R3}\left(\mathbf{\Omega}^H\mathbf{A}\mathbf{\Omega}, \mathbf{\Omega}^H\mathbf{B}\mathbf{\Omega}\right) = d_{R3}\left(\mathbf{A}, \mathbf{B}\right) \tag{4.14}$$

On the other hand metrics d_{R1},d_{R2} and d_F are not weight invariant [5]. This means that the length of the geodesic passing trough the covariance matrices $\mathbf{A}$, $\mathbf{B}$ in $\mathcal{M}$ may no longer be the same as the geodesic passing from the points $\mathbf{\Omega}^H\mathbf{A}\mathbf{\Omega}$ and $\mathbf{\Omega}^H\mathbf{B}\mathbf{\Omega}$. Such property allows us to use the prior knowledge of our training set and as a result increase the probability of correct classification. In Tables 4.9 and 4.10 the effect of weight in finding the Fréchet mean has been illustrated as well; it can be seen that by finding the appropriate weight with respect to the metrics d_{R1}, d_{R2} and d_F the result of classification has been

substantially improved.

4.3 Multi class Classification of HCI data set using Fréchet mean

In this section we perform the method of distance to the center of mass in classification of high content cell imaging (HCI) data set. To our knowledge this approach has not been applied in HCI data classification. The main purpose of this part is to demonstrate that the techniques of Fréchet means can be considered for further study in field of drug analysis. As far as the comparison of our results is concerned the aim of our work in here is to evaluate the performance of different Riemannian distances and their corresponding Fréchet means when applied on the classification of real data set. We cannot say how good is the performance of our approach to the other methods since for the time being there is no available result from other researches in this field that can be comparable to ours. However, we believe that the developed methods in this research can be considered as a first step towards analyzing the drug mechanism.

The data set that we have considered for this reason has been provided from Professor Andrew's lab under their permission at the Sunny Brook hospital, Toronto,Ontario.

4.3.1 Preprocessing and source selection of HCI data set

Human breast cancer cells (MCF7) are widely used for studies of tumor biology and drug mechanism action [13]. In this section we briefly explain the process of data preparation.

In order to set up experimental design MCF7 are plated in to three clear-bottom 384

well plates at the density of 3000 cells per well in 30 μl serum free Alpha Minimum Essential Medium. The plate is designed to include different treatments together with 32 wells per treatment. A 96 well source plate is prepared containing the drug solutions with the desired doses. At the same time the staining solutions are prepared for three mutachrome dyes. Next 10 μl from each of mutachrome solution is dispensed to each plate at time (1 plate per dye) this process is followed by adding 10 μl of drug solutions to each plate from prepared 96 well source plate. After 24 hour incubation 10 μl of The Thermo Scientific Fluorescent Probe DRAQ5 is used. The resulting plates are incubated 30 minutes prior to the imaging process. During the imaging process each plate (out of three plates) is imaged using Opera High Content imaging System (Perkin Elmar) which is an automated spinning disc microscopy system. Multiple channel images are taken in approximately 2 hours per plate using 20X magnification water lens. Figure 4.13 shows the device which is used for taking the image of the cells.

