Multichannel EEG Signal Classification on a Riemannian Manifold

MULTICHANNEL EEG SIGNAL CLASSIFICATION ON A RIEMANNIAN MANIFOLD

By

YILI LI, B.SC., M.SC.

JANUARY 2010

A Thesis

Submitted to the Department of Electrical & Computer Engineering

and the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree of

Doctor of Philosophy

McMaster University

DOCTOR OF PHILOSOPHY (2010) McMaster University

(Electrical & Computer Engineering) Hamilton, Ontario

TITLE:	Multichannel EEG Signal Classification on a Riemannian Manifold
AUTHOR:	Yili Li M.Sc. (Applied Mathematics, University of Waterloo, Canada)
SUPERVISOR:	Kon Max Wong, Canada Research Chair Professor of Signal Processing
CO-SUPERVISOR:	Hubert deBruin, Associate Professor
NUMBER OF PAGES:	1, 192

Abstract

The study of the different sleep stages of a patient using his/her recorded EEG signals falls in the area of signal classification. In general, this involves extracting from the EEG signals, a signal feature on which the classification is performed. In this thesis, we apply the techniques of signal classification to the analysis of the sleep of a patient. The feature we use is the power spectral density (PSD) matrices of a multi-channel EEG signal. This not only allows us to examine the power spectrum contents of each signal which complies with what clinical experts use in their visual judgement of EEG signals, but also allows the correlation between the multi-channel signals to be studied. To establish a metric facilitating the classification, we analyze the structure as well as exploit the specific geometric properties of the space of PSD matrices. Specifically, we study this space from the viewpoint of Riemannian manifolds. We apply a Riemannian metric and, with the aid of fibre bundle theory, develop intrinsic (geodesic) distance measures for the PSD matrix manifold. To tilize such new distance measures effectively for EEG signal classification, we need to find a suitable weighting matrix for the PSD matrices so that the distances between similar features are minimized while those between dissimilar features are maximized. A closed form of this weighting matrix is obtained by solving an equivalent convex optimization problem. The effectiveness of using these novel weighted distance measures is verified by applying them to the sleep pattern classification of a collection of recorded EEG

signals using the k-nearest neighbor decision algorithm with excellent results.

List of Acronyms

AR	Autoregression
EEG	Electroencephalogram
FFT	Fast Fourier Transform
IID	Independent and Identically Distributed
PCA	Principal Component Analysis
PSD	Positive Semi Definite
SVD	Singular Value Decomposition

List of Notations

Boldface lowercase letters are used to denote column vectors.

Boldface uppercase letters are used to denote matrices.

$(\cdot)^*$	the conjugate operator
$(\cdot)^T$	the transpose of a vector or a matrix
$(\cdot)^H$	the Hermitian transpose of a vector or a matrix
$(\cdot)^{-1}$	the inversion of a matrix
$(\hat{\cdot})$	the estimate of a parameter
$(\cdot)_j$	the jth element of a vector
$[\cdot]_{ij}$	the ijth element of a matrix
$\lvert\cdot\rvert$	the magnitude of a complex quantity or the determinant of a matrix
$\lVert\cdot\rVert$	the Euclidean norm of a vector or a matrix
$\otimes$	the tensor product
$\mathbb{E}[\cdot]$	the statistical expectation operator
diag$\{\mathbf{a}\}$	the diagonal matrix constructed from elements of $\mathbf{a}$
Tr$\{\cdot\}$	the trace of a matrix
j	$\sqrt{-1}$

$\text{vec}\{\cdot\}$	the operator stacking the columns of a matrix on top of each other
$\mathcal{M}$	Manifold
$T_x\mathcal{M}$	Tangent space of $\mathcal{M}$ at x
$\mathbb{V}(\mathcal{M})$	Vector space of $\mathcal{M}$
$\mathcal{H}$	Hilbert space
λ	eigenvalue
$\boldsymbol{I}$	Identity matrix
$\mathbb{R}$	the field of real numbers
$\mathbb{C}$	the field of complex numbers
$\mathbf{x} \sim \mathbf{y}$	$\mathbf{x}$ and $\mathbf{y}$ are similar
$\mathbf{x} \nsim \mathbf{y}$	$\mathbf{x}$ and $\mathbf{y}$ are not similar

Contents

Chapter 1

Introduction

An electroencephalogram (EEG) is the measurement of electrical activities produced by the brain. The measurement is carried out by placing electrodes on the scalp recording the electrical potentials generated by synaptic fields in the cerebral cortex. Although such an electrode would pick up the superposition of many different waves emitted from various regions of the brain, rendering it more difficult to interpret the data, EEG is still a valuable measure of the brain's electrical function. EEGs have been employed in many clinical areas such as administration of anaesthetics, detection and prediction of epileptic seizures, recognition of pathological conditions such as concussion, as well as analysis of depression, etc. [23,44,48]. In this thesis, we study an important application of EEG to the determination of the level of sleep of a patient. In particular, we determine the depth of a patient's *natural* (no anaesthetics) sleep by classifying the pattern of the recorded EEG signals.

The dependence of pattern classification on mathematics can be well characterized by the following quoting of the mathematical philosopher A. N. Whitehead: "The notion of the importance of patterns is as old as civilization. Every art is founded

on the study of patterns. Mathematics is the most powerful technique for the understanding of patterns and for the analysis of the relationships of patterns." [85]. However, mathematics being abstract, has no physical constraints. It is therefore a challenging problem to choose the proper mathematical techniques and apply them to the real-life pattern classification problems.

An EEG pattern is an entity indicating a specific state of the brain. EEG classification is the study of how machines can process the EEG signals, learn to distinguish EEG patterns in different brain states, and make reasonable decisions on the classes of the patterns. In this chapter, we first present a brief overview of sleep staging based on the contents of the different types of EEG signals. Then, we review some existing EEG signal classification methods. Finally, we present a preview of our geometric approach to EEG signal classification.

1.1 Sleep Staging by EEG Signals

The study of sleep is highly important in health care since sleep disorders affect the well-being and productivity of many individuals. However, the sleep of a person is not a homogenous state from the beginning to the end. Analysis of a patient's sleep history requires putting the patient in a sleep laboratory to acquire up to 8 hours of polysomnographic recordings which not only consists of recordings of EEG, but often also includes ocularogram (EOG) as well as other physiological data such as the activity of selected muscles, the electrocardiogram, oxygen concentration in arterial blood SaO2, and breathing rate. In this thesis, our attention is focused on the determination of the patient's sleep stages using only the recorded EEG signals.

1.1.1 Measurement of EEG signals

The recording of EEG is carried out by attaching electrodes to the scalp. A typical international 10-20 electrode placement system is shown in Fig. 1.1. The hemispheric

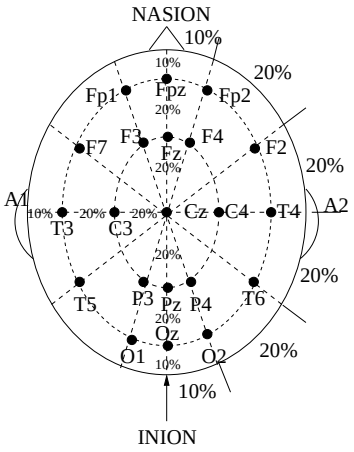


Figure 1.1: A typical international 10-20 electrode placement system: A=Ear lobe, C=Central, P=Parietal, F=Frontal, Fp=Frontal polar, and O=Occipital.

locations of the sensors are indicated by combinations of letters and numbers. The letters Fp, F, C, P, O, T correspond to *Front Polar, Frontal, Central, Parietal, Occipital,* and *Temporal*. Locations on the right and left hemispheres are indicated by even and odd numbers respectively while the letter Z shows electrode placements along the centre line. According to this system, electrodes are placed at 10% and 20% of a semi-circumference measurement on the scalp (see Fig 1.1). Instead of using the entire set of sensors, generally only M of the sensors are used for most EEG studies. In our studies of EEG classification for sleep stage determination, a differential understanding of EEG activity in the different regions of the brain is of no great value. Therefore, our measurements are usually limited to a small number

of sensors. Our decision for the placement of these sensors is based on the experience of other researchers who have carried out EEG signal measurements for sleep stage decision. The manual published by Retschaffen and Kales [74] recommends referential recording for single sensor EEG measurements (usually either C_3 or C_4, referenced to an indifferent electrode placed on the ear lobe or contralateral mastoid (A_1 or A_2). There are a number of advantages in using $(C_3 - A_2)$ or $(C_4 - A_1)$ signals. Retschaffen and Kales state [74]: "On one hand the relatively large interelectrode distance optimises EEG signal amplitudes for sleep analysis, and on the other hand most sleep grapho-elements, sleep staging criteria (vertex sharp waves, K complexes and spindles) are well visualised in these regions. Moreover, high-voltage NREM slow waves seen maximally in frontal regions minimises the contamination of ocular movements in REM sleep on EEG activity. By contrast, the alpha rhythm of relaxed wakefulness is maximal over the occipital poles." Since we are interested in all stages of sleep including relaxed wakefulness, and, in addition, the correlation of the signals at the various positions of the brain is also of importance to our studies, therefore, we choose to have $M = 4$ sensors positioned at: C_3, C_4, O_1, and O_2, each referenced to the earlobe sensor on the opposite side of the skull. Thus, our measurements will all have four channels $(C_3 - A_2)$, $(C_4 - A_1)$, $(O_1 - A_2)$, and $(O_2 - A_1)$ connected to a recording machine which displays the readings – each channel producing a time series.

The EEG signals recorded represent the effects of the superimposition of diverse processes in the brain and are often contaminated by noise and artifacts due to eye blinking or other muscular activities. Even though it is, in general, a difficult task to recognize and eliminate the artifacts in EEG recordings, it is essential to do so for the development of practical automatic sleep staging systems. The aim of artifacts removal should, on the one hand, minimize the amount of data that have

to be eliminated and, on the other hand, ensure that the results obtained are not influenced by undetected artifacts. In this thesis, we follow the usual practice of removing noise and artifacts by suitably filtering of the EEG recordings.

After filtering and the artifacts have been removed, the EEG signal is then divided into *epochs* of 30 seconds each. These are then examined by a trained clinical expert who visually determines the stage of sleep from awake to deep sleep for each of the epochs of the EEG data using the Rechtschaffen and Kales (R&K) [74] scoring system. The expert's decision is then labeled on the corresponding epoch.

1.1.2 EEG signals in sleep analysis

The different stages of the sleep process reflect the different states of the brain which are characterized by the occurrence of different EEG signals. In general, EEG signals occupy the frequency range of $0 - 60\ Hz$ which is usually separated into five constituent physiological subbands, viz., δ $(0 - 4\ Hz)$, θ $(4 - 7\ Hz)$, α $(8 - 12\ Hz)$, β $(13 - 30\ Hz)$, and γ $(30 - 60\ Hz)$ [22]. Typical EEG patterns in these subbands are shown in Figs. 1.2-1.6. Beside the occurrence of these more "stationary" patterns, during a patient's sleep, there are other more transient signal patterns such as the sleep spindles and the K-complexes. A sleep spindle consists of $12 - 16$ Hz waves that occur for $0.5 - 1.5$ seconds. A K-complex consists of a brief negative high-voltage peak followed by a slower positive complex. A K-complex occurs roughly

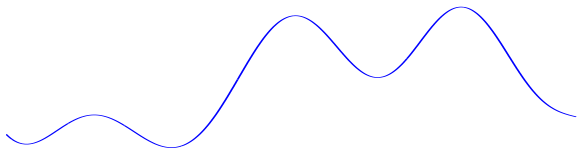


Figure 1.2: δ wave

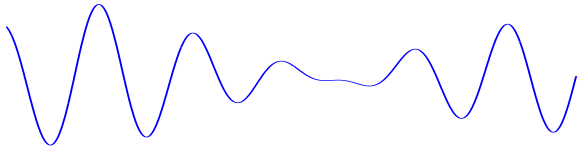


Figure 1.3: θ wave

Figure 1.4: α wave

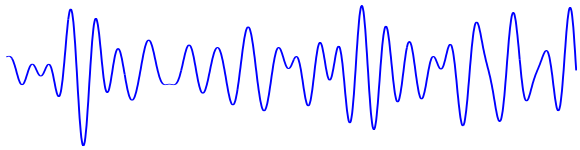


Figure 1.5: β wave

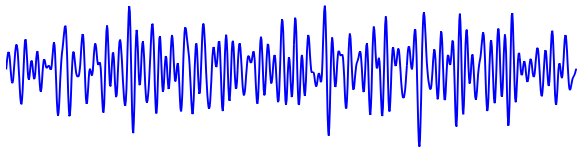


Figure 1.6: γ wave

every $1.0 - 1.7$ minutes and is often followed by bursts of sleep spindles. Figs. 1.7 depicts a K-complex followed by a sleep spindle.

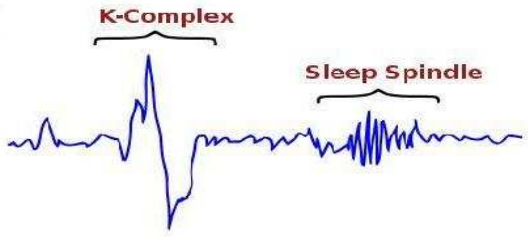


Figure 1.7: K-complex

1.1.3 Classification of sleep stages

In 1953, Kleitman and Aserinsky [6] laid the foundation of sleep classification by observing the existence of two different classes of sleep processes, viz., *slow wave sleep* which is defined by the presence of delta activity having an amplitude of at least $75\mu V$ in the EEG for more than 20% of the time, and *rapid eye movement* (REM) *sleep* which refers to altered ocular motility during sleep. Further insight into the significance of the REM stages prompted a new terminology of sleep stages that emphasized the dichotomy of two distinct neurophysiological states of sleep: *slow sleep* (non-REM sleep) and *fast sleep* (REM sleep). Non-REM sleep can be further subdivided into 4 stages according to the depth of the sleep. Table 1.1 shows the modern day classification of the various stages of sleep together with the associated EEG activities [74].

The transition from stage to stage may be somewhat imprecise. The distinction

Table 1.1: Sleep stages

Sleep stage	Frequency range	Wave patterns
1	$4-8$Hz	α, θ
2	$8-15$Hz	θ, spindles, K-complexes
3	$2-4$Hz	δ, θ
4	$0.5-2$Hz	δ, θ
REM	>12Hz	β, γ
Awake	$8-12$Hz or >12Hz	α, β, γ

between some stages needs quantitative measurements. For example, Stage 3 and Stage 4 have similar rhythms. They can only be distinguished by measuring the occupancy of delta activity.

Although the visual sleep scoring method of Rechtschaffen and Kales (R&K) [74] has been used in clinics, it is sometimes very difficult for every electroencephalographer to note exact measures for EEG phenomena as spikes, sharp waves, or other abnormal patterns. The experienced specialist is able to detect these EEG phenomena only by "eyeballing". This is a laborious and costly classification process, limiting the availability of laboratory sleep analysis in current health care with inter-rater reliability between two expert observers typically around 77% (Cohen's Kappa: 0.68) [4]. Therefore, it is necessary to develop some computer-assisted systems for sleep-staging, which is regarded as EEG signal classification.

To carry out EEG signal classification, filtering is first applied to remove interfering components, and to extract the portion of signals of interest, i.e., features. This is then followed by a signal classifier in which the signals are classified based on *similarity/dissimilarity* measured between the features of the EEG signals. A typical classification scheme is shown in Figure 1.8. Since the R&K manual was published, numerous attempts to design and implement computer-based automatic sleep staging

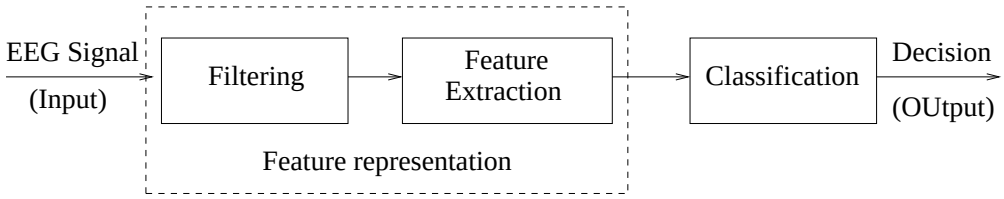


Figure 1.8: A typical classification scheme

have been proposed [4].

It should be noted that sleep staging involves a rather fuzzy process of detection and identification of EEG patterns such that it is not an easy task to perform a computer-assisted analysis since the standards are not well defined. This may be the reason why many algorithms developed in the past decades have not gained wide acceptance in practice. In general, these approaches depend very much on the population for which they were developed in the sense that the performance varies from laboratory to laboratory.

1.2 Existing methods of EEG signal classification

1.2.1 Features and feature extractions of EEG signals

After filtering is applied to remove the artifacts, we begin the extraction of the signal features. EEG signal features can be thought of as the characteristics of the EEG signal suitable for the purpose of classification. The chosen feature could be the power spectral densities (PSD), the auto regression coefficients, time frequency distributions etc.

In the design and implement of automatic sleep staging, various features of EEG

signal have been used. For example, Anderer et al [4] developed an automatic classification system based on one central EEG channel, two EOG channels and one chin EMG channel that adheres to the R&K rules for visual scoring and includes a structured quality control procedure. They achieved an 80% (Cohen's Kappa: 0.72) agreement between automatic and visual epoch staging. Other researchers have studied sleep stages using more analytic approaches such as ratios of power spectral densities measured with data from the different channels [28], inter and intra hemispheric spectral coherence analysis [2], and spectral correlation coefficients between data from the two hemispheres for each of the frequency bands [1]. Spectral power changes in the EEG bands have also been noted during and after apnea/hypopnea events during sleep [89]. Several researchers have reduced the number of channels of EEG in an effort to reduce instrumentation and computational complexity. Berthomier et al [15] validated an automatic sleep scoring system, ASEEGA, which uses spectral properties determined by Fourier analysis or autoregressive (AR) modelling, plus recognition of sleep features such as spindles, of single channel EEG and found agreement of 83% between 5 state classification by 2 experts and ASEEGA. Virkalla et al [81] developed a scoring system based on the cross-correlation of low frequency bands in the two channels of EOG, and found 72% agreement between full montage visual and their automatic scoring. Other researchers developed a continuous marker for sleep depth using the Short-time Fourier Transform and/or AR modelling [7]. Reduction of feature space dimensionality has also been addressed using the minimal redundancy method [23] or mutual information [71]. In the detection of seizures in epilepsy in patients, wavelet decomposition and chaos analysis of EEG signals have been used to provide features in which dimensionality has been reduced using Principal Component Analysis [44].

1.2.2 Recent EEG signal classification methods

After features have been selected and extracted from EEG signals, classification begins. Classification algorithms can be divided into different categories based on different perspectives such as linear classifiers, nonlinear classifiers, and combinations of classifiers [62].

The popular linear classifiers used in EEG signal classifications are linear discriminate analysis (LDA) and linear support vector machines (SVM). LDA [33] assumes that the data in each class has normal (Gaussian) distribution all having the same covariance matrix. The separating hyperplane is constructed by seeking the projection that maximizes the distance between the means of two classes and minimizes the variance of interclass. LDA classifier has a very low computational requirement which makes it suitable as online applications [42]. The main drawback of LDA is that it gives poor performance on complex nonlinear EEG data [41]. Linear SVM [33] aims to find a hyperlane that maximizes the margins, i.e., the distance from the nearest training points. Linear SVM has been successfully applied to synchronous brain computer interface (BCI) problems [42]. By using "kernel trick" the linearity restriction can be relaxed so that nonlinear decision boundaries can be created, with only a low increase of the classifier's complexity. The radial basis function (RBF) SVM also have successful applications in EEG signal classification [42]. SVM has good generalization properties due to the margin maximization and the regularization. It is insensitive to overtraining. It overcomes the problem of "curse-of-dimensionality". The drawback is the low speed of execution.

The nonlinear classifiers mostly used in EEG signal classification are the Nonlinear Bayesian classifiers [56]. Another choice is the Hidden Markov model (HMM) classifiers because it is not necessary to extract feature vectors from EEG signals for the classification. HMM has been used successfully in BCI [69, 70] and sleep staging [32].

A neural network can be viewed as universal approximator of continuous functions. Thus, it can produce nonlinear decision boundaries when used in classification [19]. However, the universality makes the classifiers sensitive to overtraining, especially with noisy and non-stationary data. Therefore, one must be careful to select the architecture and regularization [52]. Multilayer perceptron (MLP), together with linear classifiers, are the neural networks mostly used in EEG signal classifications [49, 9, 82]. Other neural network architectures have also been applied to EEG signal classifications [65].

The k-Nearest neighbor classifiers are the simplest and among the most effective nonlinear classifiers. The idea is to assign a feature vector to a class according to its nearest neighbors. The neighbors can be feature vectors from the training set if a distance measure is defined between feature vectors [20], or class prototypes if Mahalanobis distance is used [25]. The performance of a k-nearest neighbor classifier can be equal to that of a neural network classifier in the automatic scoring of human sleep recordings [11]. A more detailed introduction of k-nearest neighbor classifier will be given in Chapter 5.

There are others who suggested the use of several classifiers in cascade, each classifier focusing on the errors committed by the previous ones [33, 52]. However, the complexity of the classification will also dramatically increase.

As introduced in the above, if EEG signals can be represented by feature vectors of appropriate size in the sense of low dimensionality, then there are various choices of classifiers to carry out the classification [33]. However, feature extraction is not a trivial problem in the sense there is no way to guarantee that features extracted from EEG observations are good for classification. Furthermore, based on the same feature space, different classifiers often give very different classification performance. Examples of such difference in performance can be found in [44].

The above approaches of automatic EEG classification for sleep state determination have limited success rate and have thus not been widely used. At present, the common practice is still to have the EEG signals determined by sleep experts whose judgments are based mainly on the frequency and amplitude of the recorded EEG signals which are governing factors of the power distribution of the signal. Since modern EEG and polysomnography systems (e.g. Xltek, Oakville, Ontario, Canada) are computer-based, EEG data that have already been classified by expert clinicians are now readily available to test new approaches.

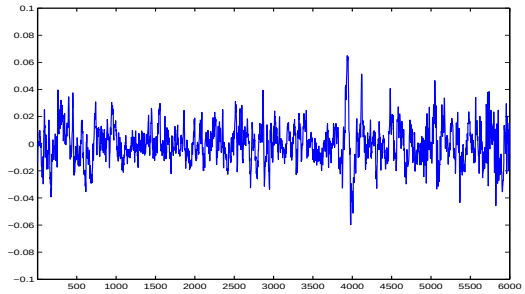
1.3 A new outlook on EEG signal classification

Signal classification is essentially a process of measuring the similarity and dissimilarity of the feature of a signal from different known feature sets. The measure of similarity/dissimilarity is generally based on the concept of *distance*. The most commonly used approaches to the problem is from a vector space point of view [36, 73] in which the selected features of the different classes of signals are treated as entities in a vector space prescribed with a distance measure. Here in this section, we will give an introduction to our outlook on the geometry of this vector space which forms the basis of the thesis. First, let us examine a feature of the EEG signal which may be attractive for practical sleep classification.

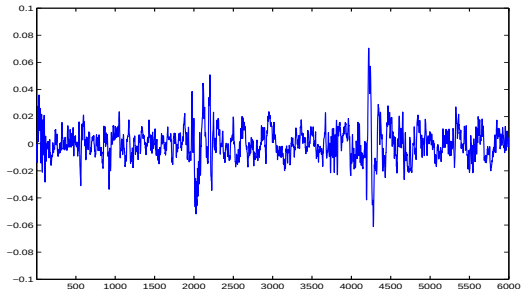
1.3.1 PSD – an EEG signal feature for sleep assessment

In the previous section, we mentioned that at present, the common practice of determining a patient's sleep stage is still by having sleep experts to visually inspect the EEG signals and to make judgments based mainly on the frequency and amplitude of the recorded EEG signals. These are the governing factors of the power distribution

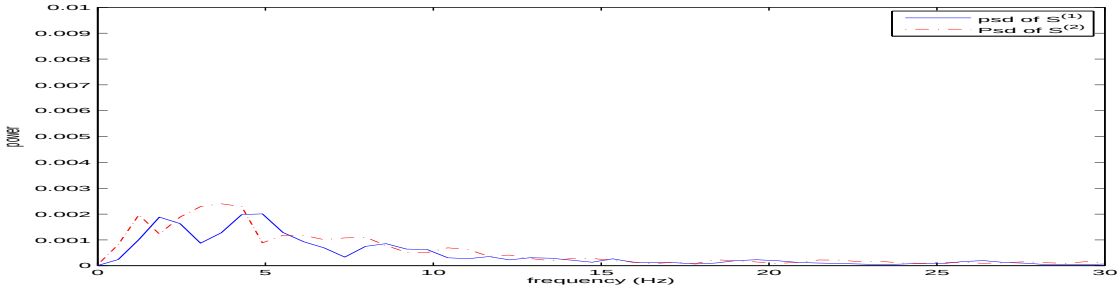
of the signals. That classification of sleep stages can be carried out by inspecting



(a) $\mathbf{s}^{(1)}$: An EEG signal from sleep state 1 of subject A



(b) $\mathbf{s}^{(2)}$: An EEG signal from sleep state 1 of subject B

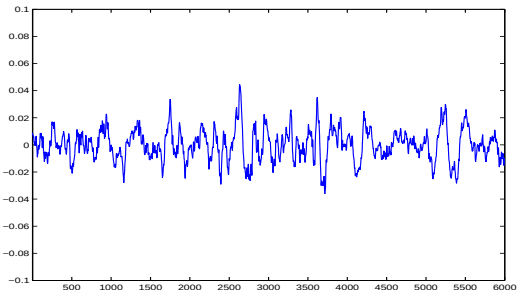


(c) Power spectral densities of the EEG signals $\mathbf{S}^{(1)}$ and $\mathbf{s}^{(2)}$

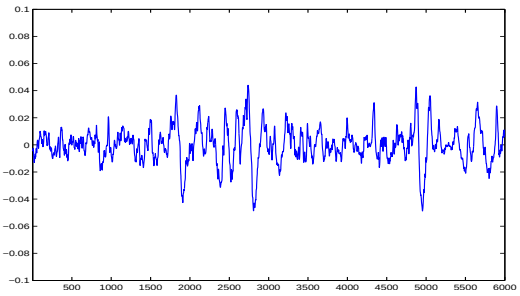
Figure 1.9: Example 1

the amount of power in certain frequency ranges can be illustrated by studying the power spectral densities of EEG signals in the following examples. Figure 1.9 shows the case of two EEG signals and their power spectral densities from one class, and Figure 1.10 shows the case of two EEG signals and their power spectral densities from a different class. To compare these power spectral densities we put them in one figure as shown in Figure 1.11.

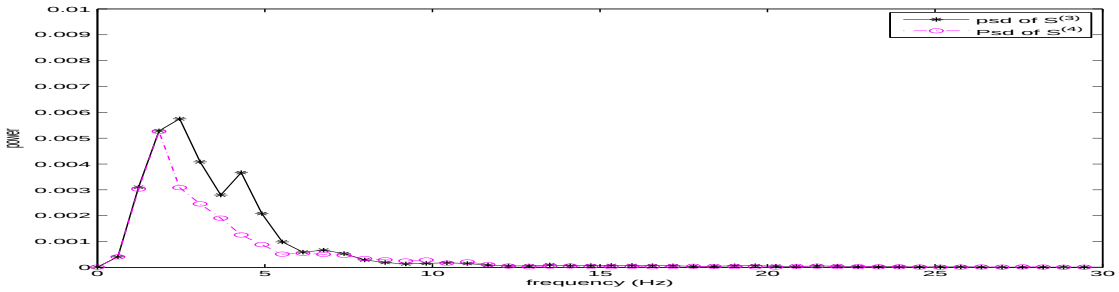
It can be seen that EEG signals from the same class have similar power spectral densities and signals from different classes have obvious different power spectral densities in the sense of their shapes. Therefore, we conclude that the judgements of the sleep experts have sound basis and that it is reasonable to represent EEG signals by



(a) $\mathbf{s}^{(3)}$: An EEG signal from sleep state 3 of subject A



(b) $\mathbf{s}^{(4)}$: An EEG signal from sleep state 3 of subject B



(c) Power spectral densities of the EEG signals $\mathbf{s}^{(3)}$ and $\mathbf{s}^{(4)}$

Figure 1.10: Example 2

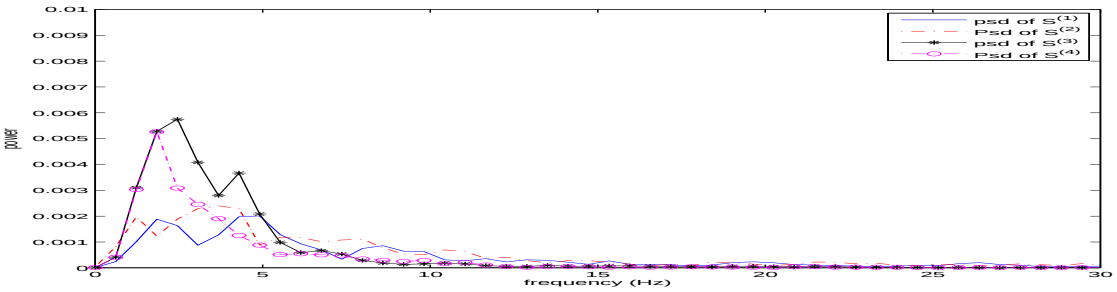


Figure 1.11: Comparison of the spectral densities of the EEG signals $\mathbf{s}^{(1)}$, $\mathbf{s}^{(2)}$, $\mathbf{s}^{(3)}$, and $\mathbf{s}^{(4)}$

their power spectral densities for sleep classification purposes. Since it is essential for our results of the automatic EEG classification to have meanings which concur with judgements of the clinical experts, in our studies, we have chosen the PSD of the EEG signals as the selected feature. In particular, since the EEG signals are collected from multi-channel measurements, we use the *PSD matrices* of the multi-channel EEG signals as our features. This will not only provide us with the power density distribution information of the signals, but will also provide us with the information of the cross power density distributions between the signals from the different channels. The use of the PSD matrix as the chosen feature can further be justified by noting that the EEG signals are generally considered to be wide-sense stationary (WSS) processes, and therefore, can be represented by their second order moments, or equivalently, their PSD matrices (see Chapter 2).

1.3.2 Distance measures for signal classification

Let us turn to the process of classification. Supervised signal classification is a comparison of similarity/dissimilarity between a signal and a standard group of signals so that a decision can be made. To this end, we define a *distance* between a pair of signals. Now, the set of signals itself begins to take on a geometric character called a *signal space* [36]. (A more detailed discussion of this concept is given in Chapter 2). The extracted features of the signals can also form a signal space on which a distance can also be defined for the purpose of classification. For example, suppose the signals (or their features) are represented by the set $\mathbb{R}^n$ (or $\mathbb{C}^n$) of ordered sequences of n real (or complex) numbers (n-tuple) such that $\mathbf{x} = [x_1, \cdots, x_n]$. Then, the totality of n-tuples of values of $\{x_1, \cdots, x_n\}$ constitutes a real (or complex) signal space of n-dimensions. Each of the n-tuples is called a point in the space. There are many ways of defining the distance between two points in a signal space, each providing the

space with different geometric characteristics and application advantages [77]. Some examples of such measures will be discussed in Chapter 3. A standard metric in this n-dimensional signal space is the *Euclidean distance* such that

$$d_2(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \tag{1.1}$$

This metric is used in a majority of engineering measurements due to the many important physical quantities it can represent.

However, while the Euclidean metric is very useful in most physical applications, it may not be the most appropriate measure for some. In our case of EEG signal classification, we use the PSD matrices of the EEG signals as features. These PSD matrices form different points in the signal space in which our classification is performed. Now, if we examine these PSD matrices, we observe that they are: 1) Hermitian symmetric, and 2) positive definite. These common properties of the PSD matrices describe a hyper-surface, called a *manifold*, in the signal space on which these points of PSD matrices are located. More specifically, these PSD matrices describe a *Riemannian manifold* [40], which is a particular kind of differentiable manifold (for further details, see in Appendix B). Now, if measurement of the distance between two points is to be carried out for the purpose of classification, a reasonable way is to measure the distance along the shortest path on the manifold between the points. This is analogous to finding the distance between two cities on a globe in which case the shortest path between two points on the globe surface has to be established and measured. The Euclidean distance which measures the straight line joining the two points may be neither appropriate nor informative. The following example further reinforces this idea:

Suppose that we have some data points which are distributed on a curved surface in $\mathbb{R}^3$ as shown in Figure 1.12. Now, consider the three points A, B, and C. Here,

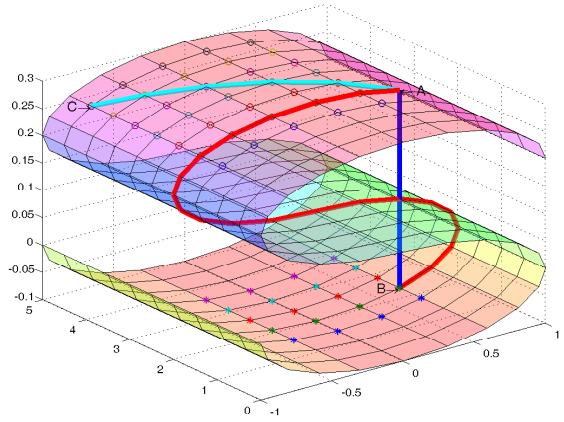


Figure 1.12: Data on a curved surface

the Euclidean distance between two points is measured in terms of the length of the straight line between them. Thus, the Euclidean distance between A and B is shorter than the Euclidean distance between A and C. On the other hand, if we measure the distance in terms of the shortest length along the curved surface joining the two points (called the *intrinsic distance*), then clearly, the distance between points A and B is much longer than the distance between points A and C. For this given data set, the use of the intrinsic metric to measure similarity/dissimilarity may yield more appropriate results than the use of Euclidean distance in practice.

Therefore, in this thesis, we explore the geometry of the Riemannian manifold of the EEG PSD matrices. By considering the tangent space at a point on the feature manifold, we can develop a suitable Riemannian metric from which a *geodesic* (the curve of minimum distance) between two points on the manifold can be established. The direct evaluation of the Riemannian (geodesic) distance may, in general, be very complicated or even untractable. However, with the help of fibre bundle theory, a straightforward derivation of the distance can be obtained.

Furthermore, even though we have seen that the PSD matrices may be a reasonable choice of a signal feature for EEG classification in the determination of sleep levels, there is no guarantee that this choice will yield the optimum separability between different signal classes. We therefore propose to weight the Riemannian distance so that the distinction between similar and dissimilar signal groups may be enhanced. To this end, we seek an optimum weighting matrix for the features using convex optimization techniques. A closed form of the weighting matrix can then be obtained.

Using the optimally weighted Riemannian distance, we can employ a classifier to carry out the EEG classification. We use the k-nearest neighbor classifier in this thesis due to its relative simplicity. The effectiveness of our new geometrical approach to EEG classification can then be thoroughly tested.

1.4 Main features of the thesis

The following are the main features of this thesis:

1. The PSD matrix is chosen as the feature representing an EEG signal for classification purpose. Therefore, EEG epochs are represented as curves on the manifold of PSD matrices.

2. Geodesic distances are developed with chosen Riemannian metrics endowed to the manifold by using elementary Riemannian geometry. Applying fibre bundle theory, complex computation is avoided in the evaluation of the geodesic distance by establishing an isometric horizontal subspace of the tangent space at the image of the point considered on the manifold.