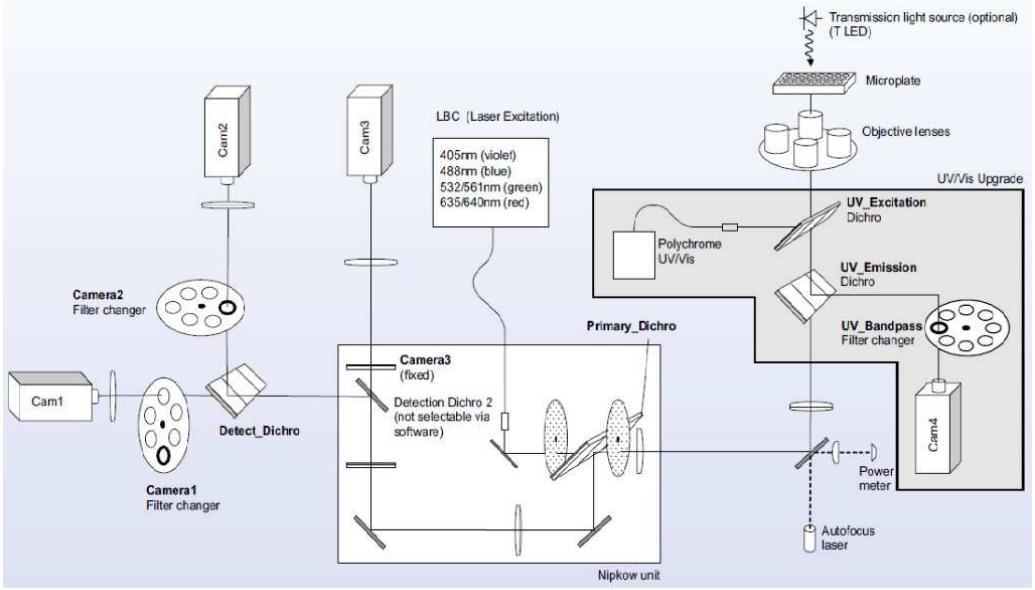


Figure 4.13: Perkin elmar High Content Imaging imaging system diagram

Once the image of the wells was captured, it can be used to extract information. For this purpose the software called CAFE (Classification and Feature Extraction of micrographs of cells) is used. The software automatically segments the image and detect the location of cells. The quality of segmentation is inspected by expert via looking at the images of few wells before running the program for entire 384 wells. It extracts 705 attributes [2] per cell and store them for further use. These attributes correspond to three channels Draq5 together with 2 channels of mutachrome dyes.

[2] Attribute refers to the number of representatives of each cell which in this data set is 705.

4.3.2 Classification result on HCI data set

The data set that we have for classification consists of 11 labels corresponding to 11 types of treatment. As mentioned earlier, depending on type of treatment, we have 32 wells per treatment. Since some of the attributes are highly correlated and redundant they can be removed from the data set [37]. To remove the correlated attributes there are several ways of approaching attribute selection namely: Principal Component Analysis (PCA), linear discriminant analysis (LDA) or removing the correlated attributes based on the mutual corrolation [38]. We perform the latter approach on the data set; the number of attributes that has been selected is ten. To our knowledge, there is not unique optimum solution based on the different methods of attribute selection to pick optimum number of attributes.

We suppose that we have $\{k\}_{k=1}^{n}$ classes, where n represents the type of treatments which is eleven according to the data set. We form the sample covariance matrices of set of cells within each class and denote them as $\{\mathbf{S}_{ik}\}_{k=1}^{n}$; where i and k represent the number of sample covariance matrices within each class and class label respectively. For each class k the geometric mean of the covariance matrices, $\left\{\hat{\mathbf{\Sigma}}_k\right\}_{k=1}^{n}$, can be obtained with respect to the population covariance matrices $\{\mathbf{S}_{ik}\}_{k=1}^{n}$ and four metrics d_{R1},d_{R2},d_{R3} and d_F using the methods discussed in Chapter 3. As far as the number of covariance matrices for training step is concerned, 50 covariance matrices of size 10×10 are considered. Each covariance matrix is formed from observation of 20 cells which is selected from training set of each class. In order to evaluate the accuracy of the classifier we train the classifier 200 times and test it against the covariance matrix of the observed cells from each class and evaluate the probability of correct classification over number of training times.

The types of consumed drug for the experiment corresponding to each data file is given

in Table 4.11.

Treatment	Dose
Untreated	null
DMSO	2.5%
Ethanol	6 μl
BFA	10 μg/ml
Rapamycin	25μ M
Tamoxifen	30 μM
Thapsigargin	40 nM
Tunicamycin	25μM
TNFalpha	10ng/ml
Starvation24	24 hours
Starvation72	72 houres

Table 4.11: The type and amount of used medications

The result of classification has been illustrated in Figure 4.14. As far as the performance of each metric in Fréchet mean based classification using Algorithm 3 is concerned, metric d_F has the lowest accuracy in comparison to the Riemannian metrics d_{R1}, d_{R2} and d_{R3}. Meanwhile, the classifier based on metric d_{R3}, even though it is weight invariant, provides highest probability of detecting the correct class across the types of the treatment with exception in class TN (Tunicamycin); this metric perform very close to the metrics d_{R1}, d_{R2} and d_F in identifying the cells which have been received Rapamycin (RAP). As expected, the classifiers based on metrics d_{R1} and d_{R2} have very close performance; it is merely due to their similar performance in obtaining the Fréchet means of population of positive definite matrices.

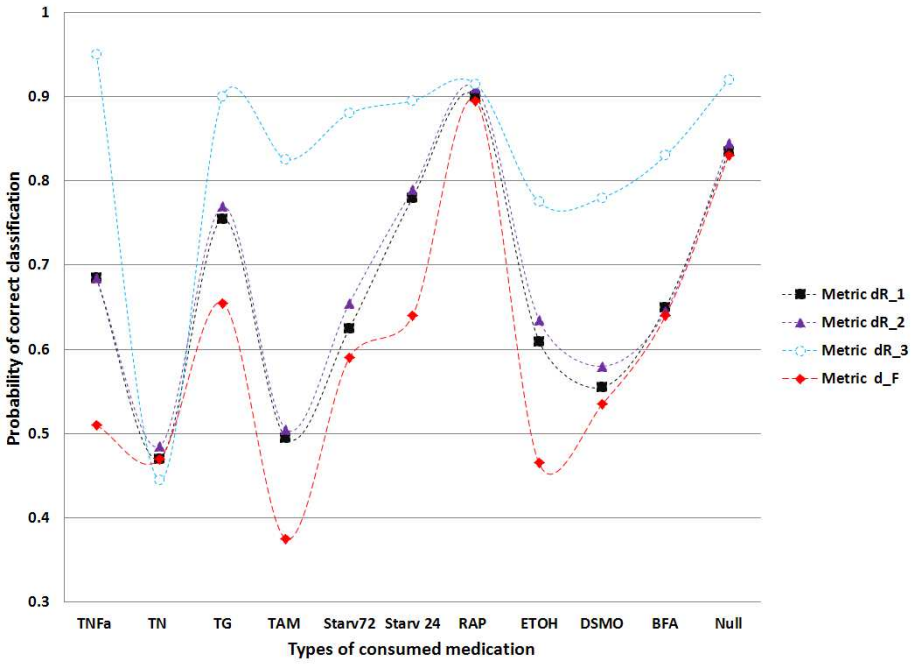


Figure 4.14: Class label versus probability of correct classification

As mentioned earlier in this chapter, the effect of the weight is to keep the covariance matrices within the same classes close to each other and at the same time keep the training covariance matrices belonging to the different classes as far as possible, according to Eq.(4.13). In order to improve the accuracy of the classifier based on the distance to the center of mass, we apply the weight with respect to each metric on the covariance matrices prior to finding the Fréchet mean. The result of this approach has been depicted in Figure 4.15. It can be seen that by using the weight the classifier performs better in terms of probability of correct classification across the metrics d_{R1},d_{R2} and d_F. Meanwhile, in classes "starvation for 24 hours" (starv 24),"starvation for 72 hours" (starv 72) , "Ethanol" (ETOH) and "Thapsigargin" (TG) the algorithm of distance to the center of mass based on metric d_{R3} performs slightly better than the metrics d_{R1} and d_{R2}. In overall, by applying

the weight, metrics d_{R1} and d_{R2} demonstrate higher accuracy in comparison to metrics d_F. The results of the classifier in either case of weighted and unweighted is compatible with the results we obtained using the simulated data in Tables 4.9 and 4.10.