3. The similarity/dissimilarity measure between two EEG signals is defined in

terms of the geodesic distances.

4. A general distance metric learning problem is proposed. In particular, an optimum weighting matrix for the geodesic distance on the manifold of PSD matrices is found in a closed form.

5. k-nearest neighbor classification rule is applied. To reduce the computational load for the Riemannian distances between a large number of points in the case of very large training sets, a multi-mean representation of classes is proposed and applied.

6. Experimental results show the power of the discrimination of the classification method developed.

Features listed in Items 2, 3, and 4 are considered major research contributions of this thesis.

Chapter 2

EEG Signals – Pre-Processing, Feature Extraction and Mathematical Representations

For efficient classification, the EEG signals have to be reasonably free from artifacts and other interference so that the signal and its feature characteristics can be determined accurately. In this chapter, we examine the collected EEG signals which are in segments of 30-sec epochs with a sampling frequency of 200Hz. The pre-processing of the EEG signals collected from the patient is first carried out so that artifacts are removed and additive noise reduced. Then, we examine the properties of the PSD as a feature of the EEG signal and describe how this feature can be extracted from the collected signals. We then present a general mathematical method of representing the EEG signals and their PSD matrices.

2.1 Pre-processing of EEG signals

As mentioned in Chapter 1, EEG signal measurements are subject to the interference of noise and other internal and external artifacts. The effects of these interference may often lead to the degradation of the classification performance. Therefore, before the process of classification, the EEG signals must first be pre-processed so as to remove the artifacts and reduce the noise in the signals.

2.1.1 Artifact removal

Artifacts in EEG signals are usually caused by movements internal or external to the patient. Artifacts removal is still one of the challenges in EEG signal processing. The problem is that there is no definite shape or size or duration of the artifacts. At present, the common practice of the clinical experts is to identify the artifacts by visual inspection and then replace the artifact samples. Other methods have been proposed to remove artifacts from EEG recordings including regression in time/frequency, and linear decomposition and reconstruction, etc. [46, 86, 5, 92, 24, 76]. Since this is not the main theme of the thesis, we will simply follow the common practice of visual inspection. A brief description of our procedure is given in the following:

We notice that the amplitudes of artifacts in the collected EEG signals are usually very large and very short in duration compared to that of the the normal EEG signals. An example of such is shown in Figure 2.1 in which an EEG signal of sleep stage 1 contains an artifact during time interval $20 - 20.1$ seconds.

In the pre-processing of such artifact infested signals, we measure the mean μ and standard deviation σ of the EEG signal epoch. Treating the distribution of the EEG signal as if it is Gaussian, any sample which has an amplitude larger than $|\mu + 3\sigma|$ will be removed [10] and replaced by random samples within the range of $\pm|\mu + 3\sigma|$.

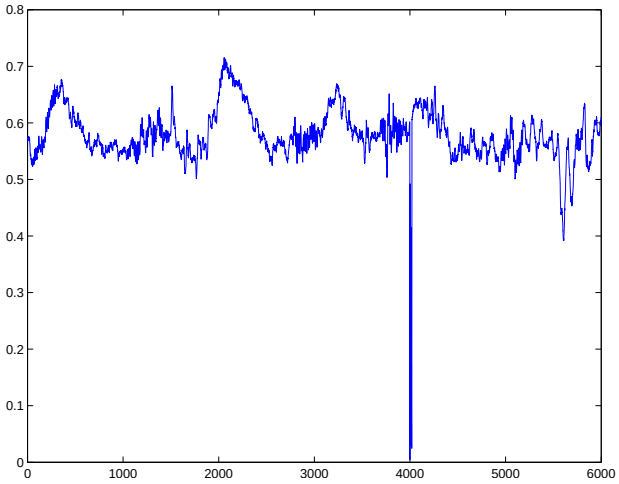


Figure 2.1: EEG signal with artifact

Figure 2.2 shows the EEG signal of Figure 2.1 after the artifact samples are removed.

2.1.2 Noise filtering

Since in our applications, all the EEG signals concentrate in the frequency range of $0 - 35$ Hz, therefore, to reduce the additive noise in the recorded EEG signals, we apply low-pass filtering to the signal epochs after the artifacts have been removed. To ensure a relatively low distortion to the signal we choose the Butterworth filter design since it has a *maximally flat* amplitude response and a relatively linear phase response in the pass-band [18]. It should also have a relatively narrow transition band. For these requirements, a tenth order low-pass Butterworth filter with the cut-off frequency of 58Hz is chosen and is realized as an Infinite-duration Impulse Response (IIR) digital filter [34]. The transfer function of such a filter is given by [34]

$$H(z) = \frac{N(z)}{D(z)} \tag{2.1}$$

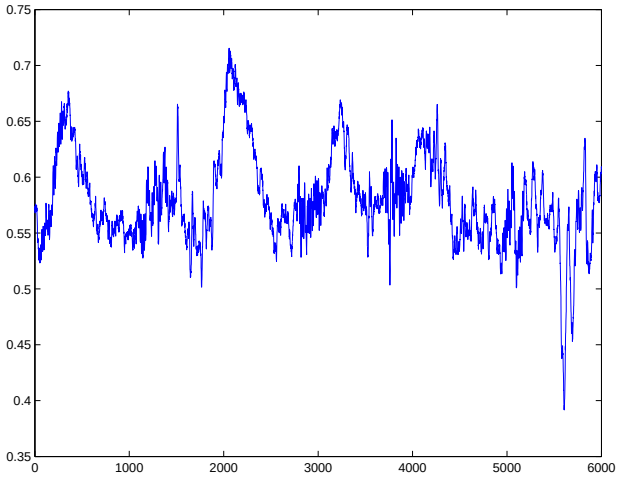


Figure 2.2: EEG signal with artifact removed

where

$$\begin{aligned} N(z) &= 0.0093z^{10} + 0.0932z^9 + 0.4193z^8 + 1.1180z^7 + 1.9565z^6 + 2.3478z^5 \\ &\quad +1.9565z^4 + 1.1180z^3 + 0.4193z^2 + 0.0932z + 0.0093 \end{aligned} \tag{2.2}$$

and

$$\begin{aligned} D(z) &= 1.0000z^{10} + 1.5938z^9 + 2.4143z^8 + 2.0262z^7 + 1.4469z^6 + 0.7003z^5 \\ &\quad +0.2712z^4 + 0.0723z^3 + 0.0139z^2 + 0.0016z + 0.0001 \end{aligned} \tag{2.3}$$

The amplitude and phase response are shown in Figure 2.3.

It can be seen that the amplitude response is flat from 0 to 35 Hz (corresponding to the normalized frequency of $2 \times \omega/\omega_s = 2 \times 35/200 = 0.35$, where ω is the frequency in radians/second and ω_s is the sampling frequency) and the phase response is approximately linear in the range of $0 - 35$ Hz.

At the transition band, although the amplitude response rolls off relatively more gently than some other designs (e.g., Chebychev, elliptic, etc.), this has insignificant

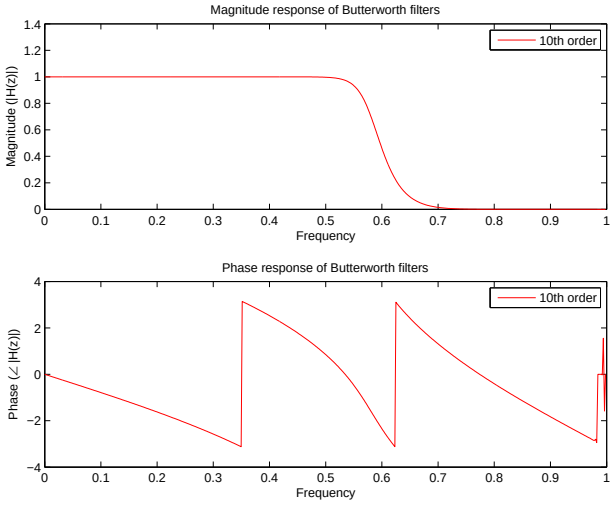


Figure 2.3: The amplitude and phase response of the filter

effects on our purpose because the EEG signals have negligible amplitudes beyond the frequency range of 35 Hz. The EEG signal of Figure 2.2 after passing through the digital Butterworth IIR low-pass filter is shown in Figure 2.4.

2.1.3 EEG signal normalization and data collection

After the above clean-up procedures, we can now collect all the M channel measurements for the ith patient and represent each of the preprocessed nth epoch of these multi-channel data at time t as a vector:

$$\mathbf{s}'^{(i)}_n(t) = [s'^{(i)}_{n1}(t), \cdots, s'^{(i)}_{nM}(t)]^T, \qquad t = 1, \cdots, T \tag{2.4}$$

Thus, the nth epoch measured data matrix (representing M channels of measured data for a duration of T seconds) for the ith patient is given by

$$\mathbf{S}'^{(i)}_n = [\mathbf{s}'^{(i)}_n(1), \cdots, \mathbf{s}'^{(i)}_n(T)], \qquad n = 1, \cdots, N^{(i)} \tag{2.5}$$

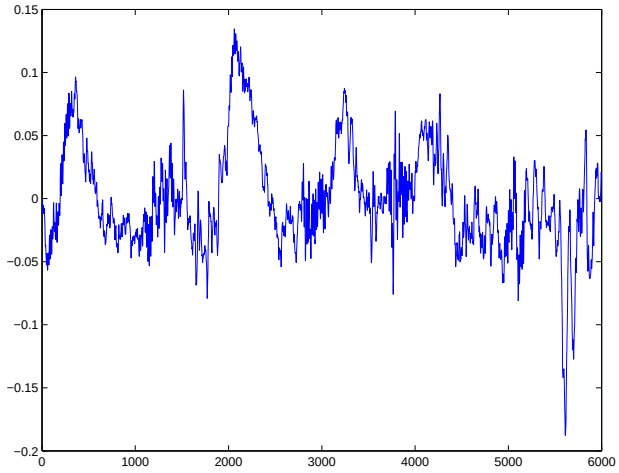


Figure 2.4: The filtered EEG signal

Each epoch of measured EEG data is then normalized such that the normalized data matrix is given by

$$\mathbf{S}_n^{(i)} = \frac{\mathbf{S}'^{(i)}_n}{\|\mathbf{S}'^{(i)}_n\|_F} = \frac{\mathbf{S}'^{(i)}_n}{\sum_{i=1}^{M}\sum_{j=1}^{T}[\mathbf{S}'^{(i)}_n]_{ij}} \tag{2.6}$$

with $[\mathbf{S}'^{(i)}_n]_{ij}$ denoting the ijth element of $\mathbf{S}'^{(i)}_n$. These normalized data matrices are then carefully inspected and classified by several clinical experts and is labeled indicating that it represents a particular state of sleep for the patient. Thus, for the ith patient, we have the following labeled sample EEG signals:

$$\mathcal{D}^{(i)} = \left\{ \begin{bmatrix} \mathbf{S}_1^{(i)} \\ \ell_1^{(i)} \end{bmatrix}, \cdots, \begin{bmatrix} \mathbf{S}_n^{(i)} \\ \ell_n^{(i)} \end{bmatrix}, \cdots, \begin{bmatrix} \mathbf{S}_{N^{(i)}}^{(i)} \\ \ell_{N^{(i)}}^{(i)} \end{bmatrix} \right\} \tag{2.7}$$

where

$$\ell_n^{(i)} \in \mathcal{L} = \{1, 2, \cdots, L\} \tag{2.8}$$

denotes the label of the nth epoch of the EEG signal belonging to any of the L states of sleep. A library of these labeled sampled EEG signals from several patients are

stored up such that $\mathcal{D} = \bigcup_i \mathcal{D}^{(i)}$ having a total of $N = \sum_i N^{(i)}$ epochs of data. Since the goal of the research in this thesis is to derive a reliable classification method to automatically determine the label ℓ_0 of a new normalized measured EEG data matrix $\mathbf{S}_0$ without having to involve expert human judgement, this collection of data will serve as a reference library as well as the supply of test data for formulating the classification measure and testing the performance of the classification methods.

2.2 EEG signals – wide-sense stationarity

We now turn our attention to the feature characterization of the EEG signals. Since there is no deterministic pattern of an EEG signal, it is usually regarded as stochastic. Let us first review some basic properties of stochastic processes [72]. A stochastic process $s(t, \xi)$ can be viewed as a real (or complex) valued function of two variables t and ξ. The domain of ξ is the set $\mathcal{S}$ of outcomes of an experiment and the time domain of t is a set of real numbers. For a specific outcome ξ_i, $s(t, \xi_i)$ signifies a single time function. For a specific time t_i, $s(t_i, \xi)$ is seen as a random variable. We usually use $s(t)$ to representation of a stochastic process with its dependence on ξ omitted.

For a real process $s(t)$, the value $s(t)$ at a specific t is a random variable. The distribution of this random variable will depend on t in general, i.e., we have

$$F(s, t) = P(s(t) \leq s) \tag{2.9}$$

where $P(\cdot)$ denotes the probability of the event. Eq. (2.9) is called the first-order distribution of the process $s(t)$. In most situations, the distribution has a probability density function which can be defined as

$$f(s, t) = \frac{\partial F(s, t)}{\partial s} \tag{2.10}$$

The joint distribution of $s(t_1)$ and $s(t_2)$ depends, in general, on t_1 and t_2, i.e.,

$$F(s_1, s_2, t_1, t_2) = P(s(t_1) \leq s_1, s(t_2) \leq s_2) \tag{2.11}$$

which is called the second order distribution of the process $s(t)$. The corresponding density function is given by

$$f(s_1, s_2, t_1, t_2) = \frac{\partial^2 F(s_1, s_2, t_1, t_2)}{\partial s_1 \partial s_2} \tag{2.12}$$

In general, it is costly to measure the distribution $F(s, t)$ experimentally and the computation of the probability density function $f(s, t)$ is extremely difficult. They are also too cumbersome to be used in practice. A simpler alternative to this form of description is to compute a number of average characteristics of a process. In other words, the moments of a probability distribution serve as simple numerical characteristics of the distribution.

The mean $\mu(t)$ of a process $s(t)$ is defined as the expected value of the random variable $s(t)$ (at a fixed t), i.e.,

$$\mu(t) = \mathbb{E}[s(t)] = \int_{-\infty}^{\infty} s f(s, t) ds \tag{2.13}$$

It is, in general, a function of t. The autocorrelation of $s(t)$ is defined as the joint moment of the random variable $s(t_1)$ and $s(t_2)$, i.e.,

$$r(t_1, t_2) \triangleq \mathbb{E}[s(t_1)s(t_2)] = \int_{-\infty}^{\infty} s_1 s_2 f(s_1, s_2, t_1, t_2) ds_1 ds_2 \tag{2.14}$$

and it is a function of t_1 and t_2. The autocovariance of $s(t)$ is the covariance of the random variable $s(t_1)$ and $s(t_2)$, i.e.,

$$c(t_1, t_2) \triangleq \mathbb{E}[(s(t_1) - \mu(t_1))(s(t_2) - \mu(t_2))] \tag{2.15a}$$

$$= r(t_1, t_2) - \mu(t_1)\mu(t_2) \tag{2.15b}$$

The relationship between autocorrelation and autocovariance in Eq. (2.15b) follow directly from Eqs. (2.14) and (2.15a).

The cross-correlation of two processes $s_1(t)$ and $s_2(t)$ is defined as

$$r_{s_1 s_2}(t_1, t_2) \triangleq \mathbb{E}[s_1(t) s_2^H(t)] \tag{2.16}$$

and their cross-covariance as

$$c_{s_1 s_2}(t_1, t_2) = r_{s_1 s_2}(t_1, t_2) - \mu_{s_1}(t_1)\mu_{s_2}^H(t_2) \tag{2.17}$$

A process $s(t)$ with distribution function $F(s, t)$ is called a wide sense stationary (WSS) process if it satisfies the following two conditions

(1) The mean value of $s(t)$ is a constant, i.e.,

$$\mu(t) = \mu, \quad \text{a constant} \tag{2.18}$$

(2) The autocorrelation function depends only on the time difference $\tau = t_1 - t_2$, i.e.,

$$r(t_1, t_2) = r(t_1 - t_2) = r(\tau) \tag{2.19}$$

We say that two processes $s_1(t)$ and $s_2(t)$ are jointly stationary in the wide sense if each of them is a WSS process and their cross-correlation depends only on the time difference $\tau = t_1 - t_2$:

$$r_{s_1 s_2}(\tau) = \mathbb{E}[s_1(t + \tau) s_2(t)] \tag{2.20}$$

These averages do not necessarily describe a stochastic signal completely, but they may be very useful for a general description of signals such as EEG. In fact, the statistical properties of EEG signals depend on both time and space. These make EEG signals complex. The temporal characteristics show that EEG signals are varying from time to time. However, each time series can be divided into epochs

which have more or less time-invariant statistical properties [84]. It is thus commonly assumeed that EEG epochs (awake condition or during sleep) of less than 32 seconds are wide-sense stationary [64].

2.3 Power spectral density – A feature characterization of EEG signals

The power spectral density (also called power spectrum) function, $p(\omega)$, and the autocorrelation, $r(\tau)$, of a WSS process $s(t)$ form a Fourier transform pair (see [91] for a rigorous treatment) such that , i.e.,

$$p(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} r(\tau)e^{-j\omega\tau}d\tau \tag{2.21}$$

$$r(\tau) = \int_{-\infty}^{\infty} p(\omega)e^{j\omega\tau}d\omega \tag{2.22}$$

whereas the cross-power spectral density, $p_{s_1s_2}(\omega)$, and the cross-correlation, $r_{s_1s_2}(\tau)$, of two WSS processes also form a Fourier transform pair:

$$p_{s_1s_2}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} r_{s_1s_2}(\tau)e^{-j\omega\tau}d\tau \tag{2.23}$$

$$r_{s_1s_2}(\tau) = \int_{-\infty}^{\infty} p_{s_1s_2}(\omega)e^{j\omega\tau}d\omega \tag{2.24}$$

The spectral representation of a univariate WSS process can be generalized to multidimensional case in a straightforward way. Let us consider a WSS M channel EEG signal

$$\mathbf{s}(t) = [s_1(t)\ s_2(t)\ \cdots\ s_M(t)]^T,\ \ t = 1, 2, \cdots, T \tag{2.25}$$

It can be written in a matrix form as

$$\mathbf{S} = [\mathbf{s}(1), \cdots, \mathbf{s}(T)] = \begin{bmatrix} s_1(1) & \cdots & s_1(T) \\ s_2(1) & \cdots & s_2(T) \\ \vdots & \ddots & \vdots \\ s_M(1) & \cdots & s_M(T) \end{bmatrix} \tag{2.26}$$

We make an $MT \times MT$ vector by stacking columns of $\mathbf{S}$ as

$$\check{\mathbf{s}} = [\mathbf{s}(1)^T, \mathbf{s}(2)^T, \cdots, \mathbf{s}(T)^T]^T. \tag{2.27}$$

Then, this signal can be characterized by its mean and variance-covariance matrix, i.e.,

$$\check{\boldsymbol{\mu}} = \mathbb{E}[\check{\mathbf{s}}] \tag{2.28}$$

and

$$\check{\mathbf{R}} = \mathbb{E}[(\check{\mathbf{s}} - \check{\boldsymbol{\mu}})^T(\check{\mathbf{s}} - \check{\boldsymbol{\mu}})], \tag{2.29}$$

which contains the $M \times M$ matrices $\mathbf{R}(\tau) = \mathbf{R}(t_1 - t_2)$, $t_1, t_2 = 1, \cdots, T$, specifically

$$\check{\mathbf{R}} = \begin{bmatrix} \mathbf{R}(0) & \mathbf{R}(1) & \cdots & \mathbf{R}(T-1) \\ \mathbf{R}(-1) & \mathbf{R}(0) & \cdots & \mathbf{R}(T-2) \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{R}(1-T) & \mathbf{R}(2-T) & \cdots & \mathbf{R}(0) \end{bmatrix}, \tag{2.30}$$

where

$$\mathbf{R}(\tau) = \mathbf{R}(t_1 - t_2) = \begin{bmatrix} r_{11}(\tau) & r_{12}(\tau) & \cdots & r_{1M}(\tau) \\ r_{21}(\tau) & r_{22}(\tau) & \cdots & r_{2M}(\tau) \\ \vdots & \vdots & \vdots & \vdots \\ r_{M1}(\tau) & r_{M2}(\tau) & \cdots & r_{MM}(\tau) \end{bmatrix}, \tag{2.31}$$

where

$$r_{ij}(\tau) = \mathbb{E}[(s_i(t) - \mu_i)(s_j(t+\tau) - \mu_j)]. \tag{2.32}$$

Since the EEG signals we considered are real, we have

$$\mathbf{R}^H(\tau) = \mathbf{R}(-\tau). \tag{2.33}$$

Thus, the matrix $\check{\mathbf{R}}$ can be completely characterized by the $2T$ elements $\{\mathbf{R}(1-T), \cdots, \mathbf{R}(-1), \mathbf{R}(0), \mathbf{R}(1), \cdots, \mathbf{R}(T-1)\}$.

The information contained in the covariances can be expressed equivalently in terms of the power spectral density matrix, $\mathbf{P}(\omega)$, of the signal which is the Fourier transform of the autocorrelation of matrix such that

$$\mathbf{P}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \boldsymbol{R}(\tau) e^{-j\omega\tau} d\tau \tag{2.34}$$

if

$$\int_{-\infty}^{\infty} \|\boldsymbol{R}(\tau)\| d\tau < \infty. \tag{2.35}$$

where $\|\cdot\|$ denotes the ℓ_1-norm of the matrix. Eq. (2.34) gives the power spectral density for a continuous signal in terms of its autocorrelation function. For observed signals in discrete time as is in the case of EEG signal measurements, we need to take the discrete Fourier transform (DFT) of the signal autocorrelation so that

$$\mathbf{P}(\omega) = \sum_{\tau=-\infty}^{\infty} e^{-j\omega\tau} \boldsymbol{R}(\tau) \tag{2.36}$$

In practice, the autocorrelation matrix is evaluated by taking the product of the finite signal sequences at different time shifts. That the dDFT of this product indeed converges to Eq. (2.36) can be seen using the following lemma [51]:

Lemma 2.1 *Let* $\rho_n = a_0 + \cdots + a_n$ *be the partial sum of a series* $\sum_{k=0}^{n} a_k$. *Let* $\beta_n = (\rho_0 + \cdots + \rho_{n-1})/n = \sum_{k=0}^{n-1}(1 - \frac{k}{n})a_k$ *be the Cesàro mean (average sum). If* $\lim_{n\to\infty} \rho_n = \rho$, *then* $\lim_{n\to\infty} \beta_n = \rho$.

Assuming $\mathbb{E}[\mathbf{s}(t)] = \mathbf{0}$ without loss of generality, we consider

$$\mathbf{I}(\omega) = \frac{1}{2\pi T}\left[\sum_{t=1}^{T}\mathbf{s}(t)e^{-j\omega t}\right]\left[\sum_{t=1}^{T}\mathbf{s}(t)e^{-j\omega t}\right]^{H}. \tag{2.37}$$

Taking the expectation we have

$$\mathbb{E}[\mathbf{I}(\omega)] = \frac{1}{2\pi}\sum_{\tau=-T+1}^{T-1}\left(1-\frac{|\tau|}{T}\right)e^{-j\omega\tau}\boldsymbol{R}(\tau), \tag{2.38}$$

which is a Cesàro mean of the series for

$$\mathbf{P}(\omega) = \sum_{\tau=-\infty}^{\infty} e^{-j\omega\tau}\boldsymbol{R}(\tau) \tag{2.39}$$

By Lemma 2.1, the series (2.38) is convergent and

$$\lim_{T\to\infty}\mathbb{E}[\mathbf{I}(\omega)] = \mathbf{P}(\omega) \succeq \mathbf{0} \tag{2.40}$$

since $\mathbb{E}[\mathbf{I}(\omega)] \succeq \mathbf{0}$. The fact that $\mathbf{P}(\omega)$ is Hermitian and positive semi-definite follows from Equation (2.34) and $\boldsymbol{R}^{H}(\tau) = \boldsymbol{R}(-\tau)$.

The spectral density measures how the power of an EEG signal is distributed with frequency and has been commonly used as a feature for EEG signal classification. However, for most of the applications, either the power spectral information of a single channel EEG signal or the collection of the power information of several single channel EEG signals is used. The inter-channel information (cross-power spectrum) which may be important for EEG signal classification has not been used. Furthermore, if power spectral density is used as the feature for EEG signal classification, then the geometrical structure of the space it describes is essential to the definition of similarity/dissimilarity between EEG signals. For these reasons, we choose to employ the power spectral density matrices as the feature characterizing the multi-channel EEG signals. Examination and analysis of the geometry of the space of the power spectral density matrices leads to novel and efficient similarity/dissimilarity measures on the space for classification.

2.4 Feature extraction – Estimation of the PSD matrix

The power spectral density of a signal can be estimated using either non-parametric methods or parametric modeling methods [55]. Non-parametric methods are simple, but are in general, not consistent in the estimate of the power spectrum. Furthermore, they are limited in their ability to resolve closely spaced variations of frequency response when the number of data samples is limited. The parametric modeling approaches usually give higher accuracy in the spectral estimation of signals if the model is chosen appropriately. There are various choices of parametric modeling. Here, because of its relative simplicity, we use the vector auto-regression (VAR) model for the estimation of EEG power spectral density, a brief outline of which is presented in the following:

The autocorrelation function (ACF) of a spectrally white multichannel noise sequence $\mathbf{n}(t)$ satisfies

$$\mathbf{R_{nn}}(\tau) = \mathbb{E}[\mathbf{n}^H(t)\mathbf{n}(t-\tau)] = \mathbf{P_{nn}}\delta(\tau) \tag{2.41}$$

where $\mathbf{P_{nn}}$ is a constant $M \times M$ matrix. Thus its PSD matrix is a constant, i.e.,

$$\mathbf{P_{nn}}(\omega) = \mathbf{P_{nn}} \tag{2.42}$$

Now, the output signal of a q-th order vector auto-regression (VAR) model can be described as

$$\mathbf{s}(t) = -\sum_{\tau=1}^{q} \mathbf{A}(\tau)\mathbf{s}(t-\tau) + \mathbf{n}(t) \tag{2.43}$$

where $\mathbf{A}(\tau)$ are the $M \times M$ coefficient matrices and $\mathbf{n}(t)$ is the $M \times 1$ vector of a

spectrally white noise. Let $\mathbf{A}(0) = \mathbf{I}$. Then, the ACF of $\mathbf{n}(t)$ is

$$\begin{aligned}\mathbf{R_{nn}}(\tau) &= \mathbb{E}[\mathbf{n}(t)\mathbf{n}^T(t+\tau)] \\ &= \mathbb{E}\Big[\sum_{\kappa=0}^{q}\sum_{\iota=0}^{q}\mathbf{A}(\kappa)\mathbf{s}(t-\kappa)\mathbf{s}^T(t+\kappa-\iota)\mathbf{A}^T(\iota)\Big] \\ &= \sum_{\kappa=0}^{q}\sum_{\iota=0}^{q}\mathbf{A}(\kappa)\mathbf{R_{ss}}(\tau+\kappa-\iota)\mathbf{A}^T(\iota) \qquad (2.44)\end{aligned}$$

Taking the z-transform of Eq. (2.44), we have

$$\begin{aligned}\mathcal{Z}[\mathbf{R_{nn}}(\tau)] &= \sum_{\tau=-\infty}^{\infty}\mathbf{R_{nn}}(\tau)z^{-\tau} \\ &= \Big(\sum_{\kappa=0}^{q}\mathbf{A}(\kappa)z^{\kappa}\Big)\Big(\sum_{\tau=-\infty}^{\infty}\mathbf{R_{ss}}(\tau+\kappa-\iota)z^{-(\tau+\kappa-\iota)}\Big)\Big(\sum_{\iota=0}^{q}\mathbf{A}^T(\iota)z^{-\iota}\Big) \qquad (2.45)\end{aligned}$$

Let $z = e^{j\omega}$ and use Eq.(2.36), we have

$$\mathbf{P_{nn}}(\omega) = \Big(\sum_{\kappa=0}^{q}\mathbf{A}(\kappa)e^{j\omega\kappa}\Big)\mathbf{P_{ss}}(\omega)\Big(\sum_{\iota=0}^{q}\mathbf{A}^T(\iota)e^{-j\omega\iota}\Big) \qquad (2.46)$$

Therefore, we have

$$\begin{aligned}\mathbf{P_{ss}}(\omega) &= \Big(\sum_{\kappa=0}^{q}\mathbf{A}(\kappa)e^{j\omega\kappa}\Big)^{-1}\mathbf{P_{nn}}(\omega)\Big(\sum_{\iota=0}^{q}\mathbf{A}^T(\iota)e^{-j\omega\iota}\Big)^{-1} \\ &= \Big(\sum_{\kappa=0}^{q}\mathbf{A}(\kappa)e^{j\omega\kappa}\Big)^{-1}\mathbf{P_{nn}}\Big(\sum_{\iota=0}^{q}\mathbf{A}^T(\iota)e^{-j\omega\iota}\Big)^{-1} \qquad (2.47)\end{aligned}$$

by Eq. (2.42). Let

$$\mathbf{A}(\omega) = \sum_{\tau=0}^{q}\mathbf{A}(\tau)e^{-j\omega\tau} \qquad (2.48)$$

Then

$$\mathbf{A}^T(\omega) = \sum_{\tau=0}^{q}\mathbf{A}^T(\tau)e^{-j\omega\tau} \qquad (2.49)$$

Thus, the Eq.(2.47) can be rewritten as

$$\mathbf{P_{ss}}(\omega) = \mathbf{A}^{-1}(-\omega)\mathbf{P_{nn}}\mathbf{A}^{-T}(\omega) \qquad (2.50)$$

From Eq. (2.50) we see that to find the power spectral density matrices $\mathbf{P_{ss}}(\omega)$ of the signal $\mathbf{s}(t)$ one need to estimate the coefficient matrices $\mathbf{A}(\tau)$ in the VAR model of Eq. (2.43). We employ the Nuttall-Strand algorithm [68] [79] which is a well-known algorithm applying the observed signal sequence to estimate the coefficient matrices $\mathbf{A}(\omega)$ and the power spectral density $\mathbf{P_{nn}}$ of the spectrally white noise. The estimate of $\mathbf{P}(\omega)$ based on the VAR model is given by

$$\hat{\mathbf{P}}(\omega) = \hat{\mathbf{A}}^{-1}(-\omega)\hat{\mathbf{P}}_{\mathbf{nn}}\hat{\mathbf{A}}^{-T}(\omega) \tag{2.51}$$

where $\hat{(\cdot)}$ denote estimated quantity. The detailed description of the algorithm is shown in Appendix C. The estimated PSD matrix $\hat{\mathbf{P}}(\omega)$ so obtained will be used as the feature for the classification of the EEG signals in our sleep analysis.

2.5 Representation of the PSD matrix in linear vector spaces

Having collected all signals (or their features) exhibiting some common property into a set, our attention naturally turns to examining the distinctive properties of elements within the set. A particular signal is interesting only in relation to other signals in the set. A general approach for studying the properties of the elements of a signal set is to add some simple algebraic and geometric structures to the set. This can be achieved through the concept of a *signal space* (normed linear vector space). In this section we first review the concept of linear spaces [77] [87], in particular, an inner product space and show how, in general, the feature of PSD matrices can be represented in such a space.

2.5.1 Inner product linear space

2.5.1.1 Metric spaces

A general approach for characterizing the difference between two elements of a signal set is to assign to each pair of elements a positive, real number. This number will be interpreted as the "distance" between the elements. The set, with a suitably defined distance, will be referred to as a *signal space*. To define a distance, we need a functional which maps all pairs of elements from the set into the real line. Such a functional, $d : \{x, y\} \rightarrow \mathbb{R}$, is called a *metric* if it posses the following properties:

$$\left.\begin{array}{ll} d(x,y) \geq 0 \text{ and } d(x,y) = 0 \text{ iff } x = y & \\ d(x,y) = d(y,x) & \text{(symmetry)} \\ d(x,z) \leq d(x,y) + d(y,z) & \text{(triangular inequality)} \end{array}\right\} \tag{2.52}$$

A set of elements $\mathcal{X}$, together with a metric d, is called a *metric space* $(\mathcal{X}, d)$. It should be noted that two different metrics, defined on the same set of elements, formed two different metric spaces. For a given metric space $(\mathcal{X}, d)$, a sequence $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots$ is Cauchy if, for every positive real number $\epsilon > 0$, there is a positive integer N such that for all natural numbers $m, n > N$,

$$d(\mathbf{x}_m, \mathbf{x}_n) \leq \epsilon \tag{2.53}$$

A metric space $(\mathcal{X}, d)$ is *complete* if every Cauchy sequence in $(\mathcal{X}, d)$ has a limit that is also in $(\mathcal{X}, d)$.

2.5.1.2 Normed linear spaces

A linear space is a set of elements called vectors with the following properties:

A. For each pair of vectors $\mathbf{x}$ and $\mathbf{y}$ in the set, there is a corresponding vector in the set $\mathbf{x} + \mathbf{y}$ called the *sum* of $\mathbf{x}$ and $\mathbf{y}$, such that

(1) Addition is commutative: $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$

(2) Addition is associative: $\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z}$

(3) The set contains a unique vector $\mathbf{0}$ such that $\mathbf{x} + \mathbf{0} = \mathbf{x} \quad \forall\, \mathbf{x}$

(4) For each $\mathbf{x}$, there is a unique vector $(-\mathbf{x})$ such that $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$

B. There is a set of elements (called scalars) which form a *field* and an operation (called scalar multiplication) such that for every scaler α and every vector $\mathbf{x}$ there is a vector $\alpha\,\mathbf{x}$, and multiplication by scalars follows:

(1) $\alpha(\beta\,\mathbf{x}) = \alpha\,\beta\,\mathbf{x}$ (associative law)

(2) $1\,\mathbf{x} = \mathbf{x}$ and $0\,\mathbf{x} = \mathbf{0} \quad \forall\, \mathbf{x}$

(3) $\alpha(\mathbf{x} + \mathbf{y}) = \alpha\,\mathbf{x} + \alpha\,\mathbf{y}$ (distributive law)

(4) $(\alpha + \beta)\mathbf{x} = \alpha\,\mathbf{x} + \beta\,\mathbf{y}$ (distributive law)

The scalars can be real or complex resulting in the linear space being a *real* or *complex* linear space. Under the above vector addition and scalar multiplication, a vector space is *closed*.[1]. The vector obtained by taking the sum of n particular vectors, each multiplied by a scalar coefficient

$$\mathbf{x} = \sum_{i=1}^{n} \alpha_i\,\mathbf{x}_i \tag{2.54}$$

is called a *linear combination*. The set of all linear combinations of $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ forms a linear space. Furthermore, if we take a subset of

$$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\},\ \text{e.g. } \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}; m < n$$

then the set of linear combinations forms a linear space which is a subset of the linear space form from linear combinations of $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. Such a subset is called a

[1]The vector resulting from the operations of addition and scalar multiplication remains in the vector space

linear subspace. A set of vectors $\{\mathbf{x}_i; i = 1, 2 \ldots, n\}$ is said to be *linearly independent* if the relation

$$\sum_{i=1}^{n} \alpha_i \, \mathbf{x}_i = \mathbf{0} \tag{2.55}$$

can only be satisfied if each of the scalars α_i is zero. In other words, a vector in a linearly independent set cannot be expressed as a linear combination of the other vectors in the set. Let $\mathcal{X}$ be the space of linear combinations of n linearly independently vectors $\{\mathbf{x}_i, i = 1, 2, \ldots n\}$. Each vector in $\mathcal{X}$ is a unique linear combination of the $\{\mathbf{x}_i\}$ (a unique set of scalar coefficients). $\mathcal{X}$ is said to be an "n-dimensional" linear space. The set $\{\mathbf{x}_i\}$ is called a *basis* for $\mathcal{X}$, and $\mathcal{X}$ is said to be *spanned* by this basis. Any set of n linearly independent vectors in $\mathcal{X}$ will serve as a basis for $\mathcal{X}$; hence a linear space does not have a unique basis. Furthermore, if we take a subset of

$$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}, \text{ e.g. } \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}; m < n$$

then the set of linear combinations forms a linear space which is a subset of the linear space form from linear combinations of $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. Such a subset is called a "linear subspace".