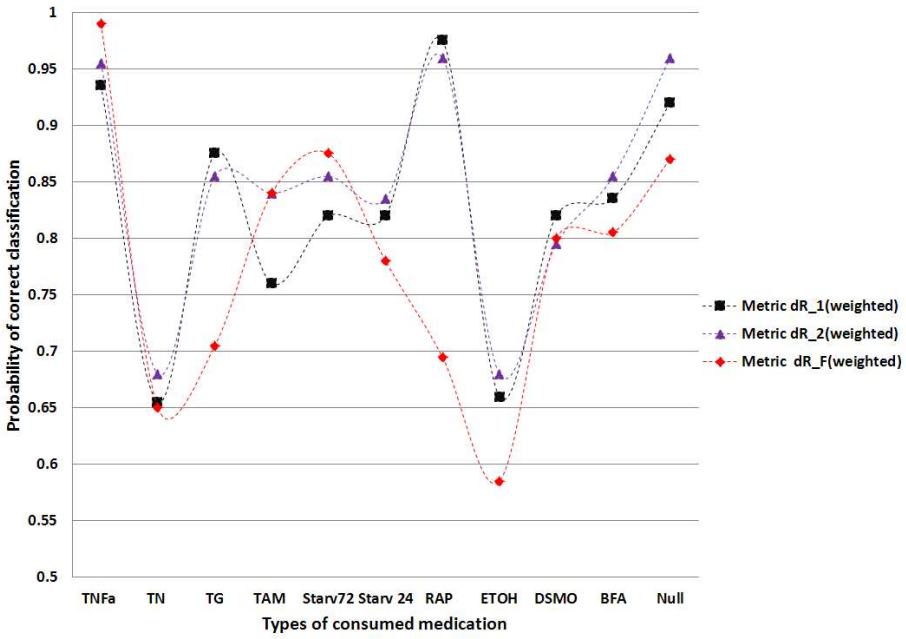


Figure 4.15: Class label versus probability of correct classification when weighting factor has been used.

Chapter 5

Conclusion and further study

5.1 Summary of research

In this thesis, our main focus is on the concept of finding mean of positive definite Hermitian matrices from mathematical point of view. It has been shown that the space of positive definite Hermitian matrices is not Euclidean with zero curvature rather than it has curvature [39]. In order to study the mean of such features we need to consider the ambient space as the manifold $\mathcal{M}$. In this thesis the notion of Fréchet mean was introduced in Chapter 3 as the basic tool of finding the mean of positive definite Hermitian matrices. Since this method depends on the type of metric, the key advantages of it was that we would be able to utilize the Riemannian distances. As a result of that we could obtain not only the arithmetic mean but also obtained the mean on the Reimannian manifold.

As far as finding the mean of positive definite Hermitian matrices are concerned, it was shown that one can unify the process of finding the mean under the concept of Fréchet

mean. Here, three Riemannian metrics were considered for derivation of Fréchet mean. For metric d_{R2} we obtained a closed form solution for the resulting Fréchet mean. For the two remaining distances it was shown that we need to resort to numerical methods to obtain the mean.

In Chapter 4 the performance of each estimator was evaluated. We considered several models based on the Cholesky factor of covariance matrix, square root of the covariance matrix and logarithm of the Cholesky factor. The performance of each Fréchet mean estimator was assessed, depending on the choice of the model, using three methods namely root mean square error of Frobinious norm, root mean square error of metric square root of matrix and the average of the loss function. We evaluate the closeness of each estimator to the nominal covariance matrix. The results showed that as far as the accuracy of each estimator is concerned, when the Fréchet mean is evaluated according to the models based on the Cholesky or square root of covariance matrix the accuracy of the resulting Fréchet mean by using Riemannian metrics d_{R1} and d_{R2} are remarkable in comparison to the Fréchet mean of metrics d_{R3} and d_F. On the other hand we have seen that when the estimators are applied to the logarithm of the Cholesky factor, Fréchet mean of metric d_{R3} demonstrate slightly better performance in terms of error among other estimators. In overall, we observed that the Riemannian metrics have better performance in finding the mean of Hermitian positive definite matrices with respect to the models.

In terms of application we performed the concept of Fréchet mean in classification task. Our main attention on this part was to evaluate this method according to different choice of Fréchet mean. Furthermore, we showed that if we use the a priori knowledge of the training features (covariance matrices) the result of classifying the unknown observation

can be improved in terms of probability missclassification. As far as the real data set is concerned, we applied the method of distance to the center of mass in classifying the high content cell imaging data set; our approach for classification of the cells according to their response to the different types of medication can be considered as a first step towards the analysis of the drug mechanism and drug interaction in this way.

5.2 Future work

There are open problems that can be addressed during the further research on finding the mean of covariance matrices: Regarding to the estimation of geometric mean of covariance matrices one can consider other models of formining the population of covariance matrices and evaluate the performance of the estimators.

It was mentioned that for the set of pairwise commutative positive definite matrices one can obtain a closed form solution for the Fréchet mean with respect to metric d_{R3}; this alternative definition for metric d_{R3} can be considered for further research. Applications of the Fréchet mean in other areas such as signal detection and signal estimation can also be considered in future.

Appendices

Appendix A

Directional derivative on manifold $\mathcal{M}$

When properly formulated, many of signal optimization problem involving matrix $\mathbf{P}$ which minimizes the functional $F(\mathbf{P})$. For an $M \times M$ matrix $\mathbf{P} \in \mathcal{M}$, a variation in $\mathbf{P}$ is also an $M \times M$ matrix in $\mathcal{M}$ and it is not possible to uniquely describe the corresponding variation in F at point $\mathbf{P}$. What can be done is to describe the variation in F with respect to variation in $\mathbf{P}$ along particular direction on the manifold. Let $\mathbf{V}$ be an element on the tangent space $\mathcal{H}$ of $\mathcal{M}$ with $\|\mathbf{V}\| = 1$. Let $D_{\mathbf{V}}$ denotes the differential operator along the direction of $\mathbf{V}$. Then for the given $\epsilon > 0$, and real continuous functional F, the directional derivative of $F(\mathbf{P})$ at $\mathbf{P}$ in the direction of $\mathbf{V}$ is defined as :

$$D_{\mathbf{V}}F(\mathbf{P}) = \lim_{\epsilon \longrightarrow 0} \frac{F(\mathbf{P} + \epsilon\mathbf{V}) - F(\mathbf{P})}{\epsilon \|\mathbf{V}\|} \tag{A.1}$$

where, since F is a functional of matrices in the manifold.

Thus $D_{\mathbf{V}}F(\mathbf{P})$ is a real quantity and both $\mathbf{P}$ and $\mathbf{V}$ can be represented as a linear

combinations in form of Eq.(2.23) such that

$$\mathbf{P}=\sum_{m=1}^{M}\sum_{n=1}^{M}p_{mn}\tilde{\mathbf{E}}_{mn};\quad \mathbf{V}=\sum_{m=1}^{M}\sum_{n=1}^{M}v_{mn}\tilde{\mathbf{E}}_{mn};\quad with\ \ \tilde{\mathbf{E}}_{mn}=\tilde{\mathbf{E}}_{mn}^{H} \tag{A.2}$$

where p_{mn} and v_{mn} are real coefficients. If F has a total differential ΔF at $\mathbf{P}$, then we have:

$$df=\lim_{\epsilon\longrightarrow 0}\Delta f=\lim_{\epsilon\longrightarrow 0}\sum_{m,n}\frac{\partial F}{\partial p_{mn}}.\epsilon v_{m,n} \tag{A.3}$$

as a consequence we have

$$D_{\mathbf{V}}F\left(\mathbf{P}\right)=\sum_{m,n}\frac{\partial F}{\partial p_{mn}}.\frac{v_{mn}}{\|\mathbf{V}\|}. \tag{A.4}$$

Now, if we define the $M\times M$ gradient operator as $\nabla\triangleq\left[\frac{\partial}{\partial p_{mn}}\right]$,$1\leq m,n\leq M$,then, the result of Eq.(A.4) is the sum of the elements of the Hadamard(element by element) product of two $M\times M$ matrices such that

$$\sum_{m,n}\left[\nabla F\bigodot\mathbf{V}\right]_{mn}=\mathrm{Tr}\left[\nabla F.\mathbf{V}\right]\triangleq\left\langle\nabla F,\mathbf{V}\right\rangle. \tag{A.5}$$

Hence the directional derivative $D_{\mathbf{V}}F\left(\mathbf{P}\right)$ can be written as

$$D_{\mathbf{V}}F\left(\mathbf{P}\right)=\left\langle\nabla F,\mathbf{V}\right\rangle \tag{A.6}$$

Appendix B

Proof of Eq.(4.12)

In chapter three we used Eq.(4.12) as another classifier in order to compare its performance with the other methods of classification . In here the proof of it has been provided.

Theorem B.0.1. *If $\mathbf{X}_1, ..., \mathbf{X}_N$ are independently identically distributed from p-variate normal distribution with $\mathbf{0}$ mean and covariance matrix $\mathbf{\Sigma}$, then $N\bar{\mathbf{X}}\mathbf{\Sigma}^{-1}\bar{\mathbf{X}}^T$ has chi-square distribution with p degrees of freedom.*