Representation of Finite-Dimensional Vectors: Now let $\mathcal{M}$ be an arbitrary n-dimensional linear space spanned by the basis $\{\mathbf{u}_i; i = 1, 2, \ldots, n\}$. Any $\mathbf{x} \in \mathcal{M}$ can be expressed uniquely by

$$\mathbf{x} = \sum_{i=1}^{n} \alpha_i \, \mathbf{u}_i \tag{2.56}$$

The ordered sequence of scalar coefficients $\{\alpha_i\}$ can be interpreted as an n-tuple. Thus there is a one-to-one correspondence between vectors in the arbitrary space $\mathcal{M}$ and the space of n-tuples, and $\mathbb{R}^n$ or $\mathbb{C}^n$ can be used as a model for any real or complex n-dimensional space. We say that the n-tuple $\boldsymbol{\alpha} = \{\alpha_i\}$ is a *representation* (in $\mathbb{R}^n$ or $\mathbb{C}^n$) for $\mathbf{x}$ relative to the basis $\{\mathbf{u}_i\}$.

We now combine the geometric concepts associated with metric spaces with the algebraic concepts associated with linear spaces. This is accomplished by assigning a real number reflecting the "size" of any element in a linear space. This number is called the *norm* of a vector (denoted by $\|\mathbf{x}\|$) and can be defined in terms of any mapping from the linear space into the real line which satisfies the following properties:

$$\begin{aligned} &\text{a.} \quad \|\mathbf{x}\| \geq 0 \text{ and } \|\mathbf{x}\| = 0 \qquad \text{iff } \mathbf{x} = 0 \\ &\text{b.} \quad \|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \\ &\text{c.} \quad \|\alpha\,\mathbf{x}\| = |\alpha|\|\,\mathbf{x}\| \end{aligned} \tag{2.57}$$

Note that the norm of a vector is its distance from the origin. A normed linear space which is also complete as a metric space is called a *Banach Space*.

2.5.1.3 Inner product spaces

The final step in the development of signal spaces is to supply additional geometric structure in the form of an *inner product* relationship between pairs of vectors. We shall henceforth deal with complex linear spaces, since the real spaces can always be treated as a special case. The inner product is a mapping of ordered pairs of vectors in the linear space into the complex plane. The mapping, with images denoted by $\langle \mathbf{x}, \mathbf{y} \rangle$ in $\mathbb{C}$, satisfies the following properties:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle^* \tag{2.58a}$$

$$\langle \alpha\,\mathbf{x} + \beta\,\mathbf{y}, \mathbf{z} \rangle = \alpha\,\langle \mathbf{x}, \mathbf{z} \rangle + \beta\,\langle \mathbf{y}, \mathbf{z} \rangle \tag{2.58b}$$

$$\langle \mathbf{x}, \mathbf{x} \rangle \geq 0 \text{ and } \langle \mathbf{x}, \mathbf{x} \rangle = 0 \text{ iff } \mathbf{x} = \mathbf{0} \tag{2.58c}$$

From (2.58a) and (2.58b) we see that $\langle \alpha\,\mathbf{x}, \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle$, $\langle \mathbf{x}, \alpha\,\mathbf{y} \rangle = \alpha^* \langle \mathbf{x}, \mathbf{y} \rangle$ and that $\langle \mathbf{x}, \mathbf{x} \rangle$ is real. An important consequence of the definition of the inner product is that

$$\|\mathbf{x}\| = \langle \mathbf{x}, \mathbf{x} \rangle^{1/2} \tag{2.59}$$

is a valid norm for the linear space. Furthermore, from the properties of Eqs. (2.57), it is easy to show that

$$d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\| = \langle(\mathbf{x}-\mathbf{y}),(\mathbf{x}-\mathbf{y})\rangle^{1/2} \tag{2.60}$$

is a metric (2.52) and this metric is implied when we refer to the inner product linear space. The inner product thus induces a norm which in turn induces a metric, by (2.60), so that an inner product space is a metric space with a particular metric implied. An inner product space which is also complete, as a metric space, is called a *Hilbert Space*.

The following properties of inner product and norm are important:

For $\mathbf{x},\mathbf{v} \in \mathcal{X}$ we have

(i) Cauchy-Schwarz inequality:

$$|\langle \mathbf{x},\mathbf{y}\rangle| \leq ||\mathbf{x}||||\mathbf{y}|| \tag{2.61}$$

with equality if and only if $\mathbf{x} = \alpha\mathbf{y}$ for some scalar α.

(ii) Parallelogram law:

$$||\mathbf{x}+\mathbf{y}||^2 + ||\mathbf{x}-\mathbf{y}||^2 = 2(||\mathbf{x}||^2 + ||\mathbf{y}||^2). \tag{2.62}$$

(iii) Pythagorean theorem: If $\mathbf{x} \perp \mathbf{y}$ then

$$||\mathbf{x}+\mathbf{y}||^2 = ||\mathbf{x}||^2 + ||\mathbf{y}||^2. \tag{2.63}$$

2.5.2 Representation of matrices in a linear vector space

The very successful development of signal theory and signal analysis has been, by and large, based on the framework of normed linear vector spaces. The EEG signal

classification problem is no exception, and most researchers have followed the same steps by defining a normed linear vector space for the single-channel EEG signals and their features and analysed their properties for classification. However, when the multi-channel PSD matrix is chosen as the EEG feature, it is natural to try converting the set of the PSD matrices into a corresponding set of vectors. In the following, we present several common ways of the such vectorization methods.

1. We can represent an $M \times M$ matrix $\mathbf{P}$ as a vector by having [45]:

$$\mathbf{v}_{\mathbf{P}_\mathrm{a}} = \mathrm{vec}(\mathbf{P}) \tag{2.64}$$

   where $\mathrm{vec}(\mathbf{P})$ is a vector of complex dimension M^2 formed by stacking up the columns of $\mathbf{P}$ upon each other.

2. Since $\mathbf{P}$ is Hermitian symmetric, its ijth element is simply the complex conjugate of the jith element. Therefore, stacking up the elements renders the dimension of $\mathrm{vec}(\mathbf{P})$ unnecessarily high containing a lot of redundant information. To eliminate the redundancy, we can stack up all the elements of only the upper (or lower) triangular part of $\mathbf{P}$ forming a vector $\mathbf{v}_{\mathbf{P}_\mathrm{b}}$ having M^2 elements of real numbers [80].

3. Since $\mathbf{P}$ is Hermitian and is positive semi-definite, its eigenvalues $\{\lambda_m\}$ are real and positive semi-definite, and its eigenvectors $\{\mathbf{u}_m\}$ form an orthonormal set such that

$$\mathbf{P} = \sum_{m=1}^{M} \lambda_m \mathbf{u}_m \mathbf{u}_m^H; \qquad \langle \mathbf{u}_m, \mathbf{u}_n \rangle = \delta_{mn} = \begin{cases} 1 & m, \\ 0 & m \neq n \end{cases} \tag{2.65}$$

   where $(\cdot)^H$ denotes the Hermitian conjugate of a vector or matrix. Since the PSD matrix represents the average cross-power between signals from different

sensors, it can be conceived that a random vector that produce this averaged cross-power is given by

$$\mathbf{x} = \sum_{m=1}^{M} b_m \mathbf{u}_m \tag{2.66}$$

where b_m is a random scalar such that

$$\mathbb{E}[b_m b_n^*] = 0, \quad m \neq n; \qquad \mathbb{E}[|b_m|^2] = \lambda_m \tag{2.67}$$

The expansion of a random vector as a linear combination of the eigen-vectors of $\mathbf{P}$ with orthogonal random coefficients as in Eq. (2.66) is called the Karhunen-Loève expansion [72]. Since the vector $\mathbf{x}$ has average power λ_m in its component $\mathbf{u}_m$, therefore, it is reasonable to represent the matrix $\mathbf{P}$ as a *power vector* having power components $\{\lambda_m\}$ in a linear combination of the eigenvectors such that

$$\mathbf{v}_{\mathbf{P}_\mathrm{c}} = \sum_{m=1}^{M} \lambda_m \mathbf{u}_m \tag{2.68}$$

We note that for a Hermitian matrix $\mathbf{P}$, the inner products of $\mathbf{v}_{\mathbf{P}_\mathrm{a}}$ and $\mathbf{v}_{\mathbf{P}_\mathrm{c}}$ are identical since

$$\langle \mathbf{v}_{\mathbf{P}_\mathrm{a}}, \mathbf{v}_{\mathbf{P}_\mathrm{a}} \rangle = \sum_{m=1}^{M} |p_{ij}|^2 = \mathrm{Tr}[\mathbf{P}\mathbf{P}^H] = \sum_{m=1}^{M} \lambda_m^2 \tag{2.69a}$$

$$\langle \mathbf{v}_{\mathbf{P}_\mathrm{c}}, \mathbf{v}_{\mathbf{P}_\mathrm{c}} \rangle = \sum_{m=1}^{M} \lambda_m^2 \tag{2.69b}$$

None of the above representations of $\mathbf{P}$ as a vector is particularly satisfactory because some of the important structures, e.g., Hermitian symmetry, of the PSD matrix are not preserved. We now take a new look at the representation of PSD matrices.

2.5.3 Representation of PSD matrices in a linear space – Lie algebra

Here, we introduce a method to represent each $M \times M$ PSD matrix by a vector so that the vector representation preserves the properties of PSD matrices. This can be accomplished by building a one-to-one correspondence between $\mathcal{M}$ and another metric space $\mathcal{N}$ such that each point in $\mathcal{M}$ has a unique vector representation in $\mathcal{N}$ and vice versa. Then we can compare two points in $\mathcal{M}$ by comparing their correspondence in $\mathcal{N}$, i.e., we define the Euclidean distance between two vectors in $\mathcal{N}$ as the distance between the corresponding PSD matrices in $\mathcal{M}$.

We now show that the space $\mathcal{N}$ exists and is in fact a product space of χ and $\tilde{\mathcal{U}}$, i.e., $\mathcal{N} = \chi \times \tilde{\mathcal{U}}$, where χ is the space of all M-dimensional positive vectors (i.e., each element of the vector is positive) and $\tilde{\mathcal{U}}$ is the space spanned by a set of basis matrices $\{\mathbf{E}_i\}_{i=1}^{M^2-1}$. Let the eigen decomposition of an $M \times M$ PSD matrix $\mathbf{P}$ be

$$\mathbf{P} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1} \tag{2.70}$$

where $\mathbf{\Lambda} = \text{diag}[\lambda_1, \cdots, \lambda_M]$. Let

$$\mathbf{U} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^{-1} \tag{2.71}$$

be the eigen decomposition of the eigenvector matrix $\mathbf{U}$ with $\mathbf{\Sigma} = \text{diag}[\sigma_1, \cdots, \sigma_M]$. Since $\mathbf{U}$ is unitary, we can rewrite the matrix $\mathbf{\Sigma}$ as

$$\mathbf{\Sigma} = \begin{bmatrix} e^{j\theta_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{j\theta_M} \end{bmatrix} \tag{2.72}$$

where the real number θ_m is the phase of σ_m, $m = 1, \cdots, M$, i.e., the eigenvalues of

$\mathbf{U}$ are all of modulus unity [2]. We form a matrix $\mathbf{\Theta}$ such that

$$\mathbf{\Theta} = \begin{bmatrix} \theta_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \theta_M \end{bmatrix} \tag{2.73}$$

Then, we have

$$\begin{aligned} e^{j\mathbf{\Theta}} &= \mathbf{I} + j\mathbf{\Theta} + \frac{(j\mathbf{\Theta})^2}{2!} + \cdots \\ &= \begin{bmatrix} 1 + j\theta_1 + \frac{(j\theta_1)^2}{2!} + \cdots & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 + j\theta_M + \frac{(j\theta_M)^2}{2!} + \cdots \end{bmatrix} \\ &= \begin{bmatrix} e^{j\theta_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{j\theta_M} \end{bmatrix} = \mathbf{\Sigma} \end{aligned} \tag{2.74}$$

Now, we create another matrix $\tilde{\mathbf{U}}$ using the eigenvectors $\mathbf{U}$ and eigenvalues θ_1 so that

$$\tilde{\mathbf{U}} = \mathbf{V}\mathbf{\Theta}\mathbf{V}^{-1}, \tag{2.75}$$

Then, since $\mathbf{V}$ is unitary, we have

$$\begin{aligned} \mathbf{U} &= \mathbf{V}\mathbf{\Sigma}\mathbf{V}^{-1} = \mathbf{V}e^{j\mathbf{\Theta}}\mathbf{V}^{-1} \\ &= \mathbf{I} + j\mathbf{V}(\mathbf{\Theta})\mathbf{V}^{-1} + \frac{(j\mathbf{V}\mathbf{\Theta}\mathbf{V}^{-1})^2}{2!} + \cdots \\ &= e^{j\mathbf{V}\mathbf{\Theta}\mathbf{V}^{-1}} = e^{j\tilde{\mathbf{U}}}. \end{aligned} \tag{2.76}$$

Therefore, the matrices $\mathbf{U}$ and $\tilde{\mathbf{U}}$ can be directly related with each other. Since $\mathbf{U}$ is unitary, we have

$$\mathbf{U}^H\mathbf{U} = (e^{j\tilde{\mathbf{U}}})^H e^{j\tilde{\mathbf{U}}} = e^{(j\tilde{\mathbf{U}})^H} e^{j\tilde{\mathbf{U}}} = e^{(j\tilde{\mathbf{U}})^H + j\tilde{\mathbf{U}}} = \mathbf{I}. \tag{2.77}$$

[2] Let $\mathbf{U}\mathbf{u} = \sigma\mathbf{u}$. Then, we have $1 = (\mathbf{U}\mathbf{u})^H(\mathbf{U}\mathbf{u}) = \sigma^*\sigma\mathbf{u}^H\mathbf{u} = |\sigma|^2$ since the eigenvectors of $\mathbf{U}$ are orthonormal. Thus, the modulus of σ is 1, i.e., $|\sigma| = 1$

Thus, $(j\tilde{\mathbf{U}})^H + j\tilde{\mathbf{U}} = \mathbf{0}$, i.e., $j\tilde{\mathbf{U}}$ is skew-Hermitian. Now,

$$(j\tilde{\mathbf{U}})^H = -j\tilde{\mathbf{U}}^H = -j\tilde{\mathbf{U}} \tag{2.78}$$

Thus, $\tilde{\mathbf{U}}^H = \tilde{\mathbf{U}}$. On the other hand, since $\tilde{\mathbf{U}}$ is Hermitian we have

$$\mathrm{Tr}[(j\tilde{\mathbf{U}})^H - j\tilde{\mathbf{U}}] = \mathrm{Tr}[-j\tilde{\mathbf{U}}^H - j\tilde{\mathbf{U}}] = -2j\mathrm{Tr}\tilde{\mathbf{U}} \tag{2.79}$$

and

$$\mathrm{Tr}[(j\tilde{\mathbf{U}})^H - j\tilde{\mathbf{U}}] = -2\Im[\mathrm{Tr}j\tilde{\mathbf{U}}] = -2\Re[\mathrm{Tr}\tilde{\mathbf{U}}] = -2\mathrm{Tr}\tilde{\mathbf{U}} \tag{2.80}$$

where $\Re(\cdot)$ and $\Im(\cdot)$ denote the real part and imaginary part respectively. Therefore we should have $2\mathrm{Tr}\tilde{\mathbf{U}} = 2j\mathrm{Tr}\tilde{\mathbf{U}}$, which implies $\mathrm{Tr}\tilde{\mathbf{U}} = 0$. Therefore, $\tilde{\mathbf{U}}$ is Hermitian and null trace. In mathematics we say that $\tilde{\mathbf{U}}$ belongs to the *Lie algebra* of the group of unitary matrices with unit determinant (see Appendix B for definition of Lie algebra and Lie group).There are methods to construct a matrix-form basis $\{\mathbf{E}_i\}_{i=1}^{M^2-1}$ for the Lie algebra of the group of unitary matrices with unit determinant [31] [26]. For example, if $M = 4$, then $\mathbf{E}_i$ can be chosen as the modified Dirac matrices with $M^2 - 1 = 15$ as shown in Table 2.1.

Note that the modified Dirac matrices satisfy the following properties [83]

(1) $\mathbf{E}_i\mathbf{E}_k + \mathbf{E}_k\mathbf{E}_i = 2\delta_{ik}\mathbf{I}$.

(2) $\mathbf{E}_i^2 = \mathbf{I}$.

(3) $|\mathbf{E}_i| = 1$.

(4) $\mathbf{E}_i = \mathbf{E}_i^H$.

(5) $Tr\mathbf{E}_i = 0$.

(6) $\{\mathbf{E}_i\}$ are linearly independent.

Table 2.1: Modified Dirac matrices

$\mathbf{E}_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & j \\ 0 & 0 & -j & 0 \end{bmatrix}$	$\mathbf{E}_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -j \\ 1 & 0 & 0 & 0 \\ 0 & j & 0 & 0 \end{bmatrix}$	$\mathbf{E}_3 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & j & 0 \\ 0 & -j & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$
$\mathbf{E}_4 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -j \\ 0 & 0 & j & 0 \end{bmatrix}$	$\mathbf{E}_5 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$	$\mathbf{E}_6 = \begin{bmatrix} 0 & 0 & 0 & -j \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ j & 0 & 0 & 0 \end{bmatrix}$
$\mathbf{E}_7 = \begin{bmatrix} 0 & 0 & j & 0 \\ 0 & 0 & 0 & 1 \\ -j & 0 & 0 & j \\ 0 & 1 & 0 & 0 \end{bmatrix}$	$\mathbf{E}_8 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & j \\ 1 & 0 & 0 & 0 \\ 0 & -j & 0 & 0 \end{bmatrix}$	$\mathbf{E}_9 = \begin{bmatrix} 0 & 0 & 0 & j \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ -j & 0 & 0 & 0 \end{bmatrix}$
$\mathbf{E}_{10} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$	$\mathbf{E}_{11} = \begin{bmatrix} 0 & -j & 0 & 0 \\ j & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$	$\mathbf{E}_{12} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -j & 0 \\ 0 & j & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$
$\mathbf{E}_{13} = \begin{bmatrix} 0 & 0 & -j & 0 \\ 0 & 0 & 0 & 1 \\ j & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$	$\mathbf{E}_{14} = \begin{bmatrix} 0 & j & 0 & 0 \\ -j & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$	$\mathbf{E}_{15} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

We note that $\tilde{\mathbf{U}}$ can be expanded as a linear combination of the basis matrices such that

$$\tilde{\mathbf{U}} = \sum_{i=1}^{M^2-1} \alpha_i \mathbf{E}_i \tag{2.81}$$

where

$$\alpha_i = \frac{1}{4} Tr \tilde{\mathbf{U}} \mathbf{E}_i. \tag{2.82}$$

That the dimension of $\tilde{\mathbf{U}}$ is $M^2 - 1$ in Eq. (2.81) is clear from the fact that there are M real parameters on its diagonal, $\frac{1}{2}M(M-1)$ complex parameters on each of the lower and upper triangular part of the matrix (being conjugates), and the trace of the matrix is zero. Thus, the total degrees of freedom for the matrix is $\{M + 2 \times \frac{1}{2}M(M-1) - 1\} = (M^2 - 1)$. That all α_i are real can be easily seen from taking the Hermitian conjugate of Eq. (2.81) and using Hermitian properties of $\tilde{\mathbf{U}}$ and $\mathbf{E}_i$, we have

$$\tilde{\mathbf{U}} = \sum_{i=1}^{M^2-1} \alpha_i^* \mathbf{E}_i \tag{2.83}$$

Subtracting Eq. (2.83) from Eq. (2.81), we have

$$\sum_{i=1}^{M^2-1} (\alpha - \alpha_i^*) \mathbf{E}_i = 0 \tag{2.84}$$

Since $\mathbf{E}_i$ are all linearly independent, then, $\alpha = \alpha_i^*$.

From the linear combination in Eqs. (2.70) and (2.81), we see that we can represent $\mathbf{P}$ as a vector such that

$$\mathbf{v}_{\mathbf{P}_\mathrm{L}} = [\boldsymbol{\lambda}, \boldsymbol{\alpha}]^T, \tag{2.85}$$

where $\boldsymbol{\lambda} = [\lambda_1, \cdots, \lambda_M]$ and $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_{M^2-1}]$. This representation of $\mathbf{P}$ is designated *Lie vector* in this thesis and is illustrated in Fig. 2.5.

It can be seen that the vector $\boldsymbol{\lambda}$ is the representation of $\mathbf{P}$ in the eigen space (spanned by the eigenvectors of $\mathbf{P}$) and the vector $\boldsymbol{\alpha}$ is the representation of the eigen space in the space spanned by the basis $\{\mathbf{E}_i\}$. In other words, we can represent

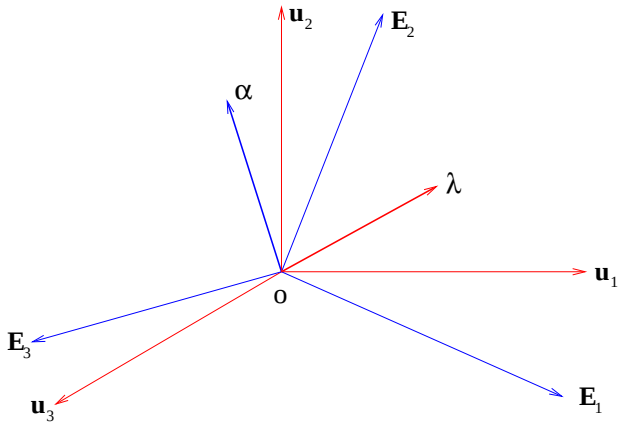


Figure 2.5: Illustrate of vector representation

the PSD matrix $\mathbf{P}$ by a vector $\mathbf{v}_{\mathbf{P}_\mathrm{L}} = [\boldsymbol{\lambda}, \boldsymbol{\alpha}]^T$ and characterize completely the $\mathbf{P}$ in the sense that the eigenvalues $\boldsymbol{\lambda}$ represent the structure of the eigenvectors and the parameters $\boldsymbol{\alpha}$ characterize the structure of the unitary matrix $\mathbf{U}$. Furthermore, $\boldsymbol{\alpha}$ are invariant under complete unitary transformations and therefore are true invariant descriptors of the system. For all power spectral density matrices, $\boldsymbol{\alpha}$ are located in the same space which makes the comparison between two power spectral density matrices in terms of their vector representations reasonable. In the sequel we adopt this representation as our vector representation of power spectral density matrices.

Let $\mathbf{v}_{\mathbf{P}_\mathrm{L}1} = [\boldsymbol{\lambda}_1, \boldsymbol{\alpha}_1]$ and $\mathbf{v}_{\mathbf{P}_\mathrm{L}2} = [\boldsymbol{\lambda}_2, \boldsymbol{\alpha}_2]$ be the vector representations of $\mathbf{P}_1$ and $\mathbf{P}_2$. Then, the distance between $\mathbf{P}_1$ and $\mathbf{P}_2$ can be defined as

$$d(\mathbf{P}_1, \mathbf{P}_2) = d(\mathbf{v}_{\mathbf{P}_\mathrm{L}1}, \mathbf{v}_{\mathbf{P}_\mathrm{L}2}) = \sqrt{\sum_{m=1}^{M} (\lambda_{1m} - \lambda_{2m})^2 + \sum_{\ell=1}^{M^2-1} (\alpha_{1\ell} - \alpha_{2\ell})^2} \tag{2.86}$$

2.6 Vector space of Hermitian matrices and manifold of PSD matrices

Since the PSD matrices of the EEG signal has been chosen to be the feature for classification and PSD matrices are a subset of Hermitian matrices, it is imperative for us to examine the structure of the vector space of these matrices. Let us first examine the vector space of Hermitian matrices.

Let $\mathcal{M}_M$ be the set of all the $M \times M$ complex matrices. Let $\mathcal{H}_H$ and $\mathcal{M}$ denote respectively the set of all Hermitian matrices and the set of positive definite Hermitian matrices, i.e.,

$$\mathcal{H}_H = \{\mathbf{A} \in \mathcal{M}_M : \mathbf{A}^H = \mathbf{A}\} \tag{2.87a}$$

$$\mathcal{M} = \{\mathbf{P} \in \mathcal{H}_H : \mathbf{P} \succ 0\} \tag{2.87b}$$

Thus, we have $\mathcal{M} \subset \mathcal{H}_H \subset \mathcal{M}_M$. We have the following proposition:

Proposition 2.1 *$\mathcal{H}_H$ is a real linear vector space. It is isomorphic to the Euclidean space $\mathbb{R}^{M \times M}$.*

Proof: We note that for an $M \times M$ Hermitian matrix $\mathbf{H}$ and a complex scalar c, $c\mathbf{H} \notin \mathcal{H}_H$ in general since $c\mathbf{H}$ may no longer be Hermitian. Therefore, $\mathcal{H}_H$ is closed only for real scalar field. Furthermore, $\mathbf{H}$ can be represented as a linear combination of a set of basis Hermitian matrices $\{\tilde{\mathbf{E}}_{ij};\ i, j = 1, \cdots, M\}$ with all coefficients being real such that

$$\mathbf{H} = \sum_{m=1}^{M} \sum_{n=1}^{M} h_{mn} \tilde{\mathbf{E}}_{mn} \tag{2.88}$$

where $\mathbf{H} = \mathbf{H}^H$, $\tilde{\mathbf{E}}_{ij} = \tilde{\mathbf{E}}_{ij}^H$ and $h_{ij} = h_{ij}^*$. This is because the matrix $\mathbf{H}$ is Hermitian, therefore there are M real elements on the diagonal and $\frac{1}{2}M(M-1)$ complex elements above and below the diagonal, respectively which are complex conjugates. Thus, the

number of real degrees of freedom is $M+\frac{1}{2}M(M-1)+\frac{1}{2}M(M-1)=M^2$. Therefore, the total degrees of freedom for the set of Hermitian matrices is M^2 and thus, there exist M^2 linearly independent and orthonormal Hermitian matrices $\tilde{\mathbf{E}}_{11},\cdots,\tilde{\mathbf{E}}_{MM}$ forming the basis of the space resulting in the linear combination of Eq. (2.88). That the coefficients h_{ij} in Eq. (2.88) are all real follows exactly the same argument as those for Eq. (2.81). These orthonormal Hermitian basis matrices can be obtained by having

$$\tilde{\mathbf{E}}_{mm}=\begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1_{(mm)} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \tag{2.89a}$$

$$\tilde{\mathbf{E}}_{mn}=\begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 1_{(mn)} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1_{(nm)} & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \end{bmatrix} \tag{2.89b}$$

$$\tilde{\mathbf{E}}_{nm}=\begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & j_{(mn)} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & -j_{(nm)} & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \end{bmatrix} \tag{2.89c}$$

where the subscripts in parentheses denote the positions of the non-zero elements in the $M \times M$ matrix. The linear combination of Eq. (2.88) with real coefficients is a reminiscence of the representation of a vector as an n-tuple in $\mathbb{R}^n$ as shown in Eq. (2.56), i.e., here we can represent an $M \times M$ Hermitian matrix as an $(M \times M)$-tuple $\{h_{mn}; m, n = 1, \cdots, M\}$ in a *real* $(M \times M)$-dimensional space $\mathbb{R}^{M \times M}$. Since the basis matrices $\{\tilde{\mathbf{E}}_{mn}\}$ in Eq. (2.88) are all orthonormal, the inner product $\langle \mathbf{H}_1, \mathbf{H}_2 \rangle$ in $\mathcal{H}_H$ is also real since

$$\langle \mathbf{H}_1, \mathbf{H}_2 \rangle = \sum_{i=1}^{M} \sum_{j=1}^{M} h_{1ij} h_{2ij} \tag{2.90}$$

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

To show that $\mathcal{H}_H$ and $\mathbb{R}^{M \times M}$ are isomorphic, we need to find a mapping $\phi : \mathcal{H}_H \rightarrow \mathbb{R}^{M \times M}$ such that

$$\phi(\mathbf{H}_1 + \mathbf{H}_2) = \phi(\mathbf{H}_1) + \phi(\mathbf{H}_2) \tag{2.91a}$$

$$\phi(k\mathbf{H}) = k\phi(\mathbf{H}) \tag{2.91b}$$

where k is a real number, and $\mathbf{H}, \mathbf{H}_1, \mathbf{H}_2 \in \mathcal{H}_H$. If we let

$$\phi(\mathbf{H}) = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1M} \\ h_{21} & h_{22} & \cdots & h_{2M} \\ \vdots & \vdots & & \vdots \\ h_{M1} & h_{M2} & \cdots & h_{MM} \end{bmatrix} \tag{2.92}$$

then, it is easy to see that the mapping ϕ satisfy Eqs. (2.91). This shows that $\mathcal{H}_H$ and $\mathbb{R}^{M \times M}$ are isomorphic, denoted by $\mathcal{H}_H \cong \mathbb{R}^{M \times M}$. □

Let us now consider the PSD matrices which are the features for EEG signal classification and are *positive definite* Hermitian. Therefore, we can likewise represent the PSD matrices as a linear combination with real coefficients $\{p_{mn}\}$ such that

$$\mathbf{P} = \sum_{m=1}^{M} \sum_{n=1}^{M} p_{mn} \tilde{\mathbf{E}}_{mn} \tag{2.93}$$

However, we cannot find a subspace in $\mathbb{R}^{M\times M}$ in which an n-tuple representation for these positive definite PSD matrices can be defined. Now, the set of PSD matrices is a subset of the set of all the Hermitian matrices, i.e.,

$$\mathbf{P} \in \mathcal{M} \subset \mathcal{H}_H \tag{2.94}$$

Therefore, we may conceive that the PSD matrices form a *manifold* [3] $\mathcal{M}$ in $\mathcal{H}_H$, the space of all Hermitian matrices. We now show that the manifold described by the PSD matrices is real:

Lemma 2.2 *[39] The exponential mapping*

$$e^{\mathbf{A}} : \mathcal{H}_H \to \mathcal{M} \tag{2.95}$$

is a bijection. In other words, if $\mathbf{A} \in \mathcal{H}_H$*, then* $e^{\mathbf{A}} \in \mathcal{M}$*; if* $\mathbf{P} \in \mathcal{M}$*, then there exists a unique* $\mathbf{A} \in \mathcal{H}_H$ *such that* $\mathbf{P} = e^{\mathbf{A}}$*.*

Theorem 2.1 $\mathcal{M}$ *is a real manifold.*

Proof. Due to Lemma 2.2, for any matrix $\mathbf{X} \in \mathcal{H}_H$ we can define a map $\mathbb{R} \to \mathcal{M}$ by

$$f(t) = e^{t\mathbf{X}+\mathbf{A}} \tag{2.96}$$

such that $f(0) = e^{\mathbf{A}} = \mathbf{P} \in \mathcal{M}$ and $\mathbf{X}, \mathbf{A} \in \mathcal{H}_H$, i.e., $f(t)$ is a path on $\mathcal{M}$ through $\mathbf{P}$. Since $t\mathbf{X}$ and $\mathbf{A}$ are Hermitian matrices, we have

$$f(t) = e^{t\mathbf{X}+\mathbf{A}} = e^{\mathbf{A}+t\mathbf{X}} = e^{t\mathbf{X}}e^{\mathbf{A}} = e^{\mathbf{A}}e^{t\mathbf{X}} \tag{2.97}$$

[3] For now, a manifold can be looked upon as [35]: "An n-dimensional manifold is a space which is not necessarily a Euclidean space nor is it a domain in a Euclidean space, but which, from the viewpoint of a short-sighted observer living in the space, looks just like such a domain of Euclidean space."

Let $\mathcal{T}_{\mathcal{M}}(\mathbf{P})$ be the tangent space formed by the set of vectors which are tangent to $\mathcal{M}$ at $\mathbf{P}$. Then, the derivative of $f(t)$ at $t = 0$ is

$$\dot{f}(t)\Big|_{t=0} = \mathbf{X}e^{t\mathbf{X}}\Big|_{t=0} e^{\mathbf{A}} = e^{\mathbf{A}}e^{t\mathbf{X}}\mathbf{X}\Big|_{t=0} \in \mathcal{T}_{\mathcal{M}}(\mathbf{P}) \tag{2.98}$$

Thus, we have

$$\dot{f}(t)\Big|_{t=0} = \mathbf{X}\mathbf{P} = \mathbf{P}\mathbf{X} \in \mathcal{T}_{\mathcal{M}}(\mathbf{P}) \tag{2.99}$$

i.e., any element of $\mathcal{T}_{\mathcal{M}}(\mathbf{P})$ is the product of $\mathbf{P}$ and any Hermitian matrix $\mathbf{X}$, and the product is commutative. Since $\mathbf{P}^{-1} \in \mathcal{M}$, we can also write

$$\mathbf{X}\mathbf{P}^{-1} = \mathbf{P}^{-1}\mathbf{X} \tag{2.100}$$

Let $\mathbf{Y} = \mathbf{P}^{-1}\mathbf{X}$. Then, using Eq. (2.100) we have

$$\mathbf{Y}^H = (\mathbf{P}^{-1}\mathbf{X})^H = \mathbf{X}\mathbf{P}^{-1} = \mathbf{P}^{-1}\mathbf{X} = \mathbf{Y} \tag{2.101}$$

i.e., $\mathbf{Y} \in \mathcal{H}_H$. Therefore, $\mathbf{P}\mathbf{Y} = \mathbf{X} \in \mathcal{T}_{\mathcal{M}}(\mathbf{P})$. Thus, we have $\mathcal{H}_H \subseteq \mathcal{T}_{\mathcal{M}}(\mathbf{P})$.
On the other hand, for any $\mathbf{X}\mathbf{P} \in \mathcal{T}_{\mathcal{M}}(\mathbf{P})$ we have

$$(\mathbf{X}\mathbf{P})^H = \mathbf{P}^H\mathbf{X}^H = \mathbf{P}\mathbf{X} = \mathbf{X}\mathbf{P} \tag{2.102}$$

i.e., $\mathbf{X}\mathbf{P} \in \mathcal{H}_H$. Thus, $\mathcal{T}_{\mathcal{M}}(\mathbf{P}) \subseteq \mathcal{H}_H$.

Therefore, we conclude that the tangent space of $\mathcal{M}$ at $\mathbf{P}$, $\mathcal{T}_{\mathcal{M}}(\mathbf{P}) = \mathcal{H}_H$, i.e., it contains all the Hermitian matrices. By Proposition 2.1, we then have $\mathcal{T}_{\mathcal{M}}(\mathbf{P}) \cong \mathbb{R}^{M\times M}$. Therefore, $\mathcal{M}$ is a *real* manifold. □

The result in Theorem 2.1 is important in the development of distance measures in the manifold of PSD matrices. This is because we only have to consider real analysis of the geometry.