Proof. Let $\mathbf{C}$ be a nonsingular matrix such that $\mathbf{C\Sigma C}^T = \mathbf{I}$. Define the random variable $\mathbf{Z} = \frac{\mathbf{C}}{\sqrt{N}}\bar{\mathbf{X}}^T$. Then $\mathbf{Z}$ is normally distributed with mean $\mathbb{E}(\mathbf{Z}) = \frac{\mathbf{C}}{\sqrt{N}}\mathbb{E}\left(\bar{\mathbf{X}}^T\right) = \mathbf{0}$ and covariance matrix $\mathbb{E}\left(\mathbf{ZZ}^T\right) = \mathbb{E}\left(\frac{\mathbf{C}}{\sqrt{N}}\bar{\mathbf{X}}^T\bar{\mathbf{X}}\mathbf{C}^T\right) = \frac{\mathbf{C}}{\sqrt{N}}\mathbb{E}\left(\bar{\mathbf{X}}^T\bar{\mathbf{X}}\right)\frac{\mathbf{C}^T}{\sqrt{N}} = N^{-1}\frac{\mathbf{C}}{\sqrt{N}}\mathbf{\Sigma}\frac{\mathbf{C}^T}{\sqrt{N}} = \mathbf{I}$. Then $N\bar{\mathbf{X}}\mathbf{\Sigma}^{-1}\bar{\mathbf{X}}^T = N\mathbf{Z}^T\left(\frac{\mathbf{C}^T}{\sqrt{N}}\right)^{-1}\mathbf{\Sigma}^{-1}\left(\frac{\mathbf{C}}{\sqrt{N}}\right)^{-1}\mathbf{Z} = \mathbf{Z}^T\left(N\frac{\mathbf{C}}{\sqrt{N}}\mathbf{\Sigma}\frac{\mathbf{C}^T}{\sqrt{N}}\right)^{-1}\mathbf{Z} = \mathbf{Z}^T\mathbf{Z}$. Which $\mathbf{Z}^T\mathbf{Z}$ is the sum of squares of the components of $\mathbf{Z}$. Thus, $N\bar{\mathbf{X}}\mathbf{\Sigma}^{-1}\bar{\mathbf{X}}^T$ is distributed as $\mathbf{Z}^T\mathbf{Z} = \sum_{i=1}^{p} Z_i^2$, where $Z_1, ..., Z_p$ are independently normally distributed with mean 0 and variance 1. Since, Z_i^2 for $i = 1, ..., p$ have chi-square distribution with 1 degree of freedom, $\sum_{i=1}^{p} Z_i^2$ has chi-square distribution with p degree of freedom.

Appendix C

Geodesic with respect to Log Reimannian metric d_{R3}

In chapter two we used the result of the theorem which states for pair of positive definite Hermitian matrices $\mathbf{A}$ and $\mathbf{B}$ of size $M \times M$ in manifold $\mathcal{M}$, such that they commute with each other, the exponential function maps the line segment $[\log \mathbf{A}, \log \mathbf{B}]$ in tangent space $T_{\mathcal{M}}$ to the geodesic $[\mathbf{A}, \mathbf{B}]$ in $\mathcal{M}$[1].

We need to show that the path:

$$\gamma(t) = \exp\left((1-t)\log \mathbf{A} + t \log \mathbf{B}\right); \quad 0 \leq t \leq 1 \tag{C.7}$$

[1]Let $\mathbf{A}$ and $\mathbf{B}$ be two points on the manifold $\mathcal{M}$. we denote the geodesic passing through the points with $[\mathbf{A}, \mathbf{B}]$. On the other hand, if $\mathbf{C}$ and $\mathbf{D}$ are two points on $T_{\mathcal{M}}$ or any space with zero curvature, we concern about the line segment connection them. However, as long as the ambient space is clear for us, we still use the similar notation to show this line segment and represent it by $[\mathbf{C}, \mathbf{D}]$.

is the unique path with shortest length (geodesic) in manifold $\mathcal{M}$ with respect to the inner-product

$$\langle \mathbf{A}, \mathbf{B} \rangle_{\mathbf{X}} = \mathrm{Tr}(\mathbf{A}\mathbf{X}^{-1}\mathbf{B}\mathbf{X}^{-1}) \tag{C.8}$$

Since $\mathbf{A}$ and $\mathbf{B}$ commute we can write $\gamma(t) = \mathbf{A}^{1-t}\mathbf{B}^{t}$. We then have

$$\gamma^{'}(t) = (\log \mathbf{B} - \log \mathbf{A})\, \gamma(t) \tag{C.9}$$

but the length of a path $\gamma : [0, 1] \longrightarrow \mathcal{M}$ is given by [7] :

$$\mathcal{L}(\gamma) := \int_0^1 \sqrt{\langle \gamma'(t), \gamma'(t) \rangle_{\gamma(t)}} dt = \int_0^1 \sqrt{\mathrm{Tr}\left(\gamma(t)^{-1}\gamma'(t)\right)^2} dt \tag{C.10}$$

We note that by definition the geodesic is defined by taking the infimum over all possible paths γ connecting $\mathbf{A}$ and $\mathbf{B}$.

According to the Eq.(C.10) and Eq.(C.9) we have :

$$\mathcal{L}(\gamma) = \int_0^1 \| \log \mathbf{A} - \log \mathbf{B} \|_2 dt = \| \log \mathbf{A} - \log \mathbf{B} \|_2 \tag{C.11}$$

On the other hand "the exponential metric increasing property" (EMI) [31] states that there is no path shorter than Eq.(C.11). To show that this path is unique we suppose that $\tilde{\gamma}$ be another path connecting $\mathbf{A}$ and $\mathbf{B}$ in $\mathcal{M}$. Then $\tilde{H}(t) = \log \tilde{\gamma}(t)$ is the path that connects $\log \mathbf{A}$ and $\log \mathbf{B}$ in $T_{\mathcal{M}}$. From [18] this path has the length $\| \log \mathbf{A} - \log \mathbf{B} \|_2$. But in Euclidean space the straight line is the unique shortest path connecting two points. As a result $\tilde{\mathbf{H}}(t)$ is another reparametrization of the line segment which connects $\log \mathbf{A}$ and $\log \mathbf{B}$ thus $\gamma(t)$ must be equal to $\tilde{\gamma}(t)$.

□

In part of the derivation of the Fréchet mean with respect to metric d_{R3}, we used the following expression:

$$\mathbf{X}^{-1}\log\left(\mathbf{A}\right)\mathbf{X} = \log\left(\mathbf{X}^{-1}\mathbf{A}\mathbf{X}\right) \tag{C.12}$$

Eq.(C.12) is valid for any invertable matrix $\mathbf{X}$. Matrix $\mathbf{A}$ is assumed to have strictly positive spectrum (i.e: $\mathbf{A}$ must be positive definite matrix). To show this we need to use the following theorem from [30].

Theorem C.0.2. *If* $\mathbf{A}$ *and* $\mathbf{B}$ *are two* $M \times M$ *matrices such that* $\mathbf{B}$ *is non-singular, then we have:*

$$\exp\left(\mathbf{B}\mathbf{A}\mathbf{B}^{-1}\right) = \mathbf{B}\exp\left(\mathbf{A}\right)\mathbf{B}^{-1} \tag{C.13}$$

According to the theorem (C.0.2) one can write:

$$\exp\left(\mathbf{B}\log\left(\mathbf{A}\right)\mathbf{B}^{-1}\right) = \mathbf{B}\mathbf{A}\mathbf{B}^{-1} \tag{C.14}$$

which immediately gives the desired result

$$\log\left(\mathbf{B}\mathbf{A}\mathbf{B}^{-1}\right) = \mathbf{B}\log\left(\mathbf{A}\right)\mathbf{B}^{-1} \tag{C.15}$$

Bibliography

[1] F. Barbaresco, "Fréchet metric space, homogeneous siegel domains & radar matrix signal processing," *Matrix Information Geometry*, p. 5.

[2] B. Jeuris, R. Vandebril, and B. Vandereycken, "A survey and comparison of contemporary algorithms for computing the matrix geometric mean," *Electronic Transactions on Numerical Analysis*, vol. 39, pp. 379–402, 2012.

[3] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass brain–computer interface classification by Riemannian geometry," *Biomedical Engineering, IEEE Transactions on*, vol. 59, no. 4, pp. 920–928, 2012.

[4] D. C. Alexander, "Multiple-fiber reconstruction algorithms for diffusion MRI," *Annals of the New York Academy of Sciences*, vol. 1064, no. 1, pp. 113–133, 2005.

[5] Y. Li and K. M. Wong, "Riemannian distances for EEG signal classification by power spectral density," *IEEE journal of selected selected topics in signal processing*, 2013.

[6] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on Riemannian manifolds," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 10, pp. 1713–1727, 2008.

[7] M. Moakher, "A differential geometric approach to the geometric mean of symmetric positive-definite matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 26, no. 3, pp. 735–747, 2005.

[8] F. Barbaresco, "Innovative tools for radar signal processing based on Cartans geometry of SPD matrices & information geometry," *Radar Conference, 2008. RADAR'08. IEEE*, pp. 1–6, 2008.