In sleep classification, we characterize an epoch of multichannel EEG signal matrix $\mathbf{S}$ by its feature PSD matrices $\mathbf{P}(\omega)$ in a frequency range $[\omega_1, \omega_2]$, therefore, we can

regard the PSD matrix as represented by a series of points forming a curve on $\mathcal{M}$ parameterized by the frequency variable ω, i.e.,

$$\mathbf{P}(\omega): \qquad [\omega_1, \omega_2] \rightarrow \mathcal{M} \tag{2.103}$$

This concept is shown in Fig. 2.6.

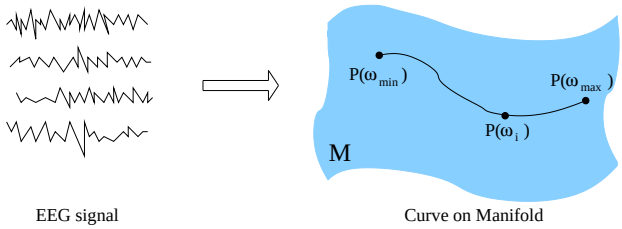


Figure 2.6: EEG signal representation

Chapter 3

Distance Measures for EEG Signal Classification

In the previous chapter, we have seen how the measured EEG signal can be collected and cleaned up, and how its PSD feature can be extracted and calculated. We have also seen that these PSD matrices can be represented as vectors and treated as points in a *linear space*, or the matrices themselves can be looked upon as points on a manifold in a linear space. For analysis of the these features, the linear space usually have certain geometric structures. These features of the collected EEG signals may then be used for classification purposes.

Now, EEG signal classification is a matter of examining the similarity/dissimilarity between the features of the signals. Similarity/dissimilarity can be quantified according to a specific measure which may not necessarily be a metric in the strict mathematical sense. The only requirement is that it quantifies the similarity or commonality between two EEG signals by taking on small values for two similar EEG signals and large values for two distinct EEG signals. However, since we have shown that the set of all the PSD matrices is a real manifold which is a mathematical space,

it is our opinion that the similarity/dissimilarity between EEG signals should take on a mathematical measure.

Intuitively two points in a mathematical space are similar if they are close to each other with respect to the metric endowed to the space, and vice verse. From the geometric point of view, the dissimilarity between two points in a space is naturally measured by some kind of *distance function*, or *distance* for short. The similarity can then be defined as a function of the dissimilarity. For example, one can define the similarity as the inverse a distance function. Therefore, it is not necessary to distinguish dissimilarity and distance. The appropriate measure of distance depends on the structure of the space. Here in this chapter, we will study the geometric structures of these linear spaces and the various metrics used to measure distances. In particular, we will examine the space of Hermitian matrices and the the Riemannian manifold in it formed by the PSD matrices of the EEG signals. From this, we will derive metrics on the Riemannian manifold and arrive at suitable distance measures suitable for the classification of EEG signals.

3.1 Distance measures in an n-dimensional inner product vector spaces

In the previous chapter, we have introduced the inner product vector space in which the distance between two vectors $\mathbf{x}$ and $\mathbf{y}$ is given by $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$. For two vectors $\mathbf{x}$ and $\mathbf{y}$ in the n-dimensional Euclidean space $\mathbb{C}^n$ in which the inner product is defined as

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i^* \tag{3.1}$$

We now present some common measures of distance induced by the inner product norm.[1]

3.1.1 Distance measures between vectors induced by the inner product

The following are commonly used distance measures in an n-dimensional inner product vector space:

1. Euclidean Distance This well-known distance measure mentioned in Eq. (1.1) is induced by the inner product norm such that:

$$d_{\mathrm{E}}(\mathbf{x},\mathbf{y}) = \sqrt{\langle(\mathbf{x}-\mathbf{y}),(\mathbf{x}-\mathbf{y})\rangle} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)}. \tag{3.2}$$

   This metric coincides with the usual concept of distance in a three-dimensional space and, due to the many important physical quantities it can represent, the Euclidean distance is a powerful measure used in the study of signals [36] [73].

2. Correlation Distance From Chapter 2, we see that the Cauchy-Schwarz inequality can be written as $\left|\langle\mathbf{x},\mathbf{y}\rangle\right| \leq \|\mathbf{x}\|\,\|\mathbf{y}\|$, we can define a real angle θ between $\mathbf{x}$ and $\mathbf{y}$ as

$$\cos\theta = \frac{|\langle\mathbf{x},\mathbf{y}\rangle|}{\|\mathbf{x}\|\,\|\mathbf{y}\|} \tag{3.3}$$

   We say that $\mathbf{x}$ and $\mathbf{y}$ are "orthogonal" if, and only if, $\langle\mathbf{x},\mathbf{y}\rangle = 0$ for which the distance between the two vectors is the greatest.[2] Thus, the angle between two

[1] There are other distance measures not induced by the inner product norm which are commonly found in engineering applications. These include the Minkowski distance defined as $d_{\mathrm{M}}(\mathbf{x},\mathbf{y}) = [\sum_{i=1}^{n}|x_i - y_i|^r]^{1/r}$ and the Chebyshev distance defined as $d_{\mathrm{Ch}}(\mathbf{x},\mathbf{y}) = \max_{i=1,\cdots,n}|x_i - y_i|$.

[2] The difficulty with (3.3) applied to complex space is apparent since we would not generate second- and third-quadrant angles. On the other hand, if we replace $\left|\langle\mathbf{x},\mathbf{y}\rangle\right|$ with $\mathrm{Re}\langle\mathbf{x},\mathbf{y}\rangle$, we could have $\theta = \pm\pi/2$ with $\langle\mathbf{x},\mathbf{y}\rangle \neq 0$.

vectors can be used as distance measures. For finite-dimensional real normalized vectors $\mathcal{X}$ such that $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 1,\ x_i > 0\}$, the *Fisher-Rao* distance [12] between $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, is defined as

$$d_{\mathrm{FR}}(\mathbf{x}, \mathbf{y}) = 2 \arccos\Big(\sum_{i=1}^{n} \sqrt{x_i y_i} \Big) \tag{3.4}$$

For two finite-dimensional complex vectors $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$ the *Fubini-Study* distance [57] between $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$d_{\mathrm{FS}}(\mathbf{x}, \mathbf{y}) = \arccos \frac{|\langle \mathbf{x}, \mathbf{y} \rangle|}{\sqrt{\langle \mathbf{x}, \mathbf{x} \rangle \langle \mathbf{y}, \mathbf{y} \rangle}} \tag{3.5}$$

We may also define a *Correlation* distance measure in terms of the angle between two vectors such that the smaller is the angle, the closer is the distance

$$d_{\mathrm{C}}(\mathbf{x}, \mathbf{y}) = 1 - \left(\frac{|\sum_{i=1}^{n} x_i y_i^*|}{\|\mathbf{x}\| \|\mathbf{y}\|} \right) \tag{3.6}$$

We note that the second term in Eq. (3.6) can refer to the argument in either $d_{\mathrm{FR}}(\mathbf{x}, \mathbf{y})$ or $d_{\mathrm{FS}}(\mathbf{x}, \mathbf{y})$ as the case of real or complex vectors may be.

Weighted Euclidean distance: Often in different applications, the Euclidean distance can also be *weighted* allowing certain parts of the signal to be accentuated. The weighted Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$ is defined as

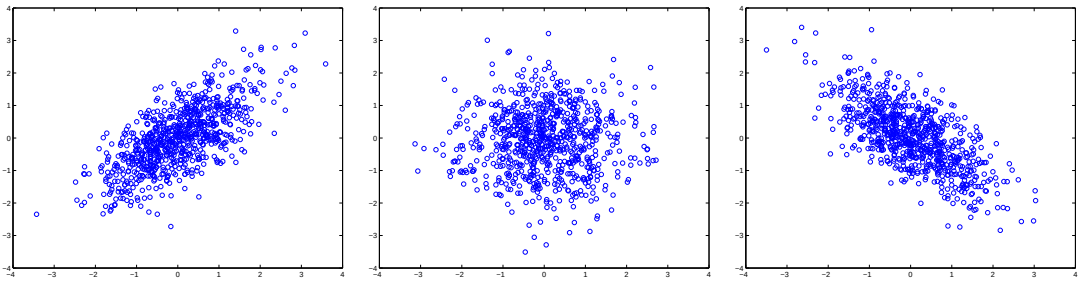
$$d_{\mathrm{WE}} = \Big\langle \mathbf{W}^{\frac{1}{2}}(\mathbf{x} - \mathbf{y}), \mathbf{W}^{\frac{1}{2}}(\mathbf{x} - \mathbf{y}) \Big\rangle^{\frac{1}{2}} = \sqrt{(\mathbf{x} - \mathbf{y})^H \mathbf{W} (\mathbf{x} - \mathbf{y})} \tag{3.7}$$

where $\mathbf{W}$ is a positive definite matrix. The choice of $\mathbf{W}$ depends on the data structure. Often, the best choice of the weighting matrix may be obtained by solving an optimization problem with a certain objective function and constraints. The following example shows how $\mathbf{W}$ is chosen according to the structure of the data:

Example: (Mahalanobis Distance) This distance is developed to fit the measurement of Gaussian data. Let $\mathbf{x}$ and $\mathbf{y}$ be two complex IID Gaussian vectors on dimension n with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Then the Mahalanobis distance is defined as

$$d_{\mathrm{M}}(\mathbf{x},\mathbf{y}) = \sqrt{(\det\boldsymbol{\Sigma})^{1/n}(\mathbf{x}-\mathbf{y})^{H}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{y})} \tag{3.8}$$

Comparing Eqs. (3.7) and (3.8), we note that the Mahalanobis distance is a weighted Euclidean distance with $\mathbf{W} = (\det\boldsymbol{\Sigma})^{1/n}\boldsymbol{\Sigma}^{-1}$ and is particularly suitable for measuring distances between Gaussian random vectors of the same distribution. This can be illustrated as follows: Suppose we have three sets of real zero-mean bivariate Gaussian data such that $\boldsymbol{\mu} = [0\ 0]^T$, $\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$, $\boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and $\boldsymbol{\Sigma}_3 = \begin{bmatrix} 1 & -0.7 \\ -0.7 & 1 \end{bmatrix}$. The data are distributed as shown in Figure 3.1 respectively.

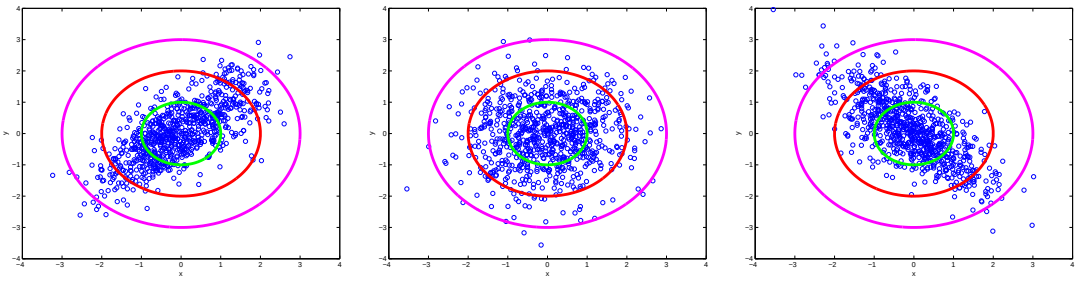


(a) μ, Σ_1 (b) μ, Σ_2 (c) μ, Σ_3

Figure 3.1: The geometries of bivariate normally distributed points with zero means and different covariance matrices.

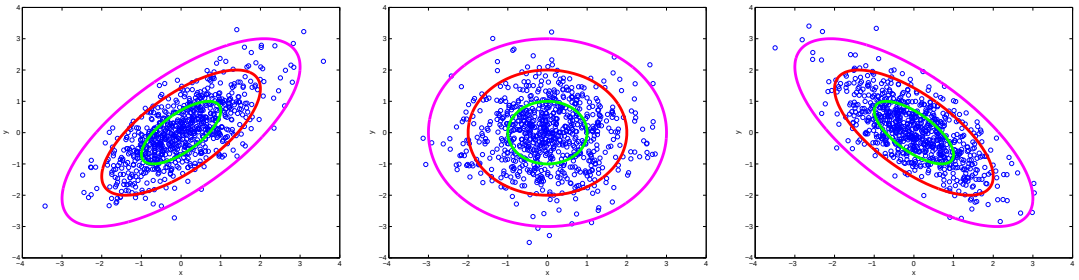
If unweighted Euclidean distance is adopted, as shown in Figure 3.2, the data distribution structure and the distance between two points relative to the distribution cannot be reflected in each of the cases. However, if Mahalanobis distance is

employed, the distances relative to the data distribution in each of the cases is completely characterized as shown in Figure 3.3.



(a) (b) (c)

Figure 3.2: Euclidean distances (solid circles) of 1, 2, and 3 from the origin, respectively.



(a) (b) (c)

Figure 3.3: Mahalanobis distances (solid circles and ellipses) of 1, 2, and 3 from the origin, respectively.

.

Here, it can be seen that points at a constant unweighted Euclidean distance from a reference point are located on the hypersphere (a circle in two dimensions), and points at a constant Mahalanobis distance to the center are located on a hyperellipsoid (an ellipse in two dimensions) following the distribution of the data. □

The above example uses the data distribution to arrive at different weighting matrices which provides us with a distance inversely proportional to the correlation of the data. This weighted Euclidean distance (Mahalanobis distance) fits very well to the Gaussian distributed data if the first- and second-order statistics are known. However, in many of the practical applications including our study of EEG signal classification, the data distribution and data structure may not be known. In that case, an optimization problem with appropriate constraints suitable to our problem may have to be defined and solved to arrive at a suitable weighting matrix. This particular problem will be considered in Chapter 4.

3.1.2 Distance measures between matrices

In Chapter 2, we have seen several ways of representing a matrix as a vector. Therefore, the above distance measures between vectors can all be applied as distance measures between matrices. For the space $\mathcal{M}$ of $M \times M$ matrices, a commonly used measure of distance between the matrices $\mathbf{A} = [a_{ij}]$ and $\mathbf{B} = [b_{ij}]$ is defined as

$$d_{\text{Fo}}(\mathbf{A}, \mathbf{B}) = \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{M} |a_{ij} - b_{ij}|^2} = \sqrt{\text{Tr}[(\mathbf{A} - \mathbf{B})(\mathbf{A} - \mathbf{B})^H]} \tag{3.9}$$

which is called the *Frobenius distance*. From Eq. (2.69a), we see that this distance is induced by the inner product of $\langle(\mathbf{A} - \mathbf{B}), (\mathbf{A} - \mathbf{B})\rangle$. Thus, the Frobenius distance can be considered as the Euclidean distance between $\mathbf{A}$ and $\mathbf{B}$ since $d_{\text{Fo}}(\mathbf{A}, \mathbf{B}) = d_{\text{E}}(\text{vec}(\mathbf{A}), \text{vec}(\mathbf{B}))$. The spaces $\mathcal{M}$ and $\mathbb{C}^n$ are thus *isometric*.

In general, if $\mathbf{v_P} \in \mathbb{C}^n$ is the vector representation of $\mathbf{P} \in \mathcal{M}$, then the Euclidean distance between $\mathbf{v_{P1}}$ and $\mathbf{v_{P2}}$ in $\mathbb{C}^n$ is the *induced Euclidean distance* between $\mathbf{P}_1$ and $\mathbf{P}_2$ in $\mathcal{M}$, i.e.,

$$d_{\text{IE}}(\mathbf{P}_1, \mathbf{P}_2) = d_{\text{E}}(\mathbf{v}_{\mathbf{P}_1}, \mathbf{v}_{\mathbf{P}_2}) \tag{3.10}$$

In the same way as for vectors, we can define the weighted distance between two matrices since we have converted the distance between two matrices as the distance between two vectors. For example, for the Frobenius distance of Eq. (3.9) between two $M \times M$ matrices, similar to Eq. (3.7), we can attach a weighting matrix $\mathbf{W}$ to the distance so that

$$d_{\text{WFo}}(\mathbf{A}, \mathbf{B}) = \sqrt{[(\mathbf{A} - \mathbf{B})\mathbf{W}(\mathbf{A} - \mathbf{B})^H]} \tag{3.11}$$

3.2 Some other interesting distances

In this section, we introduce some other distance measures which may be of interest to the application of EEG signal classification even though they are not induced by the inner product.

3.2.1 Fréchet distance

The Fréchet distance is defined for the measurement between two probability distributions. Specifically, for two random vectors $\mathbf{x}$ and $\mathbf{y}$ have distributions f and g respectively, the Fréchet distance is defined as [37]

$$d(f, g) = \sqrt{\min_{\mathbf{x},\mathbf{y}} \mathbb{E}[\| \mathbf{x} - \mathbf{y} \|^2]} \tag{3.12}$$

For two Gaussian vectors $\mathbf{x}$ and $\mathbf{y}$ having means $\boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\mu}_\mathbf{y}$ and covariance matrices $\mathbf{R}_\mathbf{x}, \mathbf{R}_\mathbf{y}$ respectively, it can be shown [**?**] that the Fréchet distance is given by

$$d(f, g) = \sqrt{\|\boldsymbol{\mu}_\mathbf{x} - \boldsymbol{\mu}_\mathbf{y}\|^2 + Tr[\mathbf{R}_\mathbf{x} + \mathbf{R}_\mathbf{y} - 2(\mathbf{R}_\mathbf{x}\mathbf{R}_\mathbf{y})^{1/2}]} \tag{3.13}$$

In the case of zero-mean Gaussian vectors, then the Fréchet distance between the distributions becomes the Fréchet distance between the covariances $\mathbf{R}_\mathbf{x}$ and $\mathbf{R}_\mathbf{y}$ such that

$$d(\mathbf{R}_\mathbf{x}, \mathbf{R}_\mathbf{y}) = \sqrt{Tr[\mathbf{R}_\mathbf{x} + \mathbf{R}_\mathbf{y} - 2(\mathbf{R}_\mathbf{x}\mathbf{R}_\mathbf{y})^{1/2}]} \tag{3.14}$$

As shown in the previous chapter, for EEG classification, our choice of the signal feature is the PSD matrices which has been shown to be the Fourier transform of the covariance matrix. Therefore, the distance in Eq. (3.14) will be of interest to our application in EEG classification.

3.2.2 Kullback-Leibler (KL) divergence and KL distance

For probability density functions $f_1(x)$ and $f_2(x)$, the KL divergence, also known as the relative entropy, is defined as [59]

$$D_{KL}(f_1||f_2) = \int f_1(x) \log \frac{f_1(x)}{f_2(x)} dx \tag{3.15}$$

The KL divergence is commonly used in statistics as a measure of similarity between two distributions. It satisfies the following properties

(a) Self-similarity: $D_{KL}(f||f) = 0$;

(b) Self-identification: $D_{KL}(f_1||f_2) = 0$ if $f_1 = f_2$;

(c) Positivity: $D_{KL}(f_1||f_2) \geq 0 \qquad \forall \;\; f_1$ and f_2.

Properties (a) and (b) are obvious. Property (c) can be shown by letting $\phi(f) = -\log \frac{f_2(x)}{f_1(x)}$, which is convex. Thus, by Jensen's inequality [47], $\mathbb{E}[\phi(f)] \geq \phi[\mathbb{E}(f)]$, i.e., $\int f_1(x) \log \frac{f_1(x)}{f_2(x)} dx \geq -\log \int f_1(x) \frac{f_2(x)}{f_1(x)} dx = 0$.

For two Gaussian vectors $\mathbf{x}$ and $\mathbf{y}$ having means $\boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\mu}_\mathbf{y}$ and covariance matrices $\mathbf{R}_\mathbf{x}, \mathbf{R}_\mathbf{y}$ respectively, the KL divergence is found to be [58]

$$D_{KL}((\boldsymbol{\mu}_1, \mathbf{R}_1), (\boldsymbol{\mu}_2, \mathbf{R}_2)) = \frac{1}{2}\Big[\log \frac{\det(\mathbf{R}_2)}{\det(\mathbf{R}_1)} + \mathrm{Tr}(\mathbf{R}_2^{-1}\mathbf{R}_1) + \mathrm{Tr}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{R}_2^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - M \Big] \tag{3.16}$$

where M is the dimension of the two random vectors. In the case of the means being zero, this simplifies to

$$D_{KL}(\mathbf{R}_1, \mathbf{R}_2) = \frac{1}{2}\Big[\log \frac{\det(\mathbf{R}_2)}{\det(\mathbf{R}_1)} + \mathrm{Tr}(\mathbf{R}_2^{-1}\mathbf{R}_1) - M \Big] \tag{3.17}$$

The KL divergence does not define a distance on the space of covariance matrices as it is neither symmetric with respect to its two arguments nor does it satisfy the triangle inequality. Its symmetrized form (also called *symmetrized divergence*) is defined as

$$D_{KLs}(\mathbf{R}_1, \mathbf{R}_2) = D_{KL}(\mathbf{R}_1, \mathbf{R}_2) + D_{KL}(\mathbf{R}_2, \mathbf{R}_1) \tag{3.18}$$

which can be expressed as

$$D_{KLs}(\mathbf{R}_1, \mathbf{R}_2) = \frac{1}{2}\mathrm{Tr}(\mathbf{R}_1\mathbf{R}_2^{-1} + \mathbf{R}_1^{-1}\mathbf{R}_2 - 2\mathbf{I}) \tag{3.19}$$

We define the *KL distance* between two covariance matrices as

$$d_{KL}(\mathbf{R}_1, \mathbf{R}_2) = \sqrt{D_{KLs}(\mathbf{R}_1, \mathbf{R}_2)} \tag{3.20}$$

Eq. (3.20) does not satisfy the axioms of distance measure [3] since the triangular inequality is still not satisfied. In spite of this, $d_{KL}(\mathbf{R}_1, \mathbf{R}_2)$ is still called the "KL distance" by convention.

As mentioned before, the PSD matrix is the Fourier transform of the covariance matrix. Therefore, we may replace $\mathbf{R}_1$ and $\mathbf{R}_2$ with $\mathbf{P}_1$ and $\mathbf{P}_2$ and define the KL distance between two PSD matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ such that

$$d_{KL}(\mathbf{P}_1, \mathbf{P}_2) = \sqrt{D_{KLs}(\mathbf{P}_1, \mathbf{P}_2)} = \sqrt{\frac{1}{2}\mathrm{Tr}(\mathbf{P}_1\mathbf{P}_2^{-1} + \mathbf{P}_1^{-1}\mathbf{P}_2 - 2\mathbf{I})} \tag{3.21}$$

Clearly, Eq. (3.21) can also be used for measuring the similarity and dissimilarity of EEG signal features.

We note that d_{KL} is weighting invariant. This is because, for a given $\mathbf{W} = \mathbf{\Omega}\mathbf{\Omega}^H \succ 0$ and weighted $\mathbf{P}_1$ and $\mathbf{P}_2$, $\mathbf{P}_{1W} = \mathbf{\Omega}^H\mathbf{P}_1\mathbf{\Omega}$ and $\mathbf{P}_{2W} = \mathbf{\Omega}^H\mathbf{P}_2\mathbf{\Omega}$, we have

$$\begin{aligned} d_{KL}(\mathbf{P}_{1W}, \mathbf{P}_{2W}) &= \sqrt{\frac{1}{2}\mathrm{Tr}[(\mathbf{\Omega}^{\mathrm{H}}\mathbf{P}_1\mathbf{\Omega})(\mathbf{\Omega}^{\mathrm{H}}\mathbf{P}_2\mathbf{\Omega})^{-1} + (\mathbf{\Omega}^{\mathrm{H}}\mathbf{P}_1\mathbf{\Omega})^{-1}(\mathbf{\Omega}^{\mathrm{H}}\mathbf{P}_2\mathbf{\Omega}) - 2\mathbf{I}]} \\ &= \sqrt{\frac{1}{2}\mathrm{Tr}[\mathbf{\Omega}^{\mathrm{H}}\mathbf{P}_1\mathbf{\Omega}\mathbf{\Omega}^{-1}\mathbf{P}_2\mathbf{\Omega}^{-\mathrm{H}} + \mathbf{\Omega}^{-1}\mathbf{P}_1^{-1}\mathbf{\Omega}^{-\mathrm{H}}\mathbf{\Omega}^{\mathrm{H}}\mathbf{P}_2\mathbf{\Omega} - 2\mathbf{I}]} \\ &= \sqrt{\frac{1}{2}\mathrm{Tr}(\mathbf{P}_1\mathbf{P}_2^{-1} + \mathbf{P}_1^{-1}\mathbf{P}_2 - 2\mathbf{I})} = d_{KL}(\mathbf{P}_1, \mathbf{P}_2) \end{aligned} \tag{3.22}$$

3.3 The geometry of the space of PSD matrices

In Chapter 2, we mentioned that the positive definite Hermitian matrices $\mathbf{P}$ describe a *real* manifold. Here, we introduce some of the concepts fundamental to the study of the geometry of the manifold. Since an n-dimensional manifold is a generalized surface, we will start by introducing some of the geometric concepts from the elementary consideration of a surface in a three-dimensional Euclidean space.

3.3.1 Intrinsic distance

Consider a surface $\mathcal{S}$ in a three-dimensional Euclidean space $\mathbb{R}^3$. Let $\mathbf{P}$ and $\mathbf{Q}$ be two points on $\mathcal{S}$. We define the intrinsic distance from $\mathbf{P}$ to $\mathbf{Q}$ denoted by $d(\mathbf{P}, \mathbf{Q})$ to be the infimum (greatest lower bound) of the length $L(C)$ of all possible arcs C on $\mathcal{S}$ joining $\mathbf{P}$ to $\mathbf{Q}$. It is clear that the intrinsic distance between two points on a surface always exists since the set of real numbers $L(C)$ is not empty ($\mathcal{S}$ is connected and hence arcwise connected) and is bounded from below by the Euclidean distance $\|\mathbf{P} - \mathbf{Q}\|$. It can easily be shown [61] that $d(\mathbf{P}, \mathbf{Q})$ satisfies all the properties (Eqs. (2.52)) of a metric.

Given $\mathbf{P}$ and $\mathbf{Q}$, if there exists a regular arc C joining $\mathbf{P}$ and $\mathbf{Q}$ whose length is equal to the intrinsic distance between $\mathbf{P}$ and $\mathbf{Q}$, then C is called an arc of minimum length. In the plane, $d(\mathbf{P}, \mathbf{Q})$ is the Euclidean distance and the arc of minimum length is unique and is the straight line segment between $\mathbf{P}$ and $\mathbf{Q}$.

3.3.2 Manifold and Riemannian geometry

As mentioned above, the space $\mathcal{M}$ is an open subset of the real vector space of Hermitian matrices. We will focus on the concepts of real manifold even though the elements of Hermitian matrices are not necessarily real numbers.

For an r times differentiable curve traced by a vector $\mathbf{x}(t)$ in the Euclidean space $\mathbb{R}^3$, consideration of the elemental segment δs yields the length of an arc of the curve to be [61]

$$s = \int_a^b \Big[\sum_{i=1}^{3} \Big(\frac{dx_i}{dt}\Big)\Big]^{1/2} dt = \int_a^b \sqrt{\langle \dot{\mathbf{x}}, \dot{\mathbf{x}} \rangle} dt \tag{3.23}$$

where $\mathbf{x}(t) = [x_1(t), x_2(t), x_3(t)]^T$ and $a \leq t \leq b$. We rewrite Eq. (3.23) as

$$s(t) = \int_{t_0}^{t} \sqrt{\langle \dot{\mathbf{x}}, \dot{\mathbf{x}} \rangle} dt \tag{3.24}$$

where $a \leq t_0 \leq b$. The function $s(t)$ is called the arc length of the curve. Taking the derivative of Eq. (3.24) with respect to t we obtain

$$(\dot{s})^2 = \langle \dot{\mathbf{x}}, \dot{\mathbf{x}} \rangle \tag{3.25}$$

Let $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ be the coordinate basis of $\mathbb{R}^3$, i.e., $\mathbf{e}_1 = [1, 0, 0]^T$, $\mathbf{e}_1 = [0, 1, 0]^T$, and $\mathbf{e}_1 = [0, 0, 1]^T$. Then, we form a matrix $\mathbf{G}$ as follows

$$\mathbf{G} = \begin{bmatrix} \langle \mathbf{e}_1, \mathbf{e}_1 \rangle & \langle \mathbf{e}_1, \mathbf{e}_2 \rangle & \langle \mathbf{e}_1, \mathbf{e}_3 \rangle \\ \langle \mathbf{e}_2, \mathbf{e}_1 \rangle & \langle \mathbf{e}_2, \mathbf{e}_2 \rangle & \langle \mathbf{e}_2, \mathbf{e}_3 \rangle \\ \langle \mathbf{e}_3, \mathbf{e}_1 \rangle & \langle \mathbf{e}_3, \mathbf{e}_2 \rangle & \langle \mathbf{e}_3, \mathbf{e}_3 \rangle \end{bmatrix}, \tag{3.26}$$

where $\langle \cdot, \cdot \rangle$ denotes inner product. Obviously $\mathbf{G}$ is an identity matrix. Thus, we can write Eq. (3.25) symbolically as

$$ds^2 = \langle d\mathbf{x}, d\mathbf{x} \rangle = d\mathbf{x}^T \mathbf{G} d\mathbf{x}. \tag{3.27}$$

We call the ds the line element of the arc.

Now we consider the curves on a surface in $\mathbb{R}^3$. A surface $\mathcal{S}$ in $\mathbb{R}^3$ can be expressed as a function $\mathbf{x}(u, v) = [x_1(u, v), x_2(u, v), x_3(u, v)]^T$ of two real variables u and v, which is defined in a simply-connected and bounded domain in the uv-plane. In other words, the surface $\mathcal{S}$ in $\mathbb{R}^3$ is parameterized by two real variables u and v. Then, an r-time

differentiable curve $C^r, r \geq 1$ on the surface $\mathcal{S}$ can be determined by a parametric representation as follows

$$\mathbf{x}(t) = \mathbf{x}(u(t), v(t)) \tag{3.28}$$

where t is a parameter of a real variable, i.e., by varying t the function $\mathbf{x}(t)$ traced a curve on $\mathcal{S}$ (Note that the parameters u and v are now functions of t). Then the direction of the tangent to the curve $C : \mathbf{x}(u(t), v(t))$ on the surface $\mathcal{S}$ is determined by the vector

$$\dot{\mathbf{x}} = \frac{d\mathbf{x}}{dt} = \frac{\partial \mathbf{x}}{\partial u}\frac{du}{dt} + \frac{\partial \mathbf{x}}{\partial v}\frac{dv}{dt} \tag{3.29}$$

i.e., the vector $\dot{\mathbf{x}}$ is a linear combination of the vectors $\frac{\partial \mathbf{x}}{\partial u}$ and $\frac{\partial \mathbf{x}}{\partial v}$ which are tangential to the coordinate curves passing through a point on $\mathcal{S}$ under consideration. The Eq. (3.29) can be written symbolically as

$$d\mathbf{x} = \frac{\partial \mathbf{x}}{\partial u}du + \frac{\partial \mathbf{x}}{\partial v}dv \tag{3.30}$$

Therefore, we find

$$\begin{aligned} ds^2 &= \langle d\mathbf{x}, d\mathbf{x}\rangle = \Big\langle \frac{\partial \mathbf{x}}{\partial u}du + \frac{\partial \mathbf{x}}{\partial v}dv, \frac{\partial \mathbf{x}}{\partial u}du + \frac{\partial \mathbf{x}}{\partial v}dv \Big\rangle \\ &= \Big\langle \frac{\partial \mathbf{x}}{\partial u}, \frac{\partial \mathbf{x}}{\partial u}\Big\rangle (du)^2 + 2\Big\langle \frac{\partial \mathbf{x}}{\partial u}, \frac{\partial \mathbf{x}}{\partial v}\Big\rangle dudv + \Big\langle \frac{\partial \mathbf{x}}{\partial v}, \frac{\partial \mathbf{x}}{\partial v}\Big\rangle (dv)^2 \end{aligned} \tag{3.31}$$

Let $\left\langle \frac{\partial \mathbf{x}}{\partial u}, \frac{\partial \mathbf{x}}{\partial u}\right\rangle = g_{11}$, $\left\langle \frac{\partial \mathbf{x}}{\partial u}, \frac{\partial \mathbf{x}}{\partial v}\right\rangle = g_{12}$, $\left\langle \frac{\partial \mathbf{x}}{\partial v}, \frac{\partial \mathbf{x}}{\partial u}\right\rangle = g_{21}$, and $\left\langle \frac{\partial \mathbf{x}}{\partial v}, \frac{\partial \mathbf{x}}{\partial v}\right\rangle = g_{22}$ and let

$$\mathbf{G} = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix} \tag{3.32}$$

Then we have

$$ds^2 = \sum_{i,j=1}^{2} g_{ij}dudv = [du, dv]^T \mathbf{G}[du, dv] \tag{3.33}$$

If $\mathbf{G} = \mathbf{I}$ in Eq. (3.32), then the surface $\mathcal{S}$ is a two-dimensional plane. Thus, we see that the matrix $\mathbf{G}$ is determined by the space, conversely, the matrix $\mathbf{G}$ reflects the

nature of the space. The matrix $\mathbf{G}$ is also the key to calculate the line element ds^2 of a curve in the space considered. With line element ds^2 given we can calculate the length of any smooth curve as well as the area of a bounded region in $D \subset \mathbb{R}^2$.

Now, let us turn our attention to the manifold generated by the PSD matrices. A manifold [35] is a topological space $\mathcal{M}$ which locally "looks" like a Euclidean space. If the Euclidean space is real, then the manifold is called a *real* manifold. Correspondingly, if the Euclidean space is complex, then the manifold is called a *complex* manifold. In other words, each point on $\mathcal{M}$ can be referred to by an element of the real or complex Euclidean space. Therefore, it is possible to characterize $\mathcal{M}$ by mapping neighboring points of $\mathcal{M}$ on neighboring points of $\mathbb{R}^n$ if it is real, or $\mathbb{C}^n$ if it is complex. In the case of the manifold being linear, then the manifold coincides with the Euclidean subspace. Since we have shown in Chapter 2 that all the $M \times M$ PSD matrices describe a real manifold in the real Euclidean space $\mathbb{R}^{M \times M}$ of all Hermitian matrices, henceforth, we will focus only on real manifolds. The concept of real manifold is roughly shown in Figure 3.4. More precisely, that a topological space locally looks like a Euclidean space means that the tangent space at every point (say x) on the manifold (denoted by $\mathcal{T}_{\mathcal{M}}(x)$) is isomorphic to $\mathbb{R}^n$, denoted by $\mathcal{T}_{\mathcal{M}}(x) \cong \mathbb{R}^n$. [3]

The study of manifolds can be from an extrinsic point of view, in which the manifolds are considered as lying in a high dimensional Euclidean space (as introduced above), or from an intrinsic point of view which started from the work of Riemann in which the manifolds are considered as given in a free-standing way.

From an extrinsic point of view, the manifold $\mathcal{M}$ can be thought of as a subset of the Euclidean space $\mathbb{R}^n$, then the distance on $\mathcal{M}$ can be defined as the induced

[3] $\mathcal{T}_{\mathcal{M}}(x) \cong \mathbb{R}^n$ means that there is a bijective map ϕ from $\mathcal{T}_{\mathcal{M}}(x)$ to $\mathbb{R}^n$ such that both ϕ and its inverse ϕ^{-1} are homomorphisms, i.e., ϕ satisfies $\phi(a+b) = \phi(a) + \phi(b)$ and $\phi(ka) = k\phi(a)$ for $a, b \in \mathcal{T}_{\mathcal{M}}(x)$ and $k \in \mathbb{R}$.

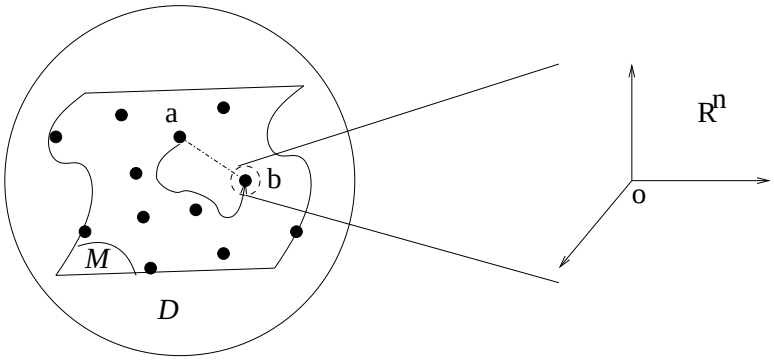


Figure 3.4: Distance on manifold

Euclidean distance. For example, the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$ of the sphere $\mathcal{S}^{n-1} = \{\mathbf{x} \in \mathbb{R}^n : \sum_{i=1}^n x_i^2 = 1\}$ is the cord joining $\mathbf{x}$ and $\mathbf{y}$. However, if $\mathcal{M}$ is a smooth connected submanifold in $\mathbb{R}^n$ then, one has the induced Riemannian distance, which is defined as the infimum of lengths of curves contained in $\mathcal{M}$ and joining $\mathbf{x}$ and $\mathbf{y}$. To see this, let $(x_1, \cdots, x_n)$ be the coordinate system in $\mathbb{R}^n$ and $\mathcal{M}$ embedded in $\mathbb{R}^n$ be parameterized in terms of coordinates $\mathbf{q} = (q_1, \cdots, q_k)$, $k \leq n$ as $x_i = x_i(q_1, \cdots, q_k)$, $i = 1, \cdots, n$. Then the Riemannian metric $g_{\mathcal{M}}$ on $\mathcal{M}$ is defined from the Euclidean length element according to

$$ds^2 = \sum_{i=1}^n (dx_i)^2 = \sum_{i=1}^n \Big(\sum_{m=1}^k \frac{\partial x_i}{\partial q_m} dq_m \Big)^2 = \sum_{m,n=1}^k g_{mn} dq_m dq_n \tag{3.34}$$

As a specific example, we consider a two-dimensional ($k = 2$) sphere $x^2+y^2+z^2 = 1$ embedded in $\mathbb{R}^3$ with coordinates (x, y, z) and length element $ds^2 = (dx)^2 + (dy)^2 + (dz)^2$. Let the parameterization of the points on the sphere be in terms of spherical coordinates (ϕ, θ) as

$$\begin{aligned} x &= r\sin(\theta)\cos(\phi) \\ y &= r\sin(\theta)\sin(\phi) \\ z &= r\cos(\theta), \qquad 0 \leq \theta \leq \pi,\ 0 \leq \phi \leq 2\pi \end{aligned} \tag{3.35}$$

Then we have

$$ds^2 = r^2(d\theta)^2 + r^2 \sin^2\theta (d\phi)^2. \tag{3.36}$$

Comparing Equations (3.34) and (3.36) we obtain the matrix elements of the metric $g_{\mathcal{M}}$ such that $g_{11} = r^2$, $g_{12} = g_{21} = 0$, and $g_{22} = r^2 \sin^2\theta$.

Riemann [78] started the study of manifold from an *intrinsic* point of view by using a quadratic formula for the infinitesimal change in distance ds. Such a structure is called a *Riemannian metric*. A manifold on which a Riemannian metric is defined is called a *Riemannian manifold*.

Specifically, a Riemannian manifold is a *differentiable manifold* in which each *tangent space*[4] is equipped with an inner product $\langle \cdot, \cdot \rangle$ in a manner which varies smoothly from point to point, i.e., a Riemannian metric is a family of positive definite inner products[5] defined by

$$g_p : \mathcal{T}_{\mathcal{M}}(p) \times \mathcal{T}_{\mathcal{M}}(p) \to \mathbb{R}, \qquad p \in \mathcal{M} \tag{3.37}$$

where $\mathcal{T}_{\mathcal{M}}(p)$ denotes the tangent space of $\mathcal{M}$ at the point $p \in \mathcal{M}$. The function defined in Eq. (3.37) is called a Riemannian metric on $\mathcal{M}$ and is a differentiable function from point to point on the Riemannian manifold $\mathcal{M}$.

With a Riemannian metric defined, the line element of a curve on the manifold is given by

$$ds^2 = \sum_{i,j=1}^{n} g_{ij}(\mathbf{x}) dx_i dx_j, \tag{3.38}$$

if $(x_1, \cdots, x_n)$ are local coordinates of class C^∞ in an open subset O of $\mathcal{M}$ at each $p \in O$.

[4]Some elementary notions of Riemannian geometry and the formal definitions of these terms in italics are given in Appendix B

[5]An interesting and important question is: "What is the best Riemannian structure on the manifold?" [13]. Even though we are not involved in seeking the answer to this question in this thesis, we still have to make a choice if several Riemannian metrics are available. This choice depends on the application. Results of applying different Riemannian distances to EEG signal classification will be presented in subsequent chapters.

3.4 Riemannian distances for matrix quantities

Having established the concept of Riemannian metric, we now develop some Riemannian distances on the manifold $\mathcal{M}$ of the PSD matrices $\{\mathbf{P}\}$ for the use in the classification of EEG signals. To achieve that, we must first endow the manifold $\mathcal{M}$ with a Riemannian metric g [40] which, as mentioned in the last section, is defined as an inner product on the tangent space $\mathcal{T}_{\mathcal{M}}(\mathbf{P})$, of each point $\mathbf{P}$ on $\mathcal{M}$. Thus, we obtain a Riemannian manifold $(\mathcal{M}, g)$. Since there are infinitely many possible Riemannian metrics on a differentiable manifold, a suitable one for our purpose of signal classification has to be chosen.

3.4.1 Riemannian distance d_{R_1}

Let $(\mathcal{M}, g_P)$ be the Riemannian manifold $\mathcal{M}$ with the Riemannian metric g_P. Let $[\theta_1, \theta_2]$ be a closed interval in $\mathbb{R}$, and let $\mathbf{\Gamma}(\theta) : [\theta_1, \theta_2] \rightarrow \mathcal{M}$ be a sufficiently smooth curve on $\mathcal{M}$ such that $\mathbf{\Gamma}(\theta_1) = \mathbf{P}_1$ and $\mathbf{\Gamma}(\theta_2) = \mathbf{P}_2$. The length of the curve $\mathbf{\Gamma}(\theta)$ is defined as [40]

$$l(\mathbf{\Gamma}) = \int_{\theta_1}^{\theta_2} \parallel \dot{\mathbf{\Gamma}}(\theta) \parallel d\theta \tag{3.39}$$

where $\dot{\mathbf{\Gamma}}(\theta) = \frac{d\mathbf{\Gamma}(\theta)}{d\theta}$. With the use of the given Riemannian metric g_P, Eq. (3.39) can be written as

$$l(\mathbf{\Gamma}) = \int_{\theta_1}^{\theta_2} \sqrt{g_{\mathbf{\Gamma}(\theta)}(\dot{\mathbf{\Gamma}}(\theta), \dot{\mathbf{\Gamma}}(\theta))} d\theta \tag{3.40}$$

Then, the global Riemannian distance between the two points $\mathbf{P}_1$ and $\mathbf{P}_2$ on the Riemannian manifold $(\mathcal{M}, g_P)$ is defined as the shortest length of curves connecting the two points such that

$$d_{R_1}(\mathbf{P}_1, \mathbf{P}_2) \triangleq \min_{\mathbf{\Gamma}:[\theta_1,\theta_2]\rightarrow\mathcal{M}} \{l(\mathbf{\Gamma})\} \tag{3.41}$$

However, since this usually leads to a set of nonlinear differential equations which is difficult to solve (see Appendix B), it is not easy to find a closed form formula for the Riemannian distance directly from Eq. (3.41). To overcome this difficulty, we resort to the theory of fiber bundles [53]. The basic idea is introduced in the following and more details can be found in Appendix B.

First, let us introduce a lemma for the representation of a point $\mathbf{P}$ in the manifold in a Hilbert space.

Lemma 3.1 *For a point $\mathbf{P} \in \mathcal{M}$, there exists a $\tilde{\mathbf{P}}$ in a Hilbert space $\mathcal{H}_M$ such that $\mathbf{P} = \tilde{\mathbf{P}}\tilde{\mathbf{P}}^H$.*

The proof of this Lemma is presented in Appendix C. □

Lemma 3.1 has important implications. It shows that for every PSD matrix $\mathbf{P}$, there exists another matrix $\tilde{\mathbf{P}} \in \mathcal{H}_M$ which though is not unique, can be viewed as a representation of $\mathbf{P}$ in the Hilbert space. (Henceforth, we will use the notation that $\tilde{\mathbf{X}}$ denote the representation in $\mathcal{H}_M$ of the matrix $\mathbf{X} \in \mathcal{M}$.)

Let the space $\tilde{\mathcal{H}}$ be defined as

$$\tilde{\mathcal{H}} = \{\tilde{\mathbf{P}} : \tilde{\mathbf{P}}\tilde{\mathbf{P}}^H = \mathbf{P} \in \mathcal{M}\} \tag{3.42}$$

The space $\tilde{\mathcal{H}}$ can be considered as a subset of $\mathcal{H}_M$, i.e., $\tilde{\mathcal{H}} \subset \mathcal{H}_M$. At any point $\tilde{\mathbf{P}} \in \tilde{\mathcal{H}}$, the tangent space, denoted by $\mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$, is the collection of vectors tangent to any smooth curve passing through $\tilde{\mathbf{P}}$. Since there are an infinite number of smooth curve that can be drawn through $\tilde{\mathbf{P}}$, the tangent space at $\tilde{\mathbf{P}}$ is therefore just the Hilbert space local to the neighborhood of $\tilde{\mathbf{P}}$. Now, let $\mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ be resolved into its horizontal and vertical subspaces $\mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ and $\mathcal{V}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ respectively, i.e., $\mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}}) = \mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}}) \oplus \mathcal{V}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$. We endow $\mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ with a real-valued inner product such that for $\tilde{\mathbf{V}}_1, \tilde{\mathbf{V}}_2 \in \mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$,

$$\langle \tilde{\mathbf{V}}_1, \tilde{\mathbf{V}}_2 \rangle_{\mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})} = \frac{1}{2}\mathrm{Tr}(\tilde{\mathbf{V}}_1^H \tilde{\mathbf{V}}_2 + \tilde{\mathbf{V}}_2^H \tilde{\mathbf{V}}_1) \tag{3.43}$$

Then, for any two $M \times M$ complex matrices $\tilde{\mathbf{A}}, \tilde{\mathbf{B}} \in \mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$, we have

$$\langle \tilde{\mathbf{A}}, \tilde{\mathbf{B}} \rangle_{\mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})} = \frac{1}{2}\mathrm{Tr}(\tilde{\mathbf{A}}^H \tilde{\mathbf{B}} + \tilde{\mathbf{B}}^H \tilde{\mathbf{A}}) \tag{3.44}$$

with the induced norm being $\|\tilde{\mathbf{A}}\|^2 = \mathrm{Tr}(\tilde{\mathbf{A}}^H \tilde{\mathbf{A}})$.

We now relate $\mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ to $\mathcal{T}_{\mathcal{M}}(\mathbf{P})$ by establishing an isometry between them so that we can let the inner product defined on $\mathcal{T}_{\mathcal{M}}(\mathbf{P})$ equals $\langle \tilde{\mathbf{A}}, \tilde{\mathbf{B}} \rangle$ in Eq. (3.44) yielding a natural Riemannian metric on $\mathcal{M}$. We have the following lemma:

Lemma 3.2 *Let $\mathbf{P} \in \mathcal{M}$ be such that $\mathbf{P} = \tilde{\mathbf{P}}\tilde{\mathbf{P}}^H$ and let $\mathbf{A}, \mathbf{B} \in \mathcal{T}_{\mathcal{M}}(\mathbf{P})$. If the Riemannian metric on $\mathcal{M}$ is given by*

$$g_{\mathbf{P}}(\mathbf{A}, \mathbf{B}) = \frac{1}{2}\mathrm{Tr}\mathbf{A}\mathbf{K} \tag{3.45}$$

where $\mathbf{K}$ is a Hermitian matrix such that $\mathbf{K}\mathbf{P} + \mathbf{P}\mathbf{K} = \mathbf{B}$, then $\mathcal{T}_{\mathcal{M}}(\mathbf{P})$ and $\mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ are isometric. □

The proof of Lemma 3.2 is presented in Appendix D. □

Let $\Delta\mathbf{P}$ be a vector on the tangent space $\mathcal{T}_{\mathcal{M}}(\mathbf{P})$ measured from $\mathbf{P}$, then, the squared distance between two very close points (say $\mathbf{P}$ and $\mathbf{P}'$) in an infinitesimal region on $\mathcal{M}$ can be approximated by the norm of $\Delta\mathbf{P}$, i.e.,

$$d^2(\mathbf{P}, \mathbf{P}') \simeq \|\Delta\mathbf{P}\|^2 = g_{\mathbf{P}}(\Delta\mathbf{P}, \Delta\mathbf{P}) = \frac{1}{2}\mathrm{Tr}\mathbf{\Delta}\mathbf{P}\mathbf{K} \tag{3.46}$$

where $\mathbf{K}\mathbf{P} + \mathbf{P}\mathbf{K} = \Delta\mathbf{P}$. In other words, the infinitesimal norm induced by the Riemannian metric in the Eq. (3.45) represents a measure of the distance between two points in $\mathcal{M}$ being infinitesimally close to each other.

Basically, the manifold $\mathcal{M}$ and the space $\tilde{\mathcal{H}}$ are considered as the base space and the total space, respectively. The projection map $\pi : \tilde{\mathcal{H}} \to \mathcal{M}$ associates each point $\mathbf{P} \in \mathcal{M}$ with $\pi^{-1}(\mathbf{P}) \subset \tilde{\mathcal{H}}$ constituting the *fiber* above $\mathbf{P} \in \mathcal{M}$. (The rigorous

definitions of italicized terms in this paragraph are given in Appendix B.) A *connection* on the fiber bundle is a rule that pairs each smooth curve through a point $\mathbf{P} \in \mathcal{M}$ with a class of corresponding smooth curves in $\tilde{\mathcal{H}}$, one through each point in the fiber above $\mathbf{P}$, known as its *lifts*. Let $\mathbf{\Gamma}(\theta) : [\theta_1, \theta_2] \to \mathcal{M}$ with $\mathbf{\Gamma}(\theta_1) = \mathbf{P}_1$ and $\mathbf{\Gamma}(\theta_2) = \mathbf{P}_2$ be a smooth curve on $\mathcal{M}$ (i.e., $\mathbf{\Gamma}(\theta)$ traces a curve on $\mathcal{M}$ with the varying of θ in $[\theta_1, \theta_2]$). Let $\tilde{\mathbf{\Gamma}}(\theta) : [\theta_1, \theta_2] \to \mathcal{H}$, with $\tilde{\mathbf{\Gamma}}(\theta_1) = \tilde{\mathbf{P}}_1$ and $\tilde{\mathbf{\Gamma}}(\theta_2) = \tilde{\mathbf{P}}_2$, be a curve on $\tilde{\mathcal{H}}$ and $\tilde{\mathbf{P}}_1, \tilde{\mathbf{P}}_2 \in \tilde{\mathcal{H}}$ being the representatives of $\mathbf{P}_1$ and $\mathbf{P}_2$ in the Hilbert space $\mathcal{H}_M$. Then we say that $\tilde{\mathbf{\Gamma}}(\theta)$ is a *horizontal lift* of $\mathbf{\Gamma}(\theta)$ if $\mathbf{\Gamma}(\theta)$ is the image of $\tilde{\mathbf{\Gamma}}(\theta)$ under π and the tangent vector to $\tilde{\mathbf{\Gamma}}(\theta)$ always lies in the horizontal subspace $\mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ of the tangent space $\mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ at each point along $\tilde{\mathbf{\Gamma}}(\theta)$. The concept of the horizontal lift is illustrated in Figure 3.5.

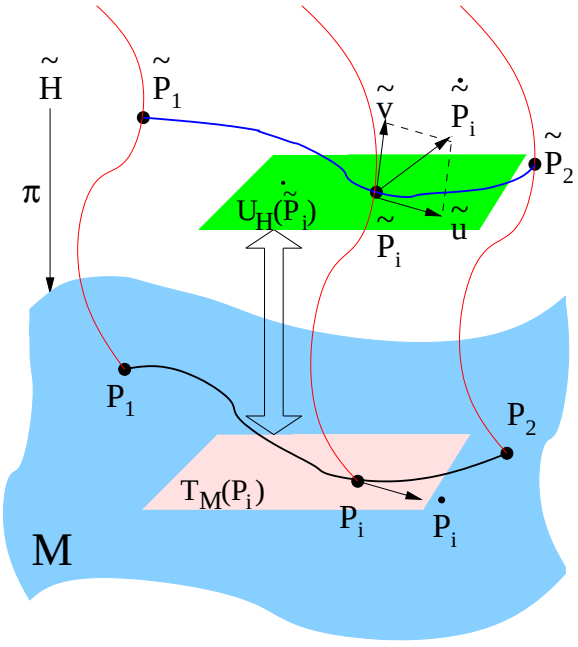


Figure 3.5: Illustration of horizontal lift

Recall that a map $\pi : \tilde{\mathcal{H}} \to \mathcal{M}$ is called a *Riemannian submersion* at $\tilde{\mathbf{P}} \in \tilde{\mathcal{H}}$

if the induced tangent map $\pi_* : \mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}}) \rightarrow \mathcal{T}_{\mathcal{M}}(\pi(\tilde{\mathbf{P}}))$ is an isometry, where $\mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ is the horizontal subspace of $\mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$. That the map $\pi : \tilde{\mathcal{H}} \rightarrow \mathcal{M}$ is a Riemannian submersion is guaranteed by the following lemma:

Lemma 3.3 *A fiber bundle with base space $\mathcal{M}$, the total space $\tilde{\mathcal{H}}$ and the projection map $\pi : \tilde{\mathcal{H}} \rightarrow \mathcal{M}$ defined by $\pi(\tilde{\mathbf{P}}) = \mathbf{P}$ is a Riemannian submersion if the horizontal subspace $\mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ is endowed with a metric defined in Eq. (3.44).* □

The proof of Lemma 3.3 follows directly from the result of Lemma 3.2.

From the above we can now conclude that the curve $\tilde{\mathbf{\Gamma}}(\theta)$ is the unique horizontal lift of the curve $\mathbf{\Gamma}(\theta)$ if we employ the map $\pi : \tilde{\mathcal{H}} \rightarrow \mathcal{M}$ such that $\pi(\tilde{\mathbf{P}}) = \mathbf{P}$ and $\pi^{-1}(\mathbf{P}) = \tilde{\mathbf{P}}\mathbf{G}$ with $\mathbf{G}$ being a unitary matrix. Furthermore, the induced tangent map $\pi_*|_{\tilde{\mathbf{P}}} : \mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}}) \rightarrow \mathcal{T}_{\mathcal{M}}(\mathbf{P})$ such that $\pi_*|_{\tilde{\mathbf{P}}}(\dot{\tilde{\mathbf{P}}}) = \dot{\mathbf{P}}$, $\dot{\tilde{\mathbf{P}}} \in \mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ and $\dot{\mathbf{P}} \in \mathcal{T}_{\mathcal{M}}(\mathbf{P})$, is an isometry between $\mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ and $\mathcal{T}_{\mathcal{M}}(\mathbf{P})$. (In this way we have made $\pi : \tilde{\mathcal{H}} \rightarrow \mathcal{M}$ a *principal G-bundle* [6] and it is also a Riemannian submersion.)

Once the Riemannian submersion $\pi : \tilde{\mathcal{H}} \rightarrow \mathcal{M}$ is established, we can endow the horizontal subspace with the metric as defined in Eq. (3.44) so that if $\mathbf{\Gamma}(\theta)$ is a geodesic curve on $\mathcal{M}$ (i.e., it has the shortest length), then its horizontal lift $\tilde{\mathbf{\Gamma}}(\theta)$ is the corresponding geodesic curve on $\tilde{\mathcal{H}}$. This is given by Lemma 3.4 [40] in the following:

Lemma 3.4 *Let $\pi : \tilde{\mathcal{H}} \rightarrow \mathcal{M}$ be a Riemannian submersion. Let $\tilde{\mathbf{\Gamma}}(\theta) : [0, 1] \rightarrow \tilde{\mathcal{H}}$ be a geodesic of $\mathcal{M}$ with $\tilde{\mathbf{\Gamma}}(0) = \tilde{\mathbf{P}}$. If $\dot{\tilde{\mathbf{\Gamma}}}(0)$ is horizontal, i.e., it lies in $\mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$, then $\dot{\tilde{\mathbf{\Gamma}}}(\theta)$ is horizontal for any θ, and the curve $\pi \circ \tilde{\mathbf{\Gamma}}(\theta)$ is a geodesic of $\mathcal{M}$, of same length as $\tilde{\mathbf{\Gamma}}(\theta)$. Conversely, let $\tilde{\mathbf{P}} \in \tilde{\mathcal{H}}$ and $\mathbf{\Gamma}(\theta) : [0, 1] \rightarrow \mathcal{M}$ be a geodesic of $\mathcal{M}$ with $\mathbf{\Gamma}(0) = \pi(\tilde{\mathbf{P}})$. Then there exists a unique local horizontal lift $\tilde{\mathbf{\Gamma}}(\theta)$ of $\mathbf{\Gamma}(\theta)$, and $\tilde{\mathbf{\Gamma}}(\theta)$*

[6] Let G be a Lie group acting on $\tilde{\mathcal{H}}$ such that $(\tilde{\mathbf{P}})$ is mapped to $\tilde{\mathbf{P}}\mathbf{G}$ and $\tilde{\mathbf{P}}\mathbf{G} \neq \tilde{\mathbf{P}}$ for $\mathbf{G} \neq \mathbf{I}$. A surjective submersion $\pi : \tilde{\mathcal{H}} \rightarrow \mathcal{M}$ is said to be a principal G-bundle if $\{\tilde{\mathbf{P}}\mathbf{G} : \mathbf{G} \in G\} = \pi^{-1}(\pi(\tilde{\mathbf{P}}))$ for any $\tilde{\mathbf{P}} \in \tilde{\mathcal{H}}$

is also a geodesic of $\tilde{\mathcal{H}}$*. Finally, the completeness of* $\tilde{\mathcal{H}}$ *implies the completeness of* $\mathcal{M}$*.* □

With the above results, we can now establish the geodesic distance between two points $\mathbf{P}_1, \mathbf{P}_2 \in \mathcal{M}$ by evaluating the geodesic distance in $\mathcal{M}$ with the corresponding geodesic distance between two point $\tilde{\mathbf{P}}_1$ and $\tilde{\mathbf{P}}_1$ on $\tilde{\mathcal{H}}$. Thus, we have the following theorem[7]:

Theorem 3.1 *For* $\mathbf{P}_1, \mathbf{P}_2 \in \mathcal{M}$ *the geodesic distance between* $\mathbf{P}_1$ *and* $\mathbf{P}_2$ *is given by*

$$d_{R_1}(\mathbf{P}_1, \mathbf{P}_2) = \sqrt{\mathrm{Tr}\mathbf{P}_1 + \mathrm{Tr}\mathbf{P}_2 - 2\mathrm{Tr}(\mathbf{P}_1^{1/2}\mathbf{P}_2\mathbf{P}_1^{1/2})^{1/2}} \tag{3.47}$$

□

Proof: Let $\mathbf{\Gamma}(\theta) : [\theta_1, \theta_2] \rightarrow \mathcal{M}$ with $\mathbf{\Gamma}(\theta_1) = \mathbf{P}_1 \in \mathcal{M}$ and $\mathbf{\Gamma}(\theta_2) = \mathbf{P}_2 \in \mathcal{M}$ be the geodesic connecting $\mathbf{P}_1$ and $\mathbf{P}_2$ on $\mathcal{M}$. Let $\tilde{\mathbf{\Gamma}}(\theta) : [\theta_1, \theta_2] \rightarrow \tilde{\mathcal{H}}$ with $\tilde{\mathbf{\Gamma}}(\theta_1) = \tilde{\mathbf{P}}_1 \in \tilde{\mathcal{H}}$ and $\tilde{\mathbf{\Gamma}}(\theta_2) = \tilde{\mathbf{P}}_2 \in \tilde{\mathcal{H}}$ be the horizontal lift of $\mathbf{\Gamma}(\theta)$. The fact that $\pi : \tilde{\mathcal{H}} \rightarrow \mathcal{M}$ is a Riemannian submersion means that the length of $\tilde{\mathbf{\Gamma}}(\theta)$ depends only on the metric associated with the horizontal subspace $\mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{\Gamma}}(\theta))$. Since $\mathcal{T}_{\mathcal{M}}(\mathbf{\Gamma}(\theta))$ and $\mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{\Gamma}}(\theta))$ are isometric we must have $l(\mathbf{\Gamma}) = l(\tilde{\mathbf{\Gamma}})$. Thus, the minimum of $l(\mathbf{\Gamma})$ can be achieved by finding the minimum of $l(\tilde{\mathbf{\Gamma}})$.

From our construction of the space $\tilde{\mathcal{H}}$, we immediately have $\mathcal{T}_{\tilde{\mathcal{H}}}(\mathbf{\Gamma}(\theta)) = \mathcal{T}_{\mathcal{H}_M}(\mathbf{\Gamma}(\theta))$. Furthermore, the metric endowed to $\mathcal{T}_{\tilde{\mathcal{H}}}(\mathbf{\Gamma}(\theta))$ is the unique real metric endowed to $\mathcal{T}_{\mathcal{H}_M}(\mathbf{\Gamma}(\theta))$ if the $\mathcal{H}_M$ is endowed with the Hilbert-Schmidt inner product. Thus, the shortest curve (geodesic) connecting two points $\tilde{\mathbf{P}}_1$ and $\tilde{\mathbf{P}}_2$ in the space $\tilde{\mathcal{H}}$ must the

[7]Bures [21] had proposed a similar distance measure for the space of positive definite matrices *with unity traces* which is not applicable to EEG classification since PSD matrices of EEG signals are not under such a constraint. The Riemannian distance d_{R_1} developed in this thesis places no constraint on the trace.

straight line segment connecting $\tilde{\mathbf{P}}_1$ and $\tilde{\mathbf{P}}_2$ in the Hilbert space $\mathcal{H}_M$. Therefore, we must have

$$\min_{\tilde{\mathbf{\Gamma}}} l(\tilde{\mathbf{\Gamma}}) = \min_{\substack{\mathbf{P}_1=\tilde{\mathbf{P}}_1\tilde{\mathbf{P}}_1^H \\ \mathbf{P}_2=\tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2^H}} \| \tilde{\mathbf{P}}_1 - \tilde{\mathbf{P}}_2 \| = \min_{\substack{\mathbf{P}_1=\tilde{\mathbf{P}}_1\tilde{\mathbf{P}}_1^H \\ \mathbf{P}_2=\tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2^H}} [\mathrm{Tr}(\tilde{\mathbf{P}}_1 - \tilde{\mathbf{P}}_2)^H(\tilde{\mathbf{P}}_1 - \tilde{\mathbf{P}}_2)]^{1/2} \tag{3.48}$$

Thus, writing $\tilde{\mathbf{P}}_1 = \mathbf{P}_1^{1/2}\mathbf{U}_1$ and $\tilde{\mathbf{P}}_2 = \mathbf{P}_2^{1/2}\mathbf{U}_2$ with $\mathbf{U}_1$ and $\mathbf{U}_2$ being unitary matrices [14], we can define the squared geodesic distance between $\mathbf{P}_1$ and $\mathbf{P}_2$ on the manifold $\mathcal{M}$ as

$$\begin{aligned} d^2_{R_1}(\mathbf{P}_1, \mathbf{P}_2) &= \min_{\mathbf{U}_1,\mathbf{U}_2} \| \tilde{\mathbf{P}}_1 - \tilde{\mathbf{P}}_2 \|^2 \\ &= \min_{\mathbf{U}_1,\mathbf{U}_2} \mathrm{Tr}((\tilde{\mathbf{P}}_1 - \tilde{\mathbf{P}}_2)^H(\tilde{\mathbf{P}}_1 - \tilde{\mathbf{P}}_2)) \\ &= \min_{\mathbf{U}_1,\mathbf{U}_2} [\mathrm{Tr}\mathbf{P}_1 + \mathrm{Tr}\mathbf{P}_2 - 2\Re(\mathrm{Tr}\mathbf{U}_2\mathbf{U}_1^H\mathbf{P}_1^{1/2}\mathbf{P}_2^{1/2})] \end{aligned} \tag{3.49}$$

where $\Re(\cdot)$ denotes the real part of a complex quantity. Minimization of Eq. (3.52) is equivalent to the maximization of the quantity $\Re(\mathrm{Tr}\mathbf{U}_2\mathbf{U}_1^H\mathbf{P}_2^{1/2}\mathbf{P}_1^{1/2})$ with respect to the unitary matrices $\mathbf{U}_1$ and $\mathbf{U}_2$. The result of this is well-known [50]

$$\max_{\mathbf{U}_1,\mathbf{U}_2} \Re(\mathrm{Tr}\mathbf{U}_2\mathbf{U}_1^H\mathbf{P}_2^{1/2}\mathbf{P}_1^{1/2}) = \mathrm{Tr}(\mathbf{P}_1^{1/2}\mathbf{P}_2\mathbf{P}_1^{1/2})^{1/2} \tag{3.50}$$

if $\mathbf{U}_2\mathbf{U}_1^H = \mathbf{V}_2\mathbf{V}_1^H$ where $\mathbf{P}_2^{1/2}\mathbf{P}_1^{1/2} = \mathbf{V}_1\mathbf{\Sigma}\mathbf{V}_2^H$ with $\mathbf{\Sigma}$ being the singular value matrix, and $\mathbf{V}_1$ and $\mathbf{V}_2$ being the left and right singular vector matrices of $\mathbf{P}_2^{1/2}\mathbf{P}_1^{1/2}$ respectively. □

Theorem 3.1 establishes a Riemannian distance between two points in $\mathcal{M}$ suitable for the measurement of distance of EEG signals represented by their power spectral density matrices. However, in applying this Riemannian distance to the classification of EEG signals, it is often desirable to weight the measured power spectral density matrices to enhance their similarity/dissimilarity. To do that, we incorporate a *positive definite Hermitian* weighting matrix $\mathbf{W}$ to the power spectral density matrices and obtain the following corollary:

Corollary 3.1 *Let* $\mathbf{W}$ *be a positive definite Hermitian matrix which can be written as* $\mathbf{W} = \mathbf{\Omega}\mathbf{\Omega}^H$. *Let* $\mathbf{P}_1, \mathbf{P}_2 \in \mathcal{M}$ *and let* $\mathbf{\Omega}^H\mathbf{P}_1\mathbf{\Omega}$ *and* $\mathbf{\Omega}^H\mathbf{P}_2\mathbf{\Omega}$ *be the weighted matrices of* $\mathbf{P}_1$ *and* $\mathbf{P}_2$, *respectively. Then the weighted geodesic distance between* $\mathbf{P}_1$ *and* $\mathbf{P}_2$ *is given by*

$$d_{R_1W}(\mathbf{P}_1, \mathbf{P}_2) = \sqrt{\mathrm{Tr}\mathbf{W}\mathbf{P}_1 + \mathrm{Tr}\mathbf{W}\mathbf{P}_2 - 2\mathrm{Tr}(\mathbf{P}_2^{1/2}\mathbf{W}\mathbf{P}_1\mathbf{W}\mathbf{P}_2^{1/2})^{1/2}} \tag{3.51}$$

□

Proof: If we denote the weighted matrices of $\mathbf{P}_1$ and $\mathbf{P}_2$ by $\mathbf{P}_{1W} = \mathbf{\Omega}^H\mathbf{P}_1\mathbf{\Omega}$ and $\mathbf{P}_{2W} = \mathbf{\Omega}^H\mathbf{P}_2\mathbf{\Omega}$ respectively, then it is easy to see that $\mathbf{P}_{1W}$ and $\mathbf{P}_{2W}$ are positive definite Hermitian matrices, i.e., $\mathbf{P}_{1W}, \mathbf{P}_{2W} \in \mathcal{M}$. Let $\tilde{\mathbf{P}}_1 = \mathbf{P}_1^{1/2}\mathbf{U}_1$ and $\tilde{\mathbf{P}}_2 = \mathbf{P}_2^{1/2}\mathbf{U}_2$, where $\mathbf{U}_1$ and $\mathbf{U}_2$ are unitary matrices. Let $\tilde{\mathbf{P}}_{1W} = \mathbf{\Omega}^H\tilde{\mathbf{P}}_1$ and $\tilde{\mathbf{P}}_{2W} = \mathbf{\Omega}^H\tilde{\mathbf{P}}_2$. Then $\mathbf{P}_{1W} = \tilde{\mathbf{P}}_{1W}\tilde{\mathbf{P}}_{1W}^H$ and $\mathbf{P}_{2W} = \tilde{\mathbf{P}}_{2W}\tilde{\mathbf{P}}_{2W}^H$ and we have,

$$\begin{aligned} d_{R_1W}^2(\mathbf{P}_1, \mathbf{P}_2) &= \min_{\mathbf{U}_1,\mathbf{U}_2} \|\tilde{\mathbf{P}}_{1W} - \tilde{\mathbf{P}}_{2W}\|^2 \\ &= \min_{\mathbf{U}_1,\mathbf{U}_2} [\mathrm{Tr}\mathbf{P}_1 + \mathrm{Tr}\mathbf{P}_2 - 2\Re(\mathrm{Tr}\mathbf{U}_2\mathbf{U}_1^H\mathbf{P}_1^{1/2}\mathbf{W}\mathbf{P}_2^{1/2})] \end{aligned} \tag{3.52}$$

Proceeding as in the derivation of Theorem 3.1, the result of Corollary 3.1 follows. □ We will use both geodesic distances in Eqs. (3.47) and (3.51) for classifying the EEG signals in the ensuing sections.

In the following we will show that the Riemannian distance developed is a true distance, i.e., it satisfies all axioms of the definition of distance function. We first introduce a well-known inequality: Let $\lambda_m(\mathbf{A})$ denote the mth eigenvalue of an $M \times M$ positive semidefinite $\mathbf{A}$. We define the the p-norm of $\mathbf{A}$ as

$$\|\mathbf{A}\|_p = \left(\sum_{m=1}^{M} \lambda_m^p(\mathbf{A}) \right)^{1/p} \tag{3.53}$$

We note that for $p = 2$, this is the same as the Frobenius norm induced by the inner product $\langle \mathbf{A}, \mathbf{A} \rangle = Tr[\mathbf{A}\mathbf{A}^H]$. It is well-known [50] that

$$\|\mathbf{A}\mathbf{B}\|_1 \leq \|\mathbf{A}\|_2 \|\mathbf{B}\|_2 \tag{3.54}$$

Theorem 3.2 *The geodesic distance given by Eq. (3.47) is a true distance, i.e., it satisfies nonnegativity, symmetry and triangle inequality.*

Proof. We need to show the nonnegativity, symmetry, and triangular inequality according to the definition of distance.

a) **Nonnegativity:** Let $\mathbf{A} = \sqrt{\mathbf{P}_1}$, $\mathbf{B} = \sqrt{\mathbf{P}_2}$ and $F(\mathbf{P}_1, \mathbf{P}_2) = \mathrm{Tr}(\mathbf{P}_1^{1/2}\mathbf{P}_2\mathbf{P}_1^{1/2})^{1/2}$. Then Since $\mathbf{P}_1$ and $\mathbf{P}_2$ are Hermitian, by Eq. (3.53), we have

$$\|\mathbf{P}_1\|_2 = (Tr\mathbf{P}_1)^{1/2} \quad \text{and} \quad \|\mathbf{P}_2\|_2 = (Tr\mathbf{P}_2)^{1/2} \tag{3.55}$$

and

$$F(\mathbf{P}_1, \mathbf{P}_2) = \left\| \sqrt{\mathbf{P}_1}\sqrt{\mathbf{P}_2} \right\|_1. \tag{3.56}$$

By applying Eq. (3.54) we obtain

$$\left\| \sqrt{\mathbf{P}_1}\sqrt{\mathbf{P}_2} \right\|_1 \leq \left\| \sqrt{\mathbf{P}_1} \right\|_2 \left\| \sqrt{\mathbf{P}_2} \right\|_2. \tag{3.57}$$

By using Eqs. (3.55) and (3.56) we have

$$d_{R_1}^2(\mathbf{P}_1, \mathbf{P}_2) \geq Tr\mathbf{P}_1 + Tr\mathbf{P}_2 - 2(Tr\mathbf{P}_1)^{1/2}(Tr\mathbf{P}_2)^{1/2} \geq 0. \tag{3.58}$$

Therefore, d_{R_1} must be a nonnegative number.

b) **Symmetry:** Let $\mathbf{v}_m$ and λ_m be an eigenvector and eigenvalue pair of $\mathbf{P}_1^{1/2}\mathbf{P}_2\mathbf{P}_1^{1/2}$. Then we have

$$\begin{aligned} \lambda_m(\mathbf{P}_2^{1/2}\mathbf{P}_1^{1/2}\mathbf{v}_m) &= \mathbf{P}_2^{1/2}\mathbf{P}_1^{1/2}(\lambda_m\mathbf{v}_m) \\ &= \mathbf{P}_2^{1/2}\mathbf{P}_1^{1/2}(\mathbf{P}_1^{1/2}\mathbf{P}_2\mathbf{P}_1^{1/2})\mathbf{v}_m \\ &= (\mathbf{P}_2^{1/2}\mathbf{P}_1\mathbf{P}_2^{1/2})(\mathbf{P}_2^{1/2}\mathbf{P}_1^{1/2}\mathbf{v}_m). \end{aligned} \tag{3.59}$$

Thus, λ_m is also an eigenvalue of $\mathbf{P}_2^{1/2}\mathbf{P}_1\mathbf{P}_2^{1/2}$. Thus we have

$$\begin{aligned} F(\mathbf{P}_1, \mathbf{P}_2) &= \mathrm{Tr}(\mathbf{P}_1^{1/2}\mathbf{P}_2\mathbf{P}_1^{1/2})^{1/2} \\ &= \sum \lambda_m^{1/2} \\ &= \mathrm{Tr}(\mathbf{P}_2^{1/2}\mathbf{P}_1\mathbf{P}_2^{1/2})^{1/2} \\ &= F(\mathbf{P}_2, \mathbf{P}_1). \end{aligned} \tag{3.60}$$

Therefore, $d_{R_1}(\mathbf{P}_1, \mathbf{P}_2)$ is symmetric.

c) **Triangle inequality:** Let $\mathbf{P}_1$, $\mathbf{P}_2$ and $\mathbf{P}_3$ be three points in $\mathcal{M}$ such that $\mathbf{P}_2$ is not on the geodesic curve c_3 connecting $\mathbf{P}_1$ and $\mathbf{P}_3$. Let c_1 be a curve connecting $\mathbf{P}_1$ and $\mathbf{P}_2$ on $\mathcal{M}$, and c_2 be a curve connecting $\mathbf{P}_2$ and $\mathbf{P}_3$ on $\mathcal{M}$, respectively, as is illustrated in Figure 3.6. Then the composite curve c_1c_2 must be different

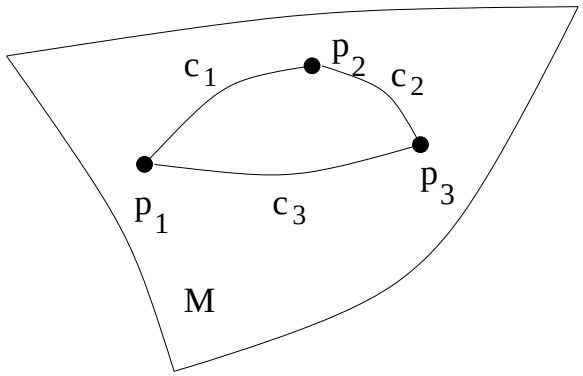


Figure 3.6: Illustration of triangle inequality

from c_3 and must connect $\mathbf{P}_1$ to $\mathbf{P}_2$ and then $\mathbf{P}_2$ to $\mathbf{P}_3$ on $\mathcal{M}$. Therefore, the length of c_1c_2 must equal to the sum of the lengths of c_1 and c_2. Now, since c_3 is the geodesic curve connecting $\mathbf{P}_1$ and $\mathbf{P}_3$, its length is the minimum between

$\mathbf{P}_1$ and $\mathbf{P}_3$ on $\mathcal{M}$. Let $l(\cdot)$ denote the length of a curve on $\mathcal{M}$. Then we have

$$\begin{aligned} d_{R_1}(\mathbf{P}_1, \mathbf{P}_3) &= l(c) \\ &\leq \min_{c_1 c_2} l(c_1 c_2) \\ &= \min_{c_1} l(c_1) + \min_{c_2} l(c_2) \\ &= d_{R_1}(\mathbf{P}_1, \mathbf{P}_2) + d_{R_1}(\mathbf{P}_2, \mathbf{P}_3) \end{aligned} \tag{3.61}$$

Similarly, we can show that $d_{R_1 W}$ of Eq. (3.52) is also a true distance measure. □

3.4.2 The Riemannian distance d_{R_1} in the special case of single sensor measurements

In the case when only one channel of the EEG signals is available, then the normalized power spectrum (at n frequency points) is usually adopted as the characterization of the EEG signals. Consider the set of normalized power spectrum in the form:

$$\mathcal{P} = \left\{ \mathbf{p} \in \mathbb{R}^n : \sum_{i=1}^{n} p_i = 1, p_i > 0 \right\} \tag{3.62}$$

We can apply the Riemannian distance d_{R_1} to measure the dissimilarity between two normalized power spectral densities $\mathbf{p}, \mathbf{q} \in \mathcal{P}$.

- **Case 1:** Let

$$\mathcal{M} = \left\{ \mathbf{P} : \mathbf{P} = \mathrm{diag}[p_1, \cdots, p_n], \mathbf{p} = [p_1, \cdots, p_n]^T \in \mathcal{P} \right\} \tag{3.63}$$

For another power spectrum $\mathbf{q} \in \mathcal{P}$, we similarly form $\mathbf{Q} \in \mathcal{M}$. Then, the Riemannian distance between $\mathbf{P}$ and $\mathbf{Q}$ is given by

$$\begin{aligned} d_{R_1}(\mathbf{P}, \mathbf{Q}) &= \sqrt{2 - 2\mathrm{Tr}\sqrt{\mathbf{P}^{1/2}\mathbf{Q}\mathbf{P}^{1/2}}} \\ &= \sqrt{2 - 2\sum_{i=1}^{n} \sqrt{p_i q_i}} \end{aligned} \tag{3.64}$$

- **Case 2:** Let

$$\mathcal{M} = \left\{ \mathbf{P} : \mathbf{P} = \mathbf{p}\mathbf{p}^T, \mathbf{p} \in \mathcal{P} \right\} \tag{3.65}$$

Then,

$$\mathbf{P}^{1/2} = (\mathbf{p}\mathbf{p}^T)^{1/2} = (\mathbf{p}\mathbf{p}^T\mathbf{p}\mathbf{p}^T)^{1/2} = \mathbf{p}\mathbf{p}^T = \mathbf{P} \tag{3.66}$$

Thus, we have

$$Tr(\mathbf{P}^{1/2}) = \mathrm{Tr}(\mathbf{P}) = \mathbf{p}^T\mathbf{p} = 1 \tag{3.67}$$

For another power spectrum $\mathbf{q} \in \mathcal{P}$, we similarly form $\mathbf{Q} \in \mathcal{M}$ and we have

$$\mathbf{Q}^{1/2} = \mathbf{Q} \ \texttt{and} \ \mathrm{Tr}(\mathbf{Q}^{1/2}) = 1. \tag{3.68}$$

Thus, the Riemannian distance between $\mathbf{P}$ and $\mathbf{Q}$ is given by Similarly

$$\begin{aligned} d_{R_1}(\mathbf{P}, \mathbf{Q}) &= \sqrt{2 - 2\mathrm{Tr}(\mathbf{P}^{1/2}\mathbf{Q}\mathbf{P}^{1/2})^{1/2}} \\ &= \sqrt{2 - 2\mathrm{Tr}(\mathbf{P}\mathbf{Q}\mathbf{P})^{1/2}} \\ &= \sqrt{2 - 2\mathrm{Tr}(\mathbf{p}\mathbf{p}^T\mathbf{q}\mathbf{q}^T\mathbf{p}\mathbf{p}^T)^{1/2}} \\ &= \sqrt{2 - 2|\mathbf{p}^T\mathbf{q}|\mathrm{Tr}(\mathbf{P}^{1/2})} \\ &= \sqrt{2 - 2|\mathbf{p}^T\mathbf{q}|}. \end{aligned} \tag{3.69}$$

We noted that for Case 1, the second term under the square-root sign in Eq. (3.64) is the argument of the Fisher-Rao distance in Eq. (3.4), and that for Case 2, the second term under the square-root sign in Eq. (3.69) is the argument of the normalized Fubini-Study distance in Eq. (3.5). Indeed, the distance measures in both Eqs. (3.64) and (3.69) are of the same form as the correlation distance in Eq. (3.6). Thus, the Riemannian distance d_{R_1} can be viewed as a generalization of the correlation distance established for single-channel measurements.

3.4.3 Alternative derivation of the Riemannian distance d_{R_1}

In Section 3.2.1, we have introduced the Fréchet distance between two probability distributions, more especially, for two zero-mean Gaussian distributions with covariance matrices $\mathbf{R}_1$ and $\mathbf{R}_2$. In this section, we start from an analog of the Fréchet distance and apply to two PSD matrices, we obtain the following result:

Theorem 3.3 *The Fréchet distance between two PSD matrices* $\mathbf{P}_1$ *and* $\mathbf{P}_2$ *is*

$$d_{\mathrm{Fe}}(\mathbf{P}_1, \mathbf{P}_2) = \sqrt{\mathrm{Tr}(\mathbf{P}_1 + \mathbf{P}_2 - 2(\mathbf{P}_1\mathbf{P}_2)^{1/2})} \tag{3.70}$$

Proof. Since $\mathbf{P}_1$ and $\mathbf{P}_1$ are nonnegative, we have

$$\mathbf{P}_1 = \sum_{i=1}^{M} \mathbf{v}_{1i}\mathbf{v}_{1i}^H \tag{3.71}$$

and

$$\mathbf{P}_2 = \sum_{i=1}^{M} \mathbf{v}_{2i}\mathbf{v}_{2i}^H \tag{3.72}$$

for some vectors $\mathbf{v}_{1i}$ and $\mathbf{v}_{2i}$. Now, following the definition of the Fréchet distance in Eq. (3.12), we define the distance between $\mathbf{P}_1$ and $\mathbf{P}_2$ as

$$d_{\mathrm{Fe}}(\mathbf{P}_1, \mathbf{P}_2) = \sqrt{\min_{\mathbf{v}_{1i}, \mathbf{v}_{2i}} \sum_{i=1}^{M} \parallel \mathbf{v}_{1i} - \mathbf{v}_{2i} \parallel^2} \tag{3.73}$$

Let

$$\mathbf{P}_{12} = \sum_{i=1}^{M} \mathbf{v}_{1i}\mathbf{v}_{2i}^H \tag{3.74}$$

Since

$$\begin{aligned}
\min_{\mathbf{v}_{1i}, \mathbf{v}_{2i}} \sum_{i=1}^{M} \parallel \mathbf{v}_{1i} - \mathbf{v}_{2i} \parallel^2 &= \min_{\mathbf{v}_{1i}, \mathbf{v}_{2i}} \sum_{i=1}^{M} (\mathbf{v}_{1i} - \mathbf{v}_{2i})^H (\mathbf{v}_{1i} - \mathbf{v}_{2i}) \\
&= \min_{\mathbf{v}_{1i}, \mathbf{v}_{2i}} \sum_{i=1}^{M} \mathrm{Tr}(\mathbf{v}_{1\mathrm{i}} - \mathbf{v}_{2\mathrm{i}})(\mathbf{v}_{1\mathrm{i}} - \mathbf{v}_{2\mathrm{i}})^{\mathrm{H}} \\
&= \min_{\mathbf{v}_{1i}, \mathbf{v}_{2i}} \mathrm{Tr}(\mathbf{P}_1 + \mathbf{P}_2 - \mathbf{P}_{12} - \mathbf{P}_{21}) \\
&= \mathrm{Tr}(\mathbf{P}_1 + \mathbf{P}_2) - \max_{\mathbf{v}_{1\mathrm{i}}, \mathbf{v}_{2\mathrm{i}}} \mathrm{Tr}(\mathbf{P}_{12} + \mathbf{P}_{21})
\end{aligned} \tag{3.75}$$

therefore, we need to solve the following problem

$$\begin{aligned} \max_{\mathbf{v}_{1i},\mathbf{v}_{2i}} \quad & \mathrm{Tr}(\mathbf{P}_{12}+\mathbf{P}_{21}) = \mathrm{Tr}\sum_{\mathrm{i=1}}^{\mathrm{M}}(\mathbf{v}_{1\mathrm{i}}\mathbf{v}_{2\mathrm{i}}^{\mathrm{H}}+\mathbf{v}_{2\mathrm{i}}\mathbf{v}_{1\mathrm{i}}^{\mathrm{H}}) \\ \text{s.t.} \quad & \mathbf{P}_1 = \sum_{i=1}^{M}\mathbf{v}_{1i}\mathbf{v}_{1i}^H \\ & \mathbf{P}_2 = \sum_{i=1}^{M}\mathbf{v}_{2i}\mathbf{v}_{2i}^H \end{aligned} \tag{3.76}$$

The Lagrangian for the above maximization problem is given by

$$\mathcal{L} = \mathrm{Tr}\sum_{i=1}^{M}(\mathbf{v}_{1i}\mathbf{v}_{2i}^H+\mathbf{v}_{2i}\mathbf{v}_{1i}^H) + \mathrm{Tr}\Big[\Big(\sum_{i=1}^{M}\mathbf{v}_{1i}\mathbf{v}_{1i}^H\Big)\mathbf{\Lambda}_1\Big] + \mathrm{Tr}\Big[\Big(\sum_{i=1}^{M}\mathbf{v}_{2i}\mathbf{v}_{2i}^H\Big)\mathbf{\Lambda}_2\Big] \tag{3.77}$$

where $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ are the Lagrange multipliers. Taking derivatives and let the result equal to zero, we have

$$\mathbf{v}_{1i} = \mathbf{\Lambda}_2\mathbf{v}_{2i} \quad \text{and} \quad \mathbf{v}_{2i} = \mathbf{\Lambda}_1\mathbf{v}_{1i} \tag{3.78}$$

Noting that $\mathbf{\Lambda}_1$ is Hermitian, we have

$$\mathbf{P}_2 = \sum_{i=1}^{M}\mathbf{v}_{2i}\mathbf{v}_{2i}^H = \sum_{i=1}^{M}\mathbf{\Lambda}_1\mathbf{v}_{1i}\mathbf{v}_{1i}^H\mathbf{\Lambda}_1 = \mathbf{\Lambda}_1\mathbf{P}_1\mathbf{\Lambda}_1 \tag{3.79}$$

and

$$\mathbf{P}_{12} = \sum_{i=1}^{M}\mathbf{v}_{1i}\mathbf{v}_{2i}^H = \sum_{i=1}^{M}\mathbf{v}_{1i}\mathbf{v}_{1i}^H\mathbf{\Lambda}_1 = \mathbf{P}_1\mathbf{\Lambda}_1 \tag{3.80}$$

Thus,

$$\mathbf{P}_{12}^2 = \mathbf{P}_1(\mathbf{\Lambda}_1\mathbf{P}_1\mathbf{\Lambda}_1) = \mathbf{P}_1\mathbf{P}_2 \tag{3.81}$$

Clearly $\mathbf{P}_1\mathbf{P}_2$ is nonnegative definite since $\mathbf{P}_1$ and $\mathbf{P}_2$ are nonnegative definite. Therefore its square root exits and is nonnegative definite, i.e.,

$$\mathbf{P}_{12} = (\mathbf{P}_1\mathbf{P}_2)^{1/2} \tag{3.82}$$

We also note that $\mathbf{P}_{21} = \mathbf{P}_{12}^H$. Thus

$$\max \mathrm{Tr}\sum_{i=1}^{M}(\mathbf{v}_{1i}\mathbf{v}_{2i}^H+\mathbf{v}_{2i}\mathbf{v}_{1i}^H) = 2\mathrm{Tr}(\mathbf{P}_1\mathbf{P}_2)^{1/2} \tag{3.83}$$

Putting this in Eq. (3.75) and taking the square root we obtain the Fréchet distance of Eq. (3.70). □

It is interesting that the distance developed in Theorem 3.3 is exactly the same as the Riemannian distance d_{R_1}, i.e., we have the following equivalence:

Assertion 3.1 *The Riemannian distance d_{R_1} of Eq. (3.47) and the Fréchet distance d_{Fe} of Eq. (3.70) are the same.*

Proof. Comparing Eqs. (3.47) and (3.70), clearly, it is sufficient to show that

$$\mathrm{Tr}(\mathbf{P}_1^{1/2}\mathbf{P}_2\mathbf{P}_1^{1/2}) = \mathrm{Tr}(\mathbf{P}_1\mathbf{P}_2) \tag{3.84}$$

In other words, we only need to show that $\mathbf{P}_1^{1/2}\mathbf{P}_2\mathbf{P}_1^{1/2}$ and $\mathbf{P}_1\mathbf{P}_2$ have the same eigenvalues since the trace of a positive-definite Hermitian matrix is equal to the sum of its eigenvalues. Let λ_i and $\mathbf{u}_i$ be the eigenvalue and eigenvector pair of $\mathbf{P}_1^{1/2}\mathbf{P}_2\mathbf{P}_1^{1/2}$. Then we have

$$\mathbf{P}_1^{1/2}\mathbf{P}_2\mathbf{P}_1^{1/2}\mathbf{u}_i = \lambda_i\mathbf{u}_i \tag{3.85}$$

Multiplying by $\mathbf{P}_1^{1/2}$ on both sides we have

$$\mathbf{P}_1\mathbf{P}_2(\mathbf{P}_1^{1/2}\mathbf{u}_i) = \lambda_i(\mathbf{P}_1^{1/2}\mathbf{u}_i) \tag{3.86}$$

Therefore, $\mathbf{P}_1^{1/2}\mathbf{P}_2\mathbf{P}_1^{1/2}$ and $\mathbf{P}_1\mathbf{P}_2$ have the same eigenvalues, and Eqs. (3.47) and (3.70) are equal. □

Remarks: Even though it is possible that the Riemannian distance d_{R_1} can be obtained by mimicking the Fréchet distance between covariance matrices, the derivation does not show that it is an intrinsic distance on the manifold $\mathcal{M}$. More importantly, this alternative derivation of d_{R_1} cannot be used to develop other Riemannian distances. By considering Riemannian metrics on the manifold as developed in Section 3.4.1, different Riemannian distances can be developed as shown in the ensuing sections.

3.4.4 Riemannian distance d_{R_2}

In this section, we employ a parallel procedure to that described in the previous section and develop another type of Riemannian distances.[8] By endowing a different Riemannian metric, we are able to arrive at a different Riemannian distance for the manifold of PSD matrices $\mathcal{M}$. First, let $\mathcal{H}_H = \{\mathbf{P} : \mathbf{P}^H = \mathbf{P}, \mathbf{P} \in \mathcal{M}_M\}$. we endow $\mathcal{H}_H$ with an inner product $\langle \tilde{\mathbf{X}}, \tilde{\mathbf{Y}} \rangle = \mathrm{Tr}(\tilde{\mathbf{X}}\tilde{\mathbf{Y}})$, $\tilde{\mathbf{X}}, \tilde{\mathbf{X}} \in \mathcal{H}_H$ (Note that this is the induced inner product by restriction of the inner product endowed to $\mathcal{H}_M$), so that $\mathcal{H}_H$ is a Hilbert space denoted by $(\mathcal{H}_H, \langle \cdot, \cdot \rangle)$. Let $\tilde{\mathcal{H}} = \{\tilde{\mathbf{P}} : \tilde{\mathbf{P}}^2 = \mathbf{P} \in \mathcal{M}\}$. Then, we have $\tilde{\mathcal{H}} \subset \mathcal{H}_M$. Since any $\mathbf{P} \in \mathcal{M}$ is positive definite Hermitian, the corresponding $\tilde{\mathbf{P}} \in \tilde{\mathcal{H}}$ is also positive definite Hermitian. [9] Then, we have the following theorem:

Theorem 3.4 *Let $(\mathcal{M}, g_P)$ be the Riemannian manifold having a Riemannian metric given by*

$$g_P(\mathbf{A}, \mathbf{B}) = \langle \mathbf{A}, \mathbf{K} \rangle \tag{3.87}$$

such that

$$\mathbf{PK} + \mathbf{KP} + 2\tilde{\mathbf{P}}\mathbf{K}\tilde{\mathbf{P}} = \mathbf{B} \tag{3.88}$$

where $\tilde{\mathbf{P}}^2 = \mathbf{P} \in \mathcal{M}$, $\mathbf{A}, \mathbf{B} \in \mathcal{T}_{\mathcal{M}}(\mathbf{P})$. Then the geodesic distance between $\mathbf{P}_1$ and $\mathbf{P}_2$ on $\mathcal{M}$ is

$$d_{R_2}(\mathbf{P}_1, \mathbf{P}_2) = \sqrt{\mathrm{Tr}(\tilde{\mathbf{P}}_1 - \tilde{\mathbf{P}}_2)^2} \tag{3.89}$$

where $\mathbf{P}_1 = \tilde{\mathbf{P}}_1^2$ and $\mathbf{P}_2 = \tilde{\mathbf{P}}_2^2$.

[8] The distance d_{R_2} was initially a conjecture proposed by Dr. K.M. Wong. Here we show that it is also a Riemannian distance.

[9] The space $\tilde{\mathcal{H}}$ here is in fact the same as $\mathcal{M}$ equipped with an inner product. Indeed, for $\mathbf{P} = \tilde{\mathbf{P}}^2$, $\mathbf{P}$ and $\tilde{\mathbf{P}}$ are in the same space since both are positive definite Hermitian matrices. Here, we show that the Euclidian distance between $\tilde{\mathbf{P}}_1$ and $\tilde{\mathbf{P}}_2$ in $\mathcal{M}$ is equal to the Riemannian distance between $\mathbf{P}_1$ and $\mathbf{P}_2$ in the same space.

Proof. Let $\mathbf{\Gamma}(r) : (-\epsilon, \epsilon) \to \mathcal{M}$ be a curve in $\mathcal{M}$ such that $\mathbf{\Gamma}(0) = \mathbf{P} \in \mathcal{M}$ and $\tilde{\mathbf{\Gamma}}(r) : (-\epsilon, \epsilon) \to \tilde{\mathcal{H}}$ be a curve in $\tilde{\mathcal{H}}$ such that $\tilde{\mathbf{\Gamma}}(0) = \tilde{\mathbf{P}} \in \tilde{\mathcal{H}}$ with

$$\mathbf{\Gamma}(r) = \tilde{\mathbf{\Gamma}}(r)\tilde{\mathbf{\Gamma}}(r) \tag{3.90}$$

Taking the derivative of Eq. (3.90) with respect to r on both sides, we have

$$\dot{\mathbf{\Gamma}}(r) = \dot{\tilde{\mathbf{\Gamma}}}(r)\tilde{\mathbf{\Gamma}}(r) + \tilde{\mathbf{\Gamma}}(r)\dot{\tilde{\mathbf{\Gamma}}}(r) \tag{3.91}$$

Let $\dot{\mathbf{P}} = \dot{\mathbf{\Gamma}}(r)\big|_{r=0}$ and $\dot{\tilde{\mathbf{P}}} = \dot{\tilde{\mathbf{\Gamma}}}(r)\big|_{r=0}$. Then, at $r = 0$ we have

$$\dot{\mathbf{P}} = \dot{\tilde{\mathbf{P}}}\tilde{\mathbf{P}} + \tilde{\mathbf{P}}\dot{\tilde{\mathbf{P}}} \tag{3.92}$$

Since $\dot{\mathbf{P}} \in \mathcal{T}_{\mathcal{M}}(\mathbf{P})$ and $\dot{\tilde{\mathbf{P}}} \in \mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$, for $\mathbf{A}, \mathbf{B} \in \mathcal{T}_{\mathcal{M}}(\mathbf{P})$ and the corresponding $\tilde{\mathbf{A}}, \tilde{\mathbf{B}} \in \mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ we have

$$\mathbf{A} = \tilde{\mathbf{A}}\tilde{\mathbf{P}} + \tilde{\mathbf{P}}\tilde{\mathbf{A}} \tag{3.93}$$

and

$$\mathbf{B} = \tilde{\mathbf{B}}\tilde{\mathbf{P}} + \tilde{\mathbf{P}}\tilde{\mathbf{B}} \tag{3.94}$$

by applying Eq. (3.92).

For a given $\tilde{\mathbf{P}} \in \tilde{\mathcal{H}}$, we define an operator $X_{\tilde{\mathbf{P}}}$ on $\mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ as

$$X_{\tilde{\mathbf{P}}}\tilde{\mathbf{A}} = \tilde{\mathbf{A}}\tilde{\mathbf{P}} + \tilde{\mathbf{P}}\tilde{\mathbf{A}} \tag{3.95}$$

where $\tilde{\mathbf{A}} \in \mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$. Then Eq. (3.93) and Eq. (3.94) can be rewritten as

$$\mathbf{A} = X_{\tilde{\mathbf{P}}}\tilde{\mathbf{A}} \tag{3.96}$$

and

$$\mathbf{B} = X_{\tilde{\mathbf{P}}}\tilde{\mathbf{B}} \tag{3.97}$$

For any $\tilde{\mathbf{P}} \in \tilde{\mathcal{H}}$, $\mathcal{T}_{\mathcal{H}_H}(\tilde{\mathbf{P}}) = \mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ and the unique metric endowed to $\mathcal{T}_{\mathcal{H}}(\tilde{P})$ is the same inner product endowed to $\mathcal{H}_H$. Since $\tilde{\mathbf{P}}$ is Hermitian, the operator $X_{\tilde{\mathbf{P}}}$ must be Hermitian, so is its inverse $X_{\tilde{\mathbf{P}}}^{-1}$. Thus, we have

$$\langle \tilde{\mathbf{A}}, \tilde{\mathbf{B}} \rangle = \langle X_{\tilde{\mathbf{P}}}^{-1}\mathbf{A}, X_{\tilde{\mathbf{P}}}^{-1}\mathbf{B} \rangle = \langle \mathbf{A}, X_{\tilde{\mathbf{P}}}^{-2}\mathbf{B} \rangle = \langle \mathbf{A}, \mathbf{K} \rangle \tag{3.98}$$

where $X_{\tilde{\mathbf{P}}}^{-1}$ is the inverse of $X_{\tilde{\mathbf{P}}}$, and $\mathbf{K} = X_{\tilde{\mathbf{P}}}^{-2}\mathbf{B}$.
On the other hand, we have

$$\begin{aligned} \mathbf{B} &= X_{\tilde{\mathbf{P}}}^2\mathbf{K} = X_{\tilde{\mathbf{P}}}(X_{\tilde{\mathbf{P}}}\mathbf{K}) = X_{\tilde{\mathbf{P}}}(\mathbf{K}\tilde{\mathbf{P}} + \tilde{\mathbf{P}}\mathbf{K}) \\ &= (\mathbf{K}\tilde{\mathbf{P}} + \tilde{\mathbf{P}}\mathbf{K})\tilde{\mathbf{P}} + \tilde{\mathbf{P}}(\mathbf{K}\tilde{\mathbf{P}} + \tilde{\mathbf{P}}\mathbf{K}) \\ &= \mathbf{P}\mathbf{K} + \mathbf{K}\mathbf{P} + 2\tilde{\mathbf{P}}\mathbf{K}\tilde{\mathbf{P}} \end{aligned} \tag{3.99}$$

Therefore, the metrics for $\mathcal{T}_{\mathcal{M}}(\mathbf{P})$ and $\mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ are the same, i.e., $\mathcal{T}_{\mathcal{M}}(\mathbf{P})$ and $\mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ are isometric. Thus, the mapping $\pi : \tilde{\mathcal{H}} \to \mathcal{M}$ such that $\pi(\tilde{\mathbf{P}}) = \mathbf{P}$ is an isometry between $\tilde{\mathcal{H}}$ and $\mathcal{M}$. As a result, the length of a geodesic connecting $\mathbf{P}_1$ and $\mathbf{P}_2$ in $\mathcal{M}$ has the same length of the geodesic connecting $\tilde{\mathbf{P}}_1$ and $\tilde{\mathbf{P}}_2$ in $\tilde{\mathcal{H}}$ (it is also the geodesic in $\mathcal{H}_H$). Since the geodesic for two points in $\mathcal{H}_H$ is measured along the straight line between the two points, we have

$$d_{R_2}(\mathbf{P}_1, \mathbf{P}_2) = \left\| \tilde{\mathbf{P}}_1 - \tilde{\mathbf{P}}_2 \right\|_2 = \sqrt{\mathrm{Tr}(\tilde{\mathbf{P}}_1 - \tilde{\mathbf{P}}_2)^2} \tag{3.100}$$

Since the PSD matrices $\mathbf{P}$ are positive definite Hermitian, $\tilde{\mathbf{P}} = \sqrt{\mathbf{P}}$ for $\mathbf{P} \in \mathcal{M}$. Then Eq. (3.100) can be written as:

$$d_{R_2}(\mathbf{P}_1, \mathbf{P}_2) = \sqrt{\mathrm{Tr}\mathbf{P}_1 + \mathrm{Tr}\mathbf{P}_2 - 2\mathrm{Tr}\sqrt{\mathbf{P}_1}\sqrt{\mathbf{P}_2}} \tag{3.101}$$

□

If desired, it is straightforward to bestow a weighting for the Riemannian distance d_{R_2} such that the weighted distance is

$$\begin{aligned} d_{R_2W}(\mathbf{P}_1, \mathbf{P}_2) &= \sqrt{\mathrm{Tr}\Big(\sqrt{\mathbf{P}_1} - \sqrt{\mathbf{P}_2}\Big)\mathbf{W}\Big(\sqrt{\mathbf{P}_1} - \sqrt{\mathbf{P}_2}\Big)} \\ &= \sqrt{\mathrm{Tr}\mathbf{W}\mathbf{P}_1 + \mathbf{W}\mathbf{P}_2 - \mathrm{Tr}\mathbf{W}\sqrt{\mathbf{P}_1}\sqrt{\mathbf{P}_2} - \mathrm{Tr}\mathbf{W}\sqrt{\mathbf{P}_2}\sqrt{\mathbf{P}_1}} \end{aligned} \tag{3.102}$$

where $\mathbf{W} = \mathbf{\Omega}^H\mathbf{\Omega} \succ 0$ be a real positive definite weighting matrix.

The above Riemannian distances d_{R_1} and d_{R_2} have been developed from the isometry of two spaces for the classification of EEG signals. As we mentioned before, there exist possibly an infinite number of Riemannian metrics. Some Riemannian metrics may similarly lead to explicit formulas for the Riemannian distances. The following is another well-known example [17]. Since the original development of the measure does not follow the geometric view proposed in this thesis, we include our own proof in Appendix F.

3.4.5 Riemannian distance d_{R_3}

Theorem 3.5 *Let $\mathcal{M}$ be the space of positive definite Hermitian matrices. If it is endowed with a Riemannian metric*

$$g_P(\mathbf{A}, \mathbf{B}) = Tr\mathbf{P}^{-1}\mathbf{A}_\mathbf{P}\mathbf{P}^{-1}\mathbf{B}_\mathbf{P}, \tag{3.103}$$

where $\mathbf{A}_\mathbf{P}, \mathbf{B}_\mathbf{P} \in T_\mathcal{M}(\mathbf{P})$, then the geodesic distance between $\mathbf{P}_1$ and $\mathbf{P}_2$ in $\mathcal{M}$ is

$$\begin{aligned} d_{R_3}(\mathbf{P}_1, \mathbf{P}_2) = &= \sqrt{Tr(\log \mathbf{P}_1^{-1/2}\mathbf{P}_2\mathbf{P}_1^{-1/2})^2} \\ &= \sqrt{\sum_{i=1}^{n} \log^2 \lambda_i} \end{aligned} \tag{3.104}$$

where λ_i are the eigenvalues of $\mathbf{P}_1^{-1}\mathbf{P}_2$.

Proof. See Appendix F. □

The distance axioms can be verified easily for d_{R_3}. We omit the verifications. An important properties of this distance is that it is weighting invariant, i.e., $d_{R_3}(\mathbf{P}_1, \mathbf{P}_2) = d_{R_3 W}(\mathbf{P}_1, \mathbf{P}_2)$ if $\mathbf{P}_1$ and $\mathbf{P}_2$ are weighted as in the previous sections. This fact has been shown in the proof of the theorem. Therefore, for enhancement of similarity and dissimilarity in EEG classification, d_{R_3} is not an appropriate choice.

3.4.6 Summary on the Riemannian distances

The development of the Riemannian distances $d_{R_1}, d_{R_2}, d_{R_3}$ in Sections 3.4.1, 3.4.4 and 3.4.5 follow a common course which can be summarized as follows:

1. We start with the space $\mathcal{H}_M$ (or $\mathcal{H}_H$), the Hilbert space formed by all the $M \times M$ complex matrices (or the $M \times M$ Hermitian matrices) equipped with the Hilbert-Schmidt inner product, and $\mathcal{M}$, a subset of $\mathcal{H}_M$ (or $\mathcal{H}_H$), containing all the PSD matrices $\{\mathbf{P}_i\}$.

2. We create a mapping $\pi(\tilde{\mathbf{P}}) = \mathbf{P} \in \mathcal{M}$ which maps the subset $\{\tilde{\mathbf{P}}\} \in \mathcal{H}_M$ (or $\{\tilde{\mathbf{P}}\} \in \mathcal{H}_H$) to $\mathcal{M}$. We denote the subset $\{\tilde{\mathbf{P}}\}$ by $\tilde{\mathcal{H}}$. Note that $\tilde{\mathcal{H}}$ is equipped with the Hilbert-Schmidt inner product and is a subset of $\mathcal{H}_M$ (or $\mathcal{H}_H$) not necessarily a complete space on its own.

3. Since the mappings created in the three cases are different, the resulting subsets $\tilde{\mathcal{H}}$ are also different. Specifically,

   (i) The mapping π for d_{R_1} results in a Riemannian submersion such that a one-to-one mapping emerges from the horizontal lift of the tangent space $\mathcal{T}_{\mathcal{M}}(\mathbf{P})$ on $\mathcal{M}$ to the tangent space $\mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ on $\tilde{\mathcal{H}}$ and an isometry between the two is established.

(ii) The mapping π for d_{R_2} results in a subset $\tilde{\mathcal{H}}$ which is the same subset as $\mathcal{M}$ of PSD matrices since $\tilde{\mathbf{P}} = \sqrt{\mathbf{P}}$ is also positive definite. Here, the inverse mapping π^{-1} is unique, and $\tilde{\mathbf{P}}$ remains in $\mathcal{M}$, we need not apply "the horizontal lift". Furthermore, $\tilde{\mathcal{H}}$, equipped with the Hilbert-Schmidt inner product, is directly isometric to $\mathcal{M}$.

(iii) The mapping π for d_{R_3} results in a subset $\tilde{\mathcal{H}}$ different from, but isometric to $\mathcal{M}$. The image $\tilde{\mathbf{P}}$ of $\mathbf{P}$ in π^{-1} is also unique.

4. In each of the three case, the subset $\tilde{\mathcal{H}}$ resulted from the mapping π is either a Riemannian submersion or is isometric to $\mathcal{M}$. This greatly facilitates the evaluation of the geodesic between two points in $\mathcal{M}$ since the geodesic between two points $\mathbf{P}_1$ and $\mathbf{P}_2$ can be evaluated by the equivalent Euclidean distance between the two image points $\tilde{\mathbf{P}}_1$ and $\tilde{\mathbf{P}}_2$. This procedure establishes the three Riemannian distances and is illustrated in Figs. 3.7, 3.8 and 3.9 where c_1 is the geodesic curve connecting two points $\mathbf{P}_1$ and $\mathbf{P}_2$ in $\mathcal{M}$, and c_2 is the Euclidean distance connecting $\tilde{\mathbf{P}}_1$ and $\tilde{\mathbf{P}}_2$ in $\tilde{\mathcal{H}}$, where $\tilde{\mathbf{P}}_1$ and $\tilde{\mathbf{P}}_2$ are the lifts of $\mathbf{P}_1$ and $\mathbf{P}_2$ through π^{-1} respectively.

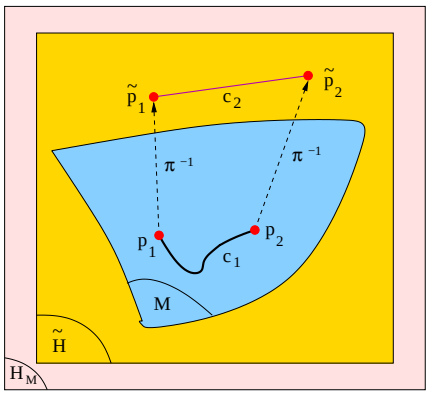


Figure 3.7: Illustration of geodesics in $\mathcal{M}$ and $\tilde{\mathcal{H}}$ for d_{R_1}

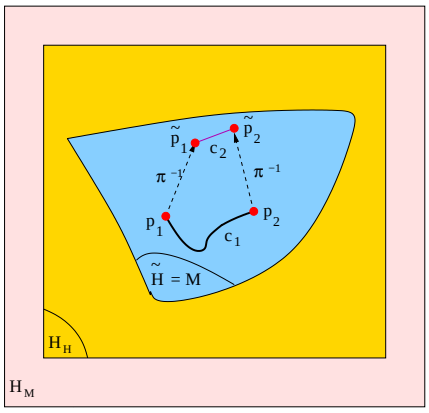


Figure 3.8: Illustration of geodesics in $\mathcal{M}$ and $\tilde{\mathcal{H}}$ for d_{R_2}

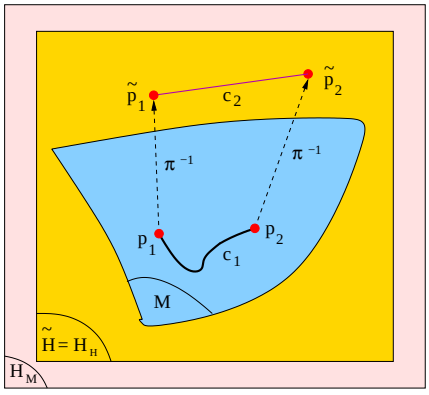


Figure 3.9: Illustration of geodesics in $\mathcal{M}$ and $\tilde{\mathcal{H}}$ for d_{R_3}

3.5 Dissimilarity measures

After representing EEG signals as curves on Riemannian manifold, we are ready to define dissimilarity measure with the use of Riemannian distances (geodesic distances):

Power spectral density is a function of the frequency ω. With the variation of ω, the PSD matrix describes a curve on the Riemannian manifold $\mathcal{M}$. Therefore, similarity/dissimilarity between two curves of PSD matrices corresponding to two multi-channel EEG signals must be established. Now, in the previous sections, we have established geodesic functions d_G between two points for the manifold $\mathcal{M}$. For two curves on the manifold described by two power density functions $\mathbf{P}_1(\omega)$ and $\mathbf{P}_2(\omega)$, due to variation of the frequency variable ω, d_G can be thought of as a non-negative real valued function of ω measuring the distance between the two curves at the frequency ω, i.e.,

$$d_G(\omega) = d_G(\mathbf{P}_1(\omega), \mathbf{P}_2(\omega)) \tag{3.105}$$

At each frequency ω_k it measures the dissimilarity between the two corresponding power spectral density matrices $\mathbf{P}_1(\omega_k)$ and $\mathbf{P}_2(\omega_k)$ on the manifold $\mathcal{M}$. As the frequency ω varies, we can define the distance between the curves $\mathbf{P}_1(\omega)$ and $\mathbf{P}_2(\omega)$ as the integral of d_G with respect to ω such that

$$d(\mathbf{P}_1(\omega), \mathbf{P}_2(\omega)) = \int_{\omega_1}^{\omega_2} d_G(\omega) d\omega \tag{3.106}$$

It is easy to show this Riemann integral satisfies the axioms of a distance function, and can be approximated as

$$\begin{aligned} d(\mathbf{P}_1(\omega), \mathbf{P}_2(\omega)) &\approx \sum_i d_G(\omega_i)\Delta\omega_i \\ &= \sum_i d_G(\mathbf{P}_1(\omega_i), \mathbf{P}_2(\omega_i))\Delta\omega_i \end{aligned} \tag{3.107}$$

If equal frequency increment is used, i.e., $\Delta\omega_i = c$, a constant, then without loss of

generality we can define the *dissimilarity* between the two PSD curves as

$$d(\mathbf{P}_1(\omega), \mathbf{P}_2(\omega)) = \sum_i d_G(\mathbf{P}_1(\omega_i), \mathbf{P}_2(\omega_i)) \tag{3.108}$$

Figure 3.10 is an illustration of the geodesic distance between any two points $\mathbf{P}_1$ and $\mathbf{P}_2$ on the manifold $\mathcal{M}$. Figure 3.11 illustrates the dissimilarity measure between two curves corresponding two EEG signals. Clearly, different geodesic distance d_G gives

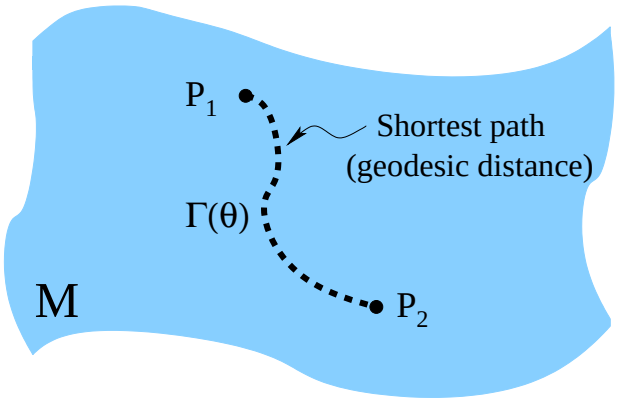


Figure 3.10: Geodesic distance between two power spectral density matrices

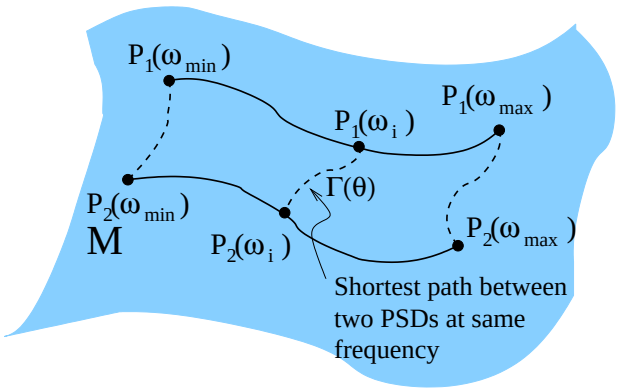


Figure 3.11: Dissimilarity measure between two EEG signals

rise to different dissimilarity measures between PSD matrices.

Chapter 4

Optimally Weighted Distances for Similarity/Dissimilarity

In Chapters 2 and 3, we have seen that the EEG signals and their features can be treated as vectors in a linear space fitted with certain structure and a distance measure to describe the relationship between them. In particular, the PSD matrices which are the selected feature of the EEG signals, can be looked upon as describing a Riemannian manifold on which the geodesic is the distance measure. We have also seen that there are various approaches in formulating both the distance in a vector space and the geodesic on a manifold, resulting in different distance measures and geodesics. Furthermore, these distances, whether they measure the distance in a linear space or the distance on a Riemannian manifold, may be weighted to enhance certain characteristics of the data so as to facilitate EEG signal classification. Here in the present chapter, we will examine the various ways of obtaining the weighting matrix to serve the final goal of classification.

4.1 Distance metric learning

If the representation of an EEG signal is treated as an abstract object such as a point in a space, then a simple illustration of class separation is shown in Figure 4.1. The idea is that one might expect that the set of points representing different events

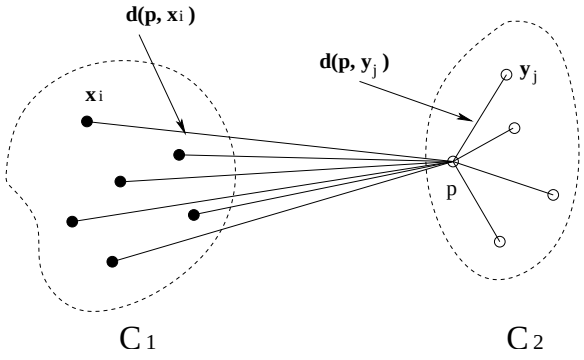


Figure 4.1: Illustration of class separation

that belong to the same class would cluster in the space in the sense that distance between members of the same class would be small, and that members of another class would also cluster, but that the two clusters representing the two classes would remain separated from one another. *Distance metric learning* [67] essentially is the term given to the learning of a distance that brings similar points closer together while staying far from the dissimilar points.

To facilitate the process of EEG signal classification, we have to establish a measure which leads to a short distance between similar power spectral densities (i.e., EEG signals of same state of sleep) and a large distance between dissimilar power spectral densities (i.e., EEG signals of different states of sleep). For data in the similar class, since the distance metric characterizes how the like data are clustered, the mean-square distance between members of the class is a measure of the size of the

cluster so formed. For classification, such a mean-square distance should be as small as possible. There are various methods of choosing a metric, most involve a transformation of the data, that minimizes the size of the cluster [75]. Figure 4.2 illustrates how a transformation applied to the data may change the distribution. On the other hand, distance as dissimilarity measure is also beneficial for data classification. However, in some cases because of the variations of the data may be large, the inclusion of the direct dissimilarity distance may not be too helpful in the classification. Again, suitable data transformation will enable the process of classification to be enhance by using the dissimilarity measure.

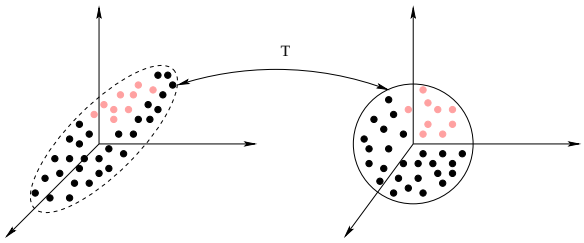


Figure 4.2: Illustration of class separation by transformation

Linear data transformation can be achieved by incorporating a positive definite weighting matrix in the distance measures. An optimally weighted distance could be obtained by optimizing a criterion of which the weighted distance is a factor. In the following, we examine the different criteria which concurrently apply both the similarity and dissimilarity measures. We also examine how each of these criteria can be optimized so that both the similarity and dissimilarity measures are jointly employed for optimal classification of EEG signal. Since we have been treating the EEG signals as both vectors in a linear space or as points on a Riemannian manifold, we will examine optimal distance weighting in both cases.

4.2 Optimally weighted Euclidean distance for similarity/dissimilary

Optimally weighted distances in a vector space [36] [66] have been studied for many years. An earlier treatment of finding an optimally weighted distance can be found in [75]. In this section, however, we would like to focus our attention on a more recent development of optimally weighted distance directly for similarity/dissimilarity proposed by Xing, Ng, and Jordan [90]. We first review the idea in this section:
Given a set of points $\{\mathbf{x}_i\}_{i=1}^{I_0} \subseteq \mathbb{R}^M$, one may form a set of *pairs of similar points* $\mathcal{S} = \{(\mathbf{x}_i, \mathbf{x}_j) : \mathbf{x}_i \sim \mathbf{x}_j\}$, and a set of *pairs of dissimilar points* $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{x}_j) : \mathbf{x}_i \nsim \mathbf{x}_j\}$. The distance metric learning is then to learn a weighted Euclidean distance (weighted L^2 distance) of the form

$$d_{\mathbf{W}}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_{\mathbf{W}} = \sqrt{(\mathbf{x} - \mathbf{y})^T \mathbf{W} (\mathbf{x} - \mathbf{y})} \tag{4.1}$$

where $\mathbf{W}$ is an $M \times M$ positive semi-definite matrix. According to the idea mentioned before, i.e., the weighted distance should minimize the distance between similar points and meanwhile maximize the distance between dissimilar points. For this purpose, one may formulate an optimization problem as:

$$\begin{aligned} \max_{\mathbf{W}} \quad & \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}} \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{W}}^2 \\ \text{s.t} \quad & \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}} \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{W}}^2 \leq 1 \\ & \mathbf{W} \succeq 0 \end{aligned} \tag{4.2}$$

Although this optimization problem has a closed form solution, the solution always gives rank-one weighting matrix $\mathbf{W}$. This can be shown as follows:

First, we formulate an optimization problem equivalent to Problem (4.2) so that

$$\begin{aligned} \max_{\mathbf{W}} \quad & \frac{\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}} \|\mathbf{x}_i-\mathbf{x}_j\|^2_{\mathbf{W}}}{\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} \|\mathbf{x}_i-\mathbf{x}_j\|^2_{\mathbf{W}}} \\ \text{s.t} \quad & \mathbf{W}\succeq 0 \end{aligned} \tag{4.3}$$

Let us sum the correlation matrices of the difference vectors within the similar and dissimilar sets forming two $M \times M$ matrices such that

$$\mathbf{M}_{\mathcal{D}} = \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}} (\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^T \tag{4.4}$$

and

$$\mathbf{M}_{\mathcal{S}} = \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} (\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^T \tag{4.5}$$

Then we can rewrite the Problem (4.3) as

$$\begin{aligned} \max_{\mathbf{W}} \quad & \frac{\mathrm{Tr}(\mathbf{W}\mathbf{M}_{\mathcal{D}})}{\mathrm{Tr}(\mathbf{W}\mathbf{M}_{\mathcal{S}})} \\ \text{s.t} \quad & \mathbf{W}\succeq 0 \end{aligned}$$

which is equivalent to

$$\begin{aligned} \max_{\mathbf{W}} \quad & \mathrm{Tr}(\mathbf{W}\mathbf{M}_{\mathcal{D}}) \\ \text{s.t} \quad & \mathrm{Tr}(\mathbf{W}\mathbf{M}_{\mathrm{S}}) = 1 \\ & \mathbf{W}\succeq 0 \end{aligned} \tag{4.6}$$

Since $\mathbf{W}\succeq 0$, we can decompose $\mathbf{W}$ as

$$\mathbf{W} = \mathbf{\Omega}\mathbf{\Omega}^T \tag{4.7}$$

where $\mathbf{\Omega}$ is an $M \times M$ square matrix. Thus Problem (4.6) becomes

$$\begin{aligned} \max_{\mathbf{\Omega}} \quad & \mathrm{Tr}(\mathbf{\Omega}^{\mathrm{T}}\mathbf{M}_{\mathcal{D}}\mathbf{\Omega}) \\ \text{s.t} \quad & \mathrm{Tr}(\mathbf{\Omega}^{\mathrm{T}}\mathbf{M}_{\mathcal{S}}\mathbf{\Omega}) = 1 \end{aligned} \tag{4.8}$$

To solve Problem (4.8), we use the Lagrange multiplier method and form the auxiliary function such that

$$\phi(\mathbf{\Omega}, \lambda) = \mathrm{Tr}(\mathbf{\Omega}^{\mathrm{T}}\mathbf{M}_{\mathcal{D}}\mathbf{\Omega}) - \lambda[\mathrm{Tr}(\mathbf{\Omega}^{\mathrm{T}}\mathbf{M}_{\mathrm{S}}\mathbf{\Omega}) - 1] \tag{4.9}$$

Taking derivative with respect to $\mathbf{\Omega}$ and setting the result equal to 0, we obtain an eigen-equation

$$\mathbf{M}_{\mathcal{D}}\mathbf{\Omega} = \lambda\mathbf{M}_{\mathcal{S}}\mathbf{\Omega} \tag{4.10}$$

Let $\mathbf{\Omega} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_M]$. Since λ is fixed, we must have $\boldsymbol{v}_1 = \boldsymbol{v}_2 = \cdots = \boldsymbol{v}_M$. Let us denote these $\boldsymbol{v}_i$ as $\boldsymbol{v}$. Then we have

$$\mathbf{W} = \mathbf{\Omega}\mathbf{\Omega}^T = M\boldsymbol{v}\boldsymbol{v}^T, \tag{4.11}$$

which is a rank one matrix and will not serve the purpose of a weighting matrix for enhancing certain aspects of the data.

The X-N-J Optimum Weighting:

To overcome the rank one problem, Xing, et. al. [90] proposed to modify the constraint of the original optimization problem from 2-norm in (4.2) to 1-norm such that

$$\begin{aligned} \min_{\mathbf{W}} \quad & \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{W}}^2 \\ s.t. \quad & \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}} \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{W}} \geq 1 \\ & \mathbf{W} \succeq 0 \end{aligned} \tag{4.12}$$

Or, equivalently,

$$\begin{aligned} \max_{\mathbf{W}} \quad & f(\mathbf{W}) = \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}} \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{W}} \\ s.t \quad & g(\mathbf{W}) = \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{W}}^2 \leq 1 \\ & \mathbf{W} \succeq 0. \end{aligned} \tag{4.13}$$

This is a convex problem [29] which can be solved numerically: Let $\nabla_{\mathbf{W}} g(\mathbf{W})$ be the gradient of $g(\mathbf{W})$ and $(\nabla_{\mathbf{W}} f(\mathbf{W}))_{\perp} \nabla_{\mathbf{W}} g(\mathbf{W})$ be the projection of $\nabla_{\mathbf{W}} f(\mathbf{W})$ onto the orthogonal subspace of $\nabla_{\mathbf{W}} g(\mathbf{W})$. Let $C_1 = \{\mathbf{W} : \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}} \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{W}}^2 \leq 1\}$ and $C_2 = \{\mathbf{W} : \mathbf{W} \succeq 0\}$. Then, denoting by $\|\cdot\|_F$ the Frobenius norm, the gradient ascent + iterative projection algorithm for solving the above optimization problem is shown in Table 4.1.

Table 4.1: The gradient ascent + iterative projection algorithm

(1) **Initialize** $\mathbf{W}^{(0)} := \mathbf{I}$, $n := 0$

(2) **Iterate**

  (a) **Project** $\mathbf{W}^{(n)}$ **onto** C_1 **and** C_2**:**

    (I) **Initialize** $\mathbf{W}_P^{(0)} := \mathbf{W}^{(n)}$, $m := 0$

    (J) **Iterate**

      (i) $\mathbf{W}_P^{(m+1)} := \arg\min_{\mathbf{W}_1} \{\|\mathbf{W}_1 - \mathbf{W}_P^{(m)}\|_F : \mathbf{W}_1 \in C_1\}$

      (j) $\mathbf{W}_P^{(m+1)} := \arg\min_{\mathbf{W}_2} \{\|\mathbf{W}_2 - \mathbf{W}_P^{(m+1)}\|_F : \mathbf{W}_2 \in C_2\}$

      (k) $m := m + 1$

    (K) **Until** $\mathbf{W}_P$ converges

  (b) $\mathbf{W}^{(n)} := \mathbf{W}_P$

  (c) $\mathbf{W}^{(n+1)} := \mathbf{W}^{(n)} + \alpha(\nabla_{\mathbf{W}^{(n)}} f(\mathbf{W}^{(n)}))_{\perp} \nabla_{\mathbf{W}^{(n)}} g(\mathbf{W}^{(n)})$

  (d) $n : +1$

(3) **Until W converges**

The optimum weighting matrix so obtained is designated the X-N-J optimum weighting (X, N, and J being the first letters of the last names of the authors). The X-N-J algorithm leads us to a numerical global optimum solution of $\mathbf{W}$. However, it does not find a closed form solution for the Problem (4.12) (or the Problem (4.13)). In the following section, we are going to generalize the Problem (4.2) so that some closed forms of weighted distances can be achieved. The generalization involves putting

$\mathbf{W} = \mathbf{\Omega}\mathbf{\Omega}^T$ and allowing the factor $\mathbf{\Omega}$ in the weighting matrix to be a full column-rank "tall" matrix.

4.3 Generalization of optimally weighted Euclidean distance

Let $\mathcal{M}_K$ be the set of all $K \times K$ matrices over the field $\mathbb{C}$. Recall that [60], for $\mathbf{B} \in \mathcal{M}_K$ and $\kappa = 0, \cdots, K$, the function $E_{K-\kappa}(\mathbf{B})$ (sometimes called the $(K-\kappa)$th trace of $\mathbf{B}$) is defined as the sum of the $(K-\kappa)$th order principal minors of $\mathbf{B}$, i.e.,

$$E_{K-\kappa}(\mathbf{B}) = \sum_{\wp} \det[\mathbf{B}_{i_1,\cdots,i_\kappa}] \tag{4.14}$$

where $\det[\cdot]$ denotes determinant and $\mathbf{B}_{i_1,\cdots,i_\kappa}$ is the principal submatrix of $\mathbf{B}$ formed by deleting the i_1th, i_2th,$\cdots$, i_κth rows and columns of $\mathbf{B}$, and $\wp$ denotes the combination set of $\{i_1, \cdots, i_\kappa\}$. We note that $E_0(\mathbf{B}) = 0, E_1(\mathbf{B}) = Tr(\mathbf{B})$, and $E_K(\mathbf{B}) = \det[\mathbf{B}]$. As an example, let

$$\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} \tag{4.15}$$

Then, we have

$$\begin{aligned} E_1(\mathbf{B}) &= \sum_{1\le i_1,i_2\le 3} \det[\mathbf{B}_{i_1,i_2}] = \det[b_{11}] + \det[b_{22}] + \det[b_{33}] \\ &= b_{11} + b_{22} + b_{33} = \mathrm{Tr}(\mathbf{B}) \qquad (4.16) \\ E_2(\mathbf{B}) &= \sum_{1\le i_1\le 3} \det[\mathbf{B}_{i_1}] \\ &= \det\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} + \det\begin{bmatrix} b_{11} & b_{13} \\ b_{31} & b_{33} \end{bmatrix} + \det\begin{bmatrix} b_{22} & b_{23} \\ b_{32} & b_{33} \end{bmatrix} \\ &= (b_{11}b_{22} - b_{12}b_{21}) + (b_{11}b_{33} - b_{13}b_{31}) + (b_{22}b_{33} - b_{23}b_{32}) \qquad (4.17) \\ \text{and } E_3(\mathbf{B}) &= \det[\mathbf{B}] \qquad (4.18) \end{aligned}$$

We can also write $E_{K-\kappa}(\mathbf{B})$ in terms of the eigenvalues $\lambda_1, \cdots, \lambda_K$ of $\mathbf{B}$, i.e.,

$$E_{K-\kappa}(\mathbf{B}) = \sum_{K-\kappa} \left(\prod \lambda_i \right) \qquad (4.19)$$

where $\sum_{K-\kappa} (\prod \lambda_i)$ denotes the sum of the products of the eigenvalues of $\mathbf{B}$ taken $(K-\kappa)$ at a time.

We now apply the function[1] $E_{K-\kappa}(\mathbf{B})$ as defined above to our problem of finding the optimum weighting for similarity/dissimilarity distance. Again, let

$$\mathbf{M}_{\mathcal{D}} = \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^H \qquad (4.20)$$

and

$$\mathbf{M}_{\mathcal{S}} = \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^H \qquad (4.21)$$

Let $\mathbf{\Omega} \in \mathcal{M}_{M\times K}$ where $\mathcal{M}_{M\times K}$ is the set of all $M \times K$ matrices with $K \le M$. Since the purpose of a weighting matrix is to enhance certain parts of the data and de-emphasize other parts, often, we require $\mathbf{\Omega}$) to be orthonormal so that the total

[1]The function has been applied in array signal processing for locating the direction of arrival of target signals [88]

energy after transformation is unchanged. Thus, if $E_{K-\kappa}(\mathbf{\Omega}^H \mathbf{M}_{\mathcal{S}} \mathbf{\Omega}) \neq 0$, then we can define the following generalized optimization problem:

$$\begin{aligned} \max_{\mathbf{\Omega}} \quad & \frac{E_{K-\kappa}(\mathbf{\Omega}^H \mathbf{M}_{\mathcal{D}} \mathbf{\Omega})}{E_{K-\kappa}(\mathbf{\Omega}^H \mathbf{M}_{\mathcal{S}} \mathbf{\Omega})} \\ s.t. \quad & \mathbf{\Omega}^H \mathbf{\Omega} = \mathbf{I}_K \end{aligned} \tag{4.22}$$

with $\mathbf{I}_K$ being a $K \times K$ identity matrix.

Let us examine the geometric meaning [88] of the objective function in Eq. (4.22). We see that both the numerator and the denominator are of the form $E_{K-\kappa}(\mathbf{\Omega}^H \mathbf{M}_0 \mathbf{\Omega})$ where $\mathbf{M}_0$ is an $M \times M$ Hermitian matrix representing either $\mathbf{M}_{\mathcal{D}}$ or $\mathbf{M}_{\mathcal{S}}$. Let

$$\mathbf{\Omega}^H \mathbf{M}_0 \mathbf{\Omega} = [\boldsymbol{\eta}_1 \ \boldsymbol{\eta}_2 \ \ \cdots \ \ \boldsymbol{\eta}_K]^H [\boldsymbol{\eta}_1 \ \boldsymbol{\eta}_2 \ \ \cdots \ \ \boldsymbol{\eta}_K] \tag{4.23}$$

where $\boldsymbol{\eta}_i$ is a K-dimensional vector. Then, it can be easily seen [16] that each of its $(K-\kappa)$-dimensional principal minors (formed by deleting κ of the corresponding rows and columns) is equal to the square of the volume of the $(K-\kappa)$-dimensional parallelepiped whose edges are the $K-\kappa$ vectors $\{\boldsymbol{\eta}_i\}$ involved in the principal minor. Therefore, the maximization of the term $E_{K-\kappa}(\mathbf{\Omega}^H \mathbf{M}_0 \mathbf{\Omega})$ can be interpreted as the maximization of the sum of the square of the volumes of all the parallelepipeds whose edges are formed by taking all the possible combinations of $K-\kappa$ of the vectors $\{\boldsymbol{\eta}_i\}, i = 1, \cdots, K$. Since the volume of a parallelepiped not only depends on the length of the vectors forming its edges, but also on the angles between them, we can see that the maximization of $E_{K-\kappa}(\mathbf{\Omega}^H \mathbf{M}_0 \mathbf{\Omega})$ is to find a weighting matrix which also maximizes the angles between the vectors, that is minimizes the correlations.

From the above geometric interpretation, we can see that maximization of the objective function $F_{\text{obj}} = \max\left(\frac{E_{K-\kappa}(\mathbf{\Omega}^H \mathbf{M}_{\mathcal{D}} \mathbf{\Omega})}{E_{K-\kappa}(\mathbf{\Omega}^H \mathbf{M}_{\mathcal{S}} \mathbf{\Omega})}\right)$ amounts to finding a weighting matrix that can minimize the correlation between the dissimilar vectors while concurrently can maximize the correlation between the similar vectors. This "ideal" weighting

matrix may be difficult to find. Hence, very often, instead of optimizing "the quotient of the functions", we may choose to approximate the objective function by "the function of the quotient" such that $F_{\text{obj}} \approx \max E_{K-\kappa}\left([\mathbf{\Omega}^H\mathbf{M}_{\mathcal{S}}\mathbf{\Omega}]^{-1}[\mathbf{\Omega}^H\mathbf{M}_{\mathcal{D}}\mathbf{\Omega}]\right)$ where of course, the "quotient" here is the inverse of the denominator matrix multiplied by the numerator matrix. While other values of $K-\kappa$ may also yield very interesting results, in the following, we only limit our examination of the optimization problems to the two special cases of $K-\kappa=1$ and $K-\kappa=K$ in which the function $E_{K-\kappa}(\mathbf{B})$ gives us respectively a trace quotient problem and a Rayleigh quotient problem.

4.3.1 Case I: $K-\kappa=1$

As seen in the discussion of the function $E_{K-\kappa}(\mathbf{B})$, for $K-\kappa=1,\ E_1(\cdot)$ is the trace of the matrix, in which case, our problem becomes:

$$\begin{aligned} \max_{\mathbf{\Omega}} \quad & \frac{\text{Tr}(\mathbf{\Omega}^{\text{H}}\mathbf{M}_{\mathcal{D}}\mathbf{\Omega})}{\text{Tr}(\mathbf{\Omega}^{\text{H}}\mathbf{M}_{\mathcal{S}}\mathbf{\Omega})} \\ s.t. \quad & \mathbf{\Omega}^H\mathbf{\Omega}=\mathbf{I}_K \end{aligned} \tag{4.24}$$

which is also difficult to solve. However, as discussed above, we can form an approximation to this problem such that

$$\begin{aligned} \max_{\mathbf{\Omega}} \quad & \text{Tr}[(\mathbf{\Omega}^{\text{H}}\mathbf{M}_{\mathcal{S}}\mathbf{\Omega})^{-1}\mathbf{\Omega}^{\text{H}}\mathbf{M}_{\mathcal{D}}\mathbf{\Omega}] \\ s.t. \quad & \mathbf{\Omega}^H\mathbf{\Omega}=\mathbf{I}_K \end{aligned} \tag{4.25}$$

To solve this Problem 4.25, we need the following lemma [63]:

Lemma 4.1 *Let* $\mathbf{P}$ *be an* $M\times M$ *positive definite matrix with eigenvalues* $\lambda_1\geq\cdots\geq\lambda_M$ *and associated orthonormal eigenvectors* $\mathbf{v}_1,\cdots,\mathbf{v}_M$*. Then the problem*

$$\begin{aligned} \max_{\mathbf{\Omega}} \quad & \text{Tr}(\mathbf{Q}^{\text{H}}\mathbf{P}\mathbf{Q}) \\ s.t. \quad & \mathbf{Q}^H\mathbf{Q}=\mathbf{I}_K \end{aligned} \tag{4.26}$$

has the solution $\sum_{i=1}^{K} \lambda_i$ *with the optimizing matrix*

$$\mathbf{Q}_{opt} = [\mathbf{v}_1, \cdots, \mathbf{v}_K]^T \tag{4.27}$$

□

Now, let $\mathbf{A} \in \mathcal{M}_{M\times K}$ and let

$$\mathbf{M}'_{\mathcal{D}} = \mathbf{A}^H \mathbf{M}_{\mathcal{D}} \mathbf{A} \tag{4.28}$$

and

$$\mathbf{M}'_{\mathcal{S}} = \mathbf{A}^H \mathbf{M}_{\mathcal{S}} \mathbf{A} \tag{4.29}$$

Let $\mathbf{\Theta}$ and $\mathbf{\Phi}$ be the eigenvalue and eigenvector matrices of $\mathbf{M}'_{\mathcal{S}}$. Then we have

$$\mathbf{M}'_{\mathcal{S}} \mathbf{\Phi} = \mathbf{\Phi} \mathbf{\Theta} \tag{4.30}$$

Let $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ be the eigenvalue and eigenvector matrices of $\mathbf{\Theta}^{-1/2} \mathbf{\Phi}^H \mathbf{M}'_{\mathcal{D}} \mathbf{\Phi} \mathbf{\Theta}^{-1/2}$. Let $\mathbf{B} = \mathbf{\Phi} \mathbf{\Theta}^{-1/2} \mathbf{\Psi}$ (Note that $\mathbf{B}$ is a nonsingular $K \times K$ matrix). Then, it is easy to verify that

$$\mathbf{B}^H \mathbf{M}'_{\mathcal{D}} \mathbf{B} = \mathbf{\Lambda} \tag{4.31}$$

and

$$\mathbf{B}^H \mathbf{M}'_{\mathcal{S}} \mathbf{B} = \mathbf{I}_K \tag{4.32}$$

Then, we have

$$\max_{\mathbf{A}} \mathrm{Tr}(\mathbf{A}^{\mathrm{H}} \mathbf{M}_{\mathcal{S}} \mathbf{A})^{-1} (\mathbf{A}^{\mathrm{H}} \mathbf{M}_{\mathcal{D}} \mathbf{A}) = \max_{\mathbf{A}} \mathrm{Tr}(\mathbf{B}^{\mathrm{H}} \mathbf{A}^{\mathrm{H}} \mathbf{M}_{\mathcal{S}} \mathbf{A} \mathbf{B})^{-1} (\mathbf{B}^{\mathrm{H}} \mathbf{A}^{\mathrm{H}} \mathbf{M}_{\mathcal{D}} \mathbf{A} \mathbf{B}) \tag{4.33}$$

Let $\mathbf{\Omega} = \mathbf{AB}$. Then, the Problems (4.25) has been transformed to the following problem

$$\begin{aligned} \max_{\mathbf{\Omega}} \quad & \mathrm{Tr} \mathbf{\Omega}^{\mathrm{H}} \mathbf{M}_{\mathcal{D}} \mathbf{\Omega} \\ s.t. \quad & \mathbf{\Omega}^H \mathbf{M}_{\mathcal{S}} \mathbf{\Omega} = \mathbf{I}_K \end{aligned} \tag{4.34}$$

Let $\mathbf{M}_{\mathcal{S}} = \mathbf{H}\mathbf{H}^H$ and $\mathbf{Y} = \mathbf{H}^H\mathbf{\Omega}$. Then,

$$\mathbf{\Omega} = \mathbf{H}^{-H}\mathbf{Y}, \tag{4.35}$$

and

$$\mathbf{\Omega}^H\mathbf{M}_{\mathcal{D}}\mathbf{\Omega} = \mathbf{Y}^H\mathbf{H}^{-1}\mathbf{M}_{\mathcal{D}}\mathbf{H}^{-H}\mathbf{Y} = \mathbf{Y}^H\tilde{\mathbf{M}}_{\mathcal{D}}\mathbf{Y}, \tag{4.36}$$

where $\tilde{\mathbf{M}}_{\mathcal{D}} = \mathbf{H}^{-1}\mathbf{M}_{\mathcal{D}}\mathbf{H}^{-H}$. Thus, Problem (4.34) becomes

$$\begin{aligned} \max_{\mathbf{Y}} \quad & \mathrm{Tr}(\mathbf{Y}^{\mathrm{T}}\tilde{\mathbf{M}}_{\mathcal{D}}\mathbf{Y}) \\ s.t. \quad & \mathbf{Y}^H\mathbf{Y} = \mathbf{I}_K \end{aligned} \tag{4.37}$$

By Lemma 4.1, if $\lambda_1 \geq \cdots \geq \lambda_K$ are the eigenvalues of $\tilde{\mathbf{M}}_{\mathcal{D}}$ associated with orthonormal eigenvectors $\mathbf{v}_1, \cdots, \mathbf{v}_K$, then the maximizing matrix is $\mathbf{Y}_{opt} = [\mathbf{v}_1, \cdots, \mathbf{v}_K]$. Note that if $\mathbf{\Lambda} = \mathrm{diag}[\lambda_1, \cdots, \lambda_K]$, then

$$\mathbf{H}^{-1}\mathbf{M}_{\mathcal{D}}\mathbf{H}^{-H}\mathbf{Y}_{opt} = \mathbf{Y}_{opt}\mathbf{\Lambda} \tag{4.38}$$

Since $\mathbf{Y}_{opt} = \mathbf{H}^H\mathbf{\Omega}_{opt}$, we have

$$\mathbf{H}^{-1}\mathbf{M}_{\mathcal{D}}\mathbf{H}^{-H}\mathbf{H}^H\mathbf{\Omega}_{opt} = \mathbf{H}^H\mathbf{\Omega}_{opt}\mathbf{\Lambda} \tag{4.39}$$

and

$$\mathbf{H}^{-H}\mathbf{H}^{-1}\mathbf{M}_{\mathcal{D}}\mathbf{\Omega}_{opt} = \mathbf{\Omega}_{opt}\mathbf{\Lambda} \tag{4.40}$$

Finally, since $\mathbf{M}_{\mathcal{S}} = \mathbf{H}\mathbf{H}^H$, we have

$$\mathbf{M}_{\mathcal{S}}^{-1}\mathbf{M}_{\mathcal{D}}\mathbf{\Omega}_{opt} = \mathbf{\Omega}_{opt}\mathbf{\Lambda}. \tag{4.41}$$

Therefore, $\mathbf{\Omega}_{opt}$ is composed of the first K eigenvectors corresponding to the first K largest eigenvalues of $\mathbf{M}_{\mathcal{S}}^{-1}\mathbf{M}_{\mathcal{D}}$, i.e.,

$$\mathbf{\Omega}_{opt} = [\mathbf{u}_1, \cdots, \mathbf{u}_K]^T \tag{4.42}$$

where $\mathbf{u}_1, \cdots, \mathbf{u}_K$ are the orthonormal eigenvectors corresponding to the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_K$ of $\mathbf{M}_{\mathcal{S}}^{-1}\mathbf{M}_{\mathcal{D}}$. It can be observed from Eqs. (4.38) and (4.41) that while $\tilde{\mathbf{M}}_{\mathcal{D}}$ and $\mathbf{M}_{\mathcal{S}}^{-1}\mathbf{M}_{\mathcal{D}}$ both have the same eigenvalues $\{\lambda_i\}$, they have different eigenvectors, being respectively given by $\{\mathbf{v}_i\}$ and $\{\mathbf{u}_i\}, i = 1, \cdots, M$.

We also note that for $K = M$, $\mathbf{\Omega}_{opt}$ will incorporate all the eigenvectors of $\mathbf{M}_{\mathcal{S}}^{-1}\mathbf{M}_{\mathcal{D}}$ and $\mathbf{W}_{opt} = \mathbf{\Omega}_{opt}\mathbf{\Omega}_{opt}^H = \mathbf{I}_M$ which will not do any weighting to the objective function. This, however, is not equivalent to the original problem of Xing *et al* in Eq. (4.8) since the constraints in the two cases are different.

4.3.2 Case II: $K - \kappa = K$

From the discussion of the function $E_{K-\kappa}(\mathbf{B})$, for $K-\kappa = K$, $E_K(\cdot)$ is the determinant of the matrix. In this case, the problem is reduced to a Rayleigh quotient problem:

$$\begin{aligned} \max_{\mathbf{\Omega}} \quad & \frac{\det[\mathbf{\Omega}^H\mathbf{M}_{\mathcal{D}}\mathbf{\Omega}]}{\det[\mathbf{\Omega}^H\mathbf{M}_{\mathcal{S}}\mathbf{\Omega}]} \\ s.t. \quad & \mathbf{\Omega} \in \mathcal{M}_{M\times K} \end{aligned} \tag{4.43}$$

The solution of Problem (4.43) necessitates the following lemma [63]:

Lemma 4.2 *Let* $\mathbf{P}$ *be an* $M \times M$ *positive definite matrix with eigenvalues* $\lambda_1 \geq \cdots \geq \lambda_M$ *and associated orthonormal eigenvectors* $\mathbf{v}_1, \cdots, \mathbf{v}_M$*. Then the problem*

$$\begin{aligned} \max_{\mathbf{Q}} \quad & \det[\mathbf{Q}^H\mathbf{P}\mathbf{Q}] \\ s.t. \quad & \mathbf{Q}^H\mathbf{Q} = \mathbf{I}_K \end{aligned} \tag{4.44}$$

has the solution

$$\mathbf{Q}_{opt} = [\mathbf{v}_1, \cdots, \mathbf{v}_K]^T \tag{4.45}$$

□

Let $\mathbf{M}_{\mathcal{S}} = \mathbf{L}\mathbf{L}^H$ be the Cholesky decomposition of $\mathbf{M}_{\mathcal{S}}$. Let $\mathbf{Z}$ be an $M \times K$ matrix such that $\mathbf{Z} = \boldsymbol{\Upsilon}\boldsymbol{\Sigma}\mathbf{V}^H$ where $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \ \cdots \ \sigma_K)$ is a $K \times K$ diagonal matrix consisting of the K singular values $\sigma_1 \geq \ \cdots \ \geq \sigma_K$, $\boldsymbol{\Upsilon} = [\boldsymbol{\upsilon}_1 \ \cdots \ \boldsymbol{\upsilon}_K]$ is an $M \times K$ matrix consisting of the first K left singular vectors corresponding to the singular values of $\mathbf{Z}$ such that $\boldsymbol{\Upsilon}^H\boldsymbol{\Upsilon} = \mathbf{I}_K$, and $\mathbf{V} = [\mathbf{v}_1 \ \cdots \ \mathbf{v}_K]$ is a $K \times K$ matrix consisting of the K right singular vectors of $\mathbf{Z}$ such that $\mathbf{V}^H\mathbf{V} = \mathbf{I}_K$. Let $\boldsymbol{\Omega} = \mathbf{L}^{-H}\mathbf{Z}$. Then, we have

$$\begin{aligned}
\frac{\det[\boldsymbol{\Omega}^H\mathbf{M}_{\mathcal{D}}\boldsymbol{\Omega}]}{\det[\boldsymbol{\Omega}^H\mathbf{M}_{\mathcal{S}}\boldsymbol{\Omega}]} &= \frac{\det[\mathbf{Z}^H\mathbf{L}^{-1}\mathbf{M}_{\mathcal{D}}\mathbf{L}^{-H}\mathbf{Z}]}{\det[\mathbf{Z}^H\mathbf{Z}]} \\
&= \frac{\det[\mathbf{V}\boldsymbol{\Sigma}\boldsymbol{\Upsilon}^H\mathbf{L}^{-1}\mathbf{M}_{\mathcal{D}}\mathbf{L}^{-H}\boldsymbol{\Upsilon}\boldsymbol{\Sigma}\mathbf{V}^H]}{\det[\mathbf{V}\boldsymbol{\Sigma}\boldsymbol{\Sigma}\mathbf{V}^H]} \\
&= \det[\boldsymbol{\Upsilon}^H\mathbf{L}^{-1}\mathbf{M}_{\mathcal{D}}\mathbf{L}^{-H}\boldsymbol{\Upsilon}] \qquad (4.46)
\end{aligned}$$

where we have used the fact that $\det(\mathbf{AB}) = \det(\mathbf{A}) \cdot \det(\mathbf{B})$. Therefore, the Problem (4.43) can be transformed to the following equivalent problem:

$$\begin{aligned}
\max_{\boldsymbol{\Upsilon}} \quad & \det[\boldsymbol{\Upsilon}^H\mathbf{L}^{-1}\mathbf{M}_{\mathcal{D}}\mathbf{L}^{-H}\boldsymbol{\Upsilon}] \\
s.t. \quad & \boldsymbol{\Upsilon}^H\boldsymbol{\Upsilon} = \mathbf{I}_K \qquad (4.47)
\end{aligned}$$

on which Lemma 4.2 can be directly applied. Let $\lambda_1 \geq \cdots \geq \lambda_M$ be the ordered eigenvalues of $\mathbf{L}^{-1}\mathbf{M}_{\mathcal{D}}\mathbf{L}^{-H}$ with the associated eigenvectors $\mathbf{u}_1, \cdots, \mathbf{u}_M$. Then, the maximizing matrix to the Problem (4.47) is

$$\boldsymbol{\Upsilon}_{\text{opt}} = [\mathbf{u}_1, \cdots, \mathbf{u}_K]^T \qquad (4.48)$$

Therefore, the maximizing matrix to the problem (4.43) is

$$\boldsymbol{\Omega}_{\text{opt}} = \mathbf{L}^{-H}\boldsymbol{\Upsilon}_{\text{opt}}\boldsymbol{\Sigma}\mathbf{V}^H \qquad (4.49)$$

4.4 Optimum weighting for Riemannian distances

In Chapter 3, we have seen that the features of EEG signals describe a manifold on which the distance between two points can best be measured by a Riemannian distance. In Section 4.1, we have further seen that EEG signal classification consists of characterizing each of the classes of signals and determining the class to which a new signal belongs, and the aim of metric learning is to find the optimum weighting matrix which minimizes the size of the cluster of similar signals while keeping the dissimilar signals at a prescribed distance. In this section, we apply some of the results of optimization introduced in the previous section to find the optimum weighting matrix $\mathbf{W}$ for the Riemannian distance developed between two weighted PSD matrices.

4.4.1 Optimum weighting for d_{R_1W}

Let $\mathbf{P}_i(\omega)$ and $\mathbf{P}_j(\omega), \omega \in [\omega_{\min}, \omega_{\max}]$, be two separate sample curves of PSD matrices as the frequency ω varies. We say that $\mathbf{P}_i(\omega)$ and $\mathbf{P}_j(\omega)$ are similar if they belong to the same class, and are dissimilar if they belong to different classes. Let $\mathbf{P}_{ik} = \mathbf{P}_i(\omega_k)$ and $\mathbf{P}_{jk} = \mathbf{P}_j(\omega_k)$ represent two separate PSD matrices from the two sample curves measured at $\omega = \omega_k$. Again, we denote the sets of similar and dissimilar PSD matrices by $\mathcal{S}$ and $\mathcal{D}$ respectively such that the set of pairs of similar PSD matrices is $\mathcal{S} = \{(\mathbf{P}_{ik}, \mathbf{P}_{jk}); \mathbf{P}_i(\omega), \mathbf{P}_j(\omega) \in \mathcal{C}_\ell\}$, whereas the set of pairs of dissimilar PSD matrices is $\mathcal{D} = \{(\mathbf{P}_{ik}, \mathbf{P}_{jk}); \mathbf{P}_i(\omega) \in \mathcal{C}_{\ell_i}, \mathbf{P}_j(\omega) \in \mathcal{C}_{\ell_j}, \ell_i \neq \ell_j\}$. The optimum $M \times M$ weighting matrix $\mathbf{W}$ may be found by maximizing the ratio of the sum of squared *interclass* distances and the sum of squared *intraclass* distances, i.e.,

$$\max_{\mathbf{W}} \quad \frac{\sum_{(\mathbf{P}_{ik}, \mathbf{P}_{jk}) \in \mathcal{D}} d^2_{R_1W}(\mathbf{P}_{ik}, \mathbf{P}_{jk})}{\sum_{(\mathbf{P}_{ik}, \mathbf{P}_{jk}) \in \mathcal{S}} d^2_{R_1W}(\mathbf{P}_{ik}, \mathbf{P}_{jk})} \tag{4.50a}$$

$$s.t. \quad \mathbf{W} = \mathbf{W}^H \succ \mathbf{0} \tag{4.50b}$$

where, from Eq. (3.51),

$$d^2_{R_1W}(\mathbf{P}_1,\mathbf{P}_2) = \mathrm{Tr}(\mathbf{W}\mathbf{P}_1) + \mathrm{Tr}(\mathbf{W}\mathbf{P}_2) - 2\mathrm{Tr}\left(\sqrt{\mathbf{P}_2^{1/2}\mathbf{W}\mathbf{P}_1\mathbf{W}\mathbf{P}_2^{1/2}}\right) \tag{4.51}$$

Direct optimization of the quantity in Eq. (4.50a) on the manifold $\mathcal{M}$ is difficult. However, from Chapter 3, we can perform the optimization equivalently using the inner product metric in the Hilbert space $\mathcal{H}_M$. To do this, we follow the steps in the development of Theorem 3.1 by letting $\mathbf{P}_{jk}^{1/2}\mathbf{P}_{ik}^{1/2} = \mathbf{V}_{ij1}\mathbf{\Sigma}_{ij}\mathbf{V}_{ij2}^H$ be the singular-value decomposition of $\mathbf{P}_{jk}^{1/2}\mathbf{P}_{ik}^{1/2}$ where $\mathbf{V}_{ij1}$ and $\mathbf{V}_{ij2}$ are respectively the left and right singular vectors, and let $\mathbf{U}_{ik}$ and $\mathbf{U}_{jk}$ be two unitary matrices such that $\mathbf{U}_{jk}\mathbf{U}_{ik}^H = \mathbf{V}_{ij2}\mathbf{V}_{ij1}^H$. Writing $\tilde{\mathbf{P}}_{ik} = \mathbf{P}_{ik}^{1/2}\mathbf{U}_{ik}$ and $\tilde{\mathbf{P}}_{jk} = \mathbf{P}_{jk}^{1/2}\mathbf{U}_{jk}$, let us now examine how $\tilde{\mathbf{P}}_{ik}$ and $\tilde{\mathbf{P}}_{jk}$ can be optimally weighted:

Following the procedure of Corollary 3.1, we let $\mathbf{W}$ be a positive definite weighting matrix so that $\mathbf{W} = \mathbf{\Omega}\mathbf{\Omega}^H$ and let $\tilde{\mathbf{P}}_{ikW} = \mathbf{\Omega}^H\tilde{\mathbf{P}}_{ik}$ and $\tilde{\mathbf{P}}_{jkW} = \mathbf{\Omega}^H\tilde{\mathbf{P}}_{jk}$, then $\mathbf{P}_{ikW} = \tilde{\mathbf{P}}_{ikW}\tilde{\mathbf{P}}_{ikW}^H$ and $\mathbf{P}_{jkW} = \tilde{\mathbf{P}}_{jkW}\tilde{\mathbf{P}}_{jkW}^H$. Since $\mathcal{T}_{\mathcal{M}}(\mathbf{P})$ and $\mathcal{U}_{\mathcal{H}}(\tilde{\mathbf{P}})$ are isometric, we have

$$\begin{aligned} d^2_{R_1W}(\mathbf{P}_{ik},\mathbf{P}_{jk}) &= \min\|\tilde{\mathbf{P}}_{ikW} - \tilde{\mathbf{P}}_{jkW}\|^2 \\ &= \mathrm{Tr}\left[(\tilde{\mathbf{P}}_{ikW} - \tilde{\mathbf{P}}_{jkW})^H(\tilde{\mathbf{P}}_{ikW} - \tilde{\mathbf{P}}_{jkW})\right] \\ &= \mathrm{Tr}\left[(\tilde{\mathbf{P}}_{ik} - \tilde{\mathbf{P}}_{jk})^H\mathbf{W}(\tilde{\mathbf{P}}_{ik} - \tilde{\mathbf{P}}_{jk})\right] \\ &= \mathrm{Tr}\left[\mathbf{\Omega}^H(\tilde{\mathbf{P}}_{ik} - \tilde{\mathbf{P}}_{jk})(\tilde{\mathbf{P}}_{ik} - \tilde{\mathbf{P}}_{jk})^H\mathbf{\Omega}\right] \end{aligned} \tag{4.52}$$

Let $\tilde{\mathcal{S}}_1 = \{(\tilde{\mathbf{P}}_{ik},\tilde{\mathbf{P}}_{jk}); \mathbf{P}_i(\omega),\mathbf{P}_j(\omega) \in \mathcal{C}_\ell\}$ and $\tilde{\mathcal{D}}_1 = \{(\tilde{\mathbf{P}}_{ik},\tilde{\mathbf{P}}_{jk}); \mathbf{P}_i(\omega) \in \mathcal{C}_{\ell_i}, \mathbf{P}_{jk}(\omega) \in \mathcal{C}_{\ell_j}, \ell_i \neq \ell_j\}$. Then, writing

$$\tilde{\mathbf{M}}_{\tilde{\mathcal{S}}_1} = \sum_{(\tilde{\mathbf{P}}_{ik},\tilde{\mathbf{P}}_{jk})\in\tilde{\mathcal{S}}_1} (\tilde{\mathbf{P}}_{ik} - \tilde{\mathbf{P}}_{jk})(\tilde{\mathbf{P}}_{ik} - \tilde{\mathbf{P}}_{jk})^H \tag{4.53}$$

and

$$\tilde{\mathbf{M}}_{\tilde{\mathcal{D}}_1} = \sum_{(\tilde{\mathbf{P}}_{ik},\tilde{\mathbf{P}}_{jk})\in\tilde{\mathcal{D}}_1} (\tilde{\mathbf{P}}_{ik} - \tilde{\mathbf{P}}_{jk})(\tilde{\mathbf{P}}_{ik} - \tilde{\mathbf{P}}_{jk})^H \tag{4.54}$$

and substituting into Eq. (4.50), the optimization problem becomes

$$\begin{aligned} \max_{\mathbf{\Omega}} \quad & \frac{\mathrm{Tr}\left(\mathbf{\Omega}^H \tilde{\mathbf{M}}_{\tilde{\mathcal{D}}_1} \mathbf{\Omega}\right)}{\mathrm{Tr}\left(\mathbf{\Omega}^H \tilde{\mathbf{M}}_{\tilde{\mathcal{S}}_1} \mathbf{\Omega}\right)} \\ s.t. \quad & \mathbf{\Omega}\mathbf{\Omega}^H \succ 0 \end{aligned} \tag{4.55}$$

As discussed in the last section, this problem may be difficult to solve and we turn to solve an approximation problem such that

$$\begin{aligned} \max_{\mathbf{\Omega}} \quad & \mathrm{Tr}[(\mathbf{\Omega}^H \tilde{\mathbf{M}}_{\tilde{\mathcal{S}}_1} \mathbf{\Omega})^{-1} \mathbf{\Omega}^H \tilde{\mathbf{M}}_{\tilde{\mathcal{D}}_1} \mathbf{\Omega}] \\ s.t. \quad & \mathbf{\Omega} \in \mathcal{M}_{M \times K} \end{aligned} \tag{4.56}$$

which is in the same form as the Problem (4.25). Therefore,

$$\mathbf{\Omega}_{\mathrm{op1}} = [\tilde{\mathbf{v}}_1, \cdots, \tilde{\mathbf{v}}_K]^T \tag{4.57}$$

where $\tilde{\mathbf{v}}_1, \cdots, \tilde{\mathbf{v}}_K$ are the orthonormal eigenvectors corresponding to the eigenvalues $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_K$ of $\tilde{\mathbf{M}}_{\tilde{\mathcal{S}}_1}^{-1} \tilde{\mathbf{M}}_{\tilde{\mathcal{D}}_1}$. Thus, the optimum weighting matrix $\mathbf{W}_{\mathrm{op1}}$ is given by

$$\mathbf{W}_{\mathrm{op1}} = \mathbf{\Omega}_{\mathrm{op1}} \mathbf{\Omega}_{\mathrm{op1}}^H \tag{4.58}$$

4.4.2 Optimum weighting for d_{R_2W}

Let $\tilde{\mathcal{S}}_2 = \{(\mathbf{P}_{ik}^{1/2}, \mathbf{P}_{jk}^{1/2}); \mathbf{P}_i(\omega), \mathbf{P}_j(\omega) \in \mathcal{C}_\ell\}$ and $\tilde{\mathcal{D}}_2 = \{(\mathbf{P}_{ik}^{1/2}, \mathbf{P}_{jk}^{1/2}); \mathbf{P}_i(\omega) \in \mathcal{C}_{\ell_i}, \mathbf{P}_{jk}(\omega) \in \mathcal{C}_{\ell_j}, \ell_i \neq \ell_j\}$. Then, writing

$$\tilde{\mathbf{M}}_{\check{\mathcal{S}}_2} = \sum_{(\mathbf{P}_{ik}^{1/2}, \mathbf{P}_{jk}^{1/2}) \in \tilde{\mathcal{S}}_2} (\mathbf{P}_{ik}^{1/2} - \mathbf{P}_{jk}^{1/2})(\mathbf{P}_{ik}^{1/2} - \mathbf{P}_{jk}^{1/2})^H \tag{4.59}$$

and

$$\tilde{\mathbf{M}}_{\check{\mathcal{D}}_2} = \sum_{(\mathbf{P}_{ik}^{1/2}, \mathbf{P}_{jk}^{1/2}) \in \tilde{\mathcal{D}}_2} (\mathbf{P}_{ik}^{1/2} - \mathbf{P}_{jk}^{1/2})(\mathbf{P}_{ik}^{1/2} - \mathbf{P}_{jk}^{1/2})^H \tag{4.60}$$

and using the same reasoning as the previous section we can find the optimum

$$\mathbf{\Omega}_{\text{op2}} = [\tilde{\mathbf{v}}_1, \cdots, \tilde{\mathbf{v}}_K]^T \tag{4.61}$$

where $\tilde{\mathbf{v}}_1, \cdots, \tilde{\mathbf{v}}_K$ are the orthonormal eigenvectors corresponding to the eigenvalues $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_K$ of $\tilde{\mathbf{M}}_{\mathcal{S}_2}^{-1}\check{\mathbf{M}}_{\mathcal{D}_2}$. Thus, the optimum weighting matrix $\mathbf{W}_{\text{op2}}$ is given by

$$\mathbf{W}_{\text{op2}} = \mathbf{\Omega}_{\text{op2}}\mathbf{\Omega}_{\text{op2}}^H \tag{4.62}$$

In this chapter, we have examined the use of weighting in the signal classification. The principle established here is that an optimum weighting matrix, while keeping the data of dissimilarity to be as distant as possible, should keep the data of similarity to be within the vicinity. Using this principle, we have derived approximate optimum weightings for both d_{R_1W} and d_{R_2W}. (It has been shown that d_{R_3} is weight invariant.) In the ensuing chapter, we will apply both the weighted and unweighted distance measures to the classification of EEG signals. The effects of optimum weighting will be apparent from those results.

Chapter 5

Geometric EEG signal classification

In this Chapter, we apply the optimally weighted Riemannian distance derived in Chapters 3 and 4 to the classification of EEG signals for the determination of a patient's sleep stage. Since our similarity/dissimilarity measure is defined by considering the geometric structure of the feature space, our classification method will be called *Geometric EEG Signal Classification* to emphasize this aspect. Specifically, our method is k-nearest neighbor (k-NN) rule coupled with the similarity/dissimilarity measures based on (both unweighted and weighted) Riemannian distances. In the following, we will describe each part of the classification method in details.

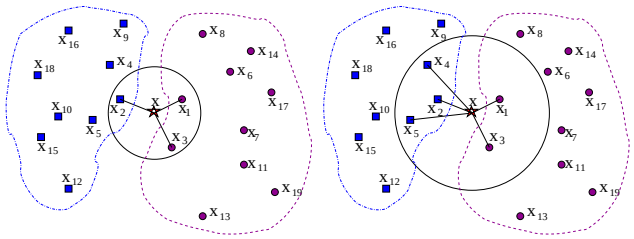
5.1 Nearest neighbor classification methods

As in any pattern classification problem, once the feature space has been determined, there are various choices of classifiers as mentioned in Chapter 1. Conceptually, the simplest classifier is perhaps the k-nearest-neighbor rule, which is a sub-optimal

procedure. It only requires a finite reference sample of N ($N > k$) feature matrices (feature vectors are special cases) labeled according to the pattern class of origin, and a dissimilarity measure in the space of feature matrices. For a given input feature matrix, the algorithm uses the given dissimilarity measure to first identify the k feature matrices from the reference samples which are closest to the input matrix and then assigns the input feature matrix to the pattern class that appears most frequently amongst the k nearest neighbors.

5.1.1 k-nearest neighbor classification algorithm [20, 25]

For our case of EEG signal classification, we take the feature PSD matrix of a test signal epoch not being part of the library, and compare the Riemannian distance of this test feature PSD matrix to its k nearest neighbors. Then we assign it to a class according to majority decision among these k neighbor matrices. Fig. 5.1.1 shows an example of 3-NN and an example of 5-NN in a two-class case. (In our case, the neighbors are the feature PSD matrices from the library signal sets.)



(a) Line segment (b) Circle

Figure 5.1: k-Nearest Neighbor Decision a) k = 3; b) k = 5

We can see that the assignment of the object $\mathbf{x}$ may vary with the choice of different values of k, regardless of whether the distributions of the objects are similar or different. However, there is no general rule to choose the best value of k in the

$k-$nearest neighbor algorithm. If *the sample size is infinite*, the larger is k the better is the performance of the k-nearest neighbor classifier. In fact, for infinitely large sample-size, the performance of the k-nearest neighbor algorithm has been shown to approach the optimum Bayesian classifier with $k \to \infty$ and $k/N \to 0$ (N being the sample size) [30].

In our tests, we first set up a library of epochs of EEG signals and categorize them into $L = 6$ classes, each representing a particular stage of sleep. Each epoch of EEG signals has been examined by clinical experts and classification agreements have been obtained. Using the procedure described in Section 2.4, the PSD matrices of these signal epochs in each of the categories are evaluated at each frequency point within the range $\omega \in [0\text{Hz}, 30\text{Hz}]$ forming different categories of curves (sequences of points), $\{\mathbf{P}_n(\omega), n = 1, \cdots N\}$. These are the PSD matrix curves to which we apply the k-nearest neighbor algorithm coupled with the Riemannian (or otherwise) distances for classification of the EEG signals. Since our sample size is finite, we found that, by choosing a small value of k, the results are very satisfactory. In the following, we summarize the classification of EEG signals using the k-nearest neighbor algorithm coupled with the weighted Riemannian distance d_{R1W}. (For classification using the k-nearest neighbor with other weighted or unweighted distances, the procedure will be identical):

1. With all the PSD matrices of EEG signal epochs of the L states of sleep in the library, the optimum weighting matrix $\mathbf{W}$ of similarity/dissimilarity is evaluated for d_{R_1W} according the description in Section 4.4.1.

2. For the PSD matrix curve $\mathbf{P}_0(\omega)$ of a test EEG signal, we calculate the dissimilarity measures $\{d_{ni} = d_{R_1W}(\mathbf{P}_0(\omega_i), \mathbf{P}_n(\omega_i)),\ n = 1, \cdots, N\}$ at each frequency ω_i according to Eqs. (4.52) and (4.57), and then calculate the total distance d_n

between the two curves according to Eq. (3.108). For a chosen value of k, the k nearest neighbors of $\mathbf{P}_0(\omega)$ of the test signal (k PSD matrices at same ω having shortest weighted distances from $\mathbf{P}_0(\omega)$) are then identified.

3. $\mathbf{P}_0(\omega)$ is then assigned to class $\mathcal{C}_{\ell_0}$ if $\ell_0 = \text{maj}(\ell_1, \cdots, \ell_k)$ where $\ell_1, \cdots, \ell_k$ are the class labels of the k-nearest neighbors of $\mathbf{P}_0(\omega)$ among the members of the library, and $\text{maj}(\cdot)$ denotes the majority vote function, i.e., its value is the element which has occurred most in $\{\ell_1, \cdots, \ell_k\}$.

5.2 Q-fold cross-validation method

The above description outlines the procedure of applying the k-nearest neighbor algorithm together with the Riemannian distance to classify an unknown signal to a particular state of sleep. In this section, we will examine the performance of our classification method. Ideally, the performance accuracy of our EEG classification algorithm should be measured in terms of its probability of error which necessitates the knowledge of the ground truth of the patient's state of sleep. However, since the ground truth of the state of sleep of a patient measured from the signal epoch is not really known, we will therefore treat the library of signal epochs classified by clinical experts as the ground truth. From the library of collected signal epochs, we will randomly select some as training signals and some as test signals so that the validation of our classification methods is carried out as follows:

i) For each of the classes $\mathcal{C}_\ell, \ell = 1, \cdots, L$, containing N_ℓ feature PSD matrix curves (being functions of ω) of the same state of sleep, we randomly choose $N_{\ell T}$ matrix curves as the test set and the rest ($N_\ell - N_{\ell T}$) as the training (library) set.[1]

[1]In an actual clinical test in which a patient's EEG signal is under examination, the number $N_{\ell T}$ of test curves does not affect the test outcome since the epochs are tested one at a time against a

ii) As described in the previous section, for all the L states of sleep, the weighting matrix $\mathbf{W}$ is first evaluated using the training sets, each containing $(N_\ell - N_{\ell T})$ selected feature matrix curves. For each member matrix curve of the test sets, the dissimilarity measures from the library sets are calculated and its classification is carried out according to the k-nearest neighbor algorithm.

iii) The above steps are repeated Q times (Q-fold cross-validation), each time choosing different sets of training and test feature matrix curves in $\mathcal{C}_\ell$. The probability of correct classification in each state can then be estimated by $\hat{P}_{c\ell} = \frac{1}{Q}\sum_{q=1}^{Q}\hat{P}_{c\ell q}$ where $\hat{P}_{c\ell q}$ denotes the estimated probability of correct classification of class ℓ at the qth trial, $q = 1, \cdots, Q$, i.e. $\hat{P}_{c\ell q} = \frac{N_{c\ell}}{N_{\ell T}}$ with $N_{\ell c}$ being the number of correct classification for class $\mathcal{C}_\ell$ at the qth trial.

5.3 Validation test results

We now perform some tests using the collected sleep data to validate our classification algorithm employing the Riemannian distance developed in Chapter 3 and Chapter 4. The test results are based on the data collected from five patients. For each patient, we collect the multichannel recordings for each sleep state. The recordings were selected from channels $(C_3 - A_2)$, $(C_4 - A_1)$, $(O_1 - A_2)$, and $(O_2 - A_1)$[2] for all patients. As described in the previous chapters, for our validation tests, the raw EEG recordings were first pre-processed by removing the DC values, and the frequency components of the signals were kept to within the range of $0.5 - 30Hz$ by using a bandpass filter. We sectioned the recording length to $30s$ epochs. Each epoch was examined by clinical

library of reference signals. However, in our validation test here, taking away $N_{\ell T}$ signals from a group results in a depletion of the library reference signals. For a finite number of reference signals, the performance of the classification algorithm may well be affected by $N_{\ell T}$ as will be demonstrated in the next section.

[2]Please refer to Section 1.1.1 for the positioning of these sensors.

experts, and upon agreement, classified into one of six states. For each sleep state we collected 75 epochs for a total of 450 epochs in all states and used them in the verification of the methods. Also, the power spectral density matrix of each epoch was estimated by the Nuttall-Strand algorithm [68,79]. In each trial, we randomly choose $N_{\ell T}$ PSD matrices from each state as test signals while the remaining $(75 - N_{\ell T})$ PSD matrices form the training data set so that the total number of the training feature signals in each trial is $6 \times (75 - N_{\ell T})$ for each trial.

The following are examples of the tests of the effectiveness of various dissimilarity measures in the classification of EEG signals we carried out under different environments.

5.3.1 Example 5.1

We first examine the performance of our classification algorithm using either the Riemannian distances d_{R_1} and d_{R_2} or the weighted Riemannian distances d_{R_1W} and d_{R_2W}. Our experiments are carried out with $N_{\ell T} = 1, 5$, and 15, for each of which we employ the parameter $k = 1, 3, 5$, and 7 for the nearest neighbor tests. Each test is repeated $Q = 75$ times. Fig. 5.2, Fig. 5.3 and Fig. 5.4 show the performance of the methods using d_{R_1} and d_{R_1W} under different parameter values. Comparing the results in the three figures, clearly, the weighted Riemannian distance outperforms the unweighted one by a margin of 8% to 10% in accuracy of classification. It is also observed that the cases of having the number of nearest neighbors being $k = 5$ and 3 seem to have, on average, good performance in the three figures shown. As the number of selected test signals $N_{\ell T}$ increases, we can see that the performance of all the cases deteriorate. In the case of the weighted Riemannian distance, the accuracy deteriorates from a high-90% for $N_{\ell T} = 1$ to a low-90% using $k = 5$ for $N_{\ell T} = 15$. Other cases of k have similar drops in performance as $N_{\ell T}$ increases. This is because

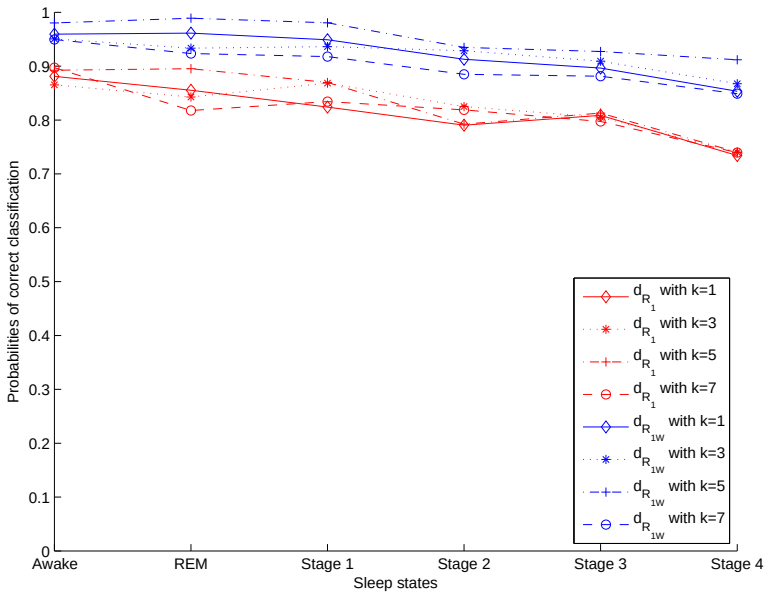


Figure 5.2: Classification results using d_{R_1} ($N_{\ell T} = 1$, $k = 1, 