[9] H. Karcher, "Riemannian center of mass and mollifier smoothing," *Communications on pure and applied mathematics*, vol. 30, no. 5, pp. 509–541, 1977.

[10] R. Bhatia and J. Holbrook, "Riemannian geometry and matrix geometric means," *Linear algebra and its applications*, vol. 413, no. 2, pp. 594–618, 2006.

[11] J. Lawson and Y. Lim, "Weighted means and Karcher equations of positive operators," *Proceedings of the National Academy of Sciences*, vol. 110, no. 39, pp. 15 626–15 632, 2013.

[12] R. J. Muirhead, *Aspects of multivariate statistical theory*. John Wiley & Sons, 2009, vol. 197.

[13] C. K. Osborne, K. Hobbs, and J. M. Trent, "Biological differences among mcf-7 human breast cancer cell lines from different laboratories," *Breast cancer research and treatment*, vol. 9, no. 2, pp. 111–121, 1987.

[14] R. A. Monzingo and T. W. Miller, *Introduction to adaptive arrays*. SciTech Publishing, 1980.

[15] J. R. Magnus and H. Neudecker, *Matrix differential calculus with applications in statistics and econometrics*. Wiley, 1988.

[16] L. Lin, N. J. Higham, and J. Pan, "Covariance structure regularization via entropy loss function," *Computational Statistics & Data Analysis*, vol. 72, pp. 315–327, 2014.

[17] T. W. Anderson, *An Introduction To Multivariate Statistical Analysis*. Wiley Eastern Private Limited; New Delhi, 1954.

[18] R. Bhatia, *Positive definite matrices*. Princeton University Press, 2009.

[19] G. Casella and R. L. Berger, *Statistical inference*. Duxbury Press Belmont, CA, 1990, vol. 70.

[20] M. Congedo, "EEG source analysis," http://tel.archives-ouvertes.fr/docs/00/91/58/24/PDF/Congedo_Marco_HDR_V2.pdf.

[21] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate analysis*. Academic Press, 1979.

[22] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[23] F. Crosilla and A. Beinat, "Use of generalised procrustes analysis for the photogrammetric block adjustment by independent models," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 56, no. 3, pp. 195–209, 2002.

[24] F. Hiai and D. Petz, "Riemannian metrics on positive definite matrices related to means," *Linear Algebra and its Applications*, vol. 430, no. 11, pp. 3105–3130, 2009.

[25] M. P. Do Carmo, *Riemannian geometry*. Springer, 1992.

[26] C. Niculescu and L.-E. Persson, *Convex functions and their applications: A contemporary approach*. Springer, 2006, vol. 23.

[27] N. J. Higham, *Functions of matrices: theory and computation*. SIAM, 2008.

[28] W. Rudin, *Principles of mathematical analysis*. McGraw-Hill New York, 1964, vol. 3.

[29] D. A. Bini, B. Iannazzo, B. Jeuris, and R. Vandebril, "Geometric means of structured matrices," *BIT Numerical Mathematics*, pp. 1–29, 2012.

[30] M. L. Curtis and M. Curtis, *Matrix groups*. Springer, 1984, vol. 2.

[31] R. Bhatia, "On the exponential metric increasing property," *Linear Algebra and its applications*, vol. 375, pp. 211–220, 2003.

[32] D. J. MacKay, "Introduction to Monte Carlo methods," in *Learning in graphical models*. Springer, 1998, pp. 175–204.

[33] D. Pigoli, J. A. Aston, I. L. Dryden, and P. Secchi, "Distances and inference for covariance operators," *Biometrika*, 2014.

[34] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Riemannian geometry applied to BCI classification," in *Latent Variable Analysis and Signal Separation*. Springer, 2010, pp. 629–636.

[35] R. A. Johnson and D. W. Wichern, *Applied multivariate statistical analysis*. Prentice Hall Upper Saddle River, NJ, 2002, vol. 5, no. 8.

[36] D. Ciochina, M. Pesavento, and K. M. Wong, "Worst case robust downlink beamforming on the Riemannian manifold," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3801–3805.

[37] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[38] M. Haindl, P. Somol, D. Ververidis, and C. Kotropoulos, "Feature selection based on mutual correlation," in *Progress in Pattern Recognition, Image Analysis and Applications*. Springer, 2006, pp. 569–577.

[39] S. Kobayashi and K. Nomizu, *Foundations of differential geometry*. New York, 1969, vol. 2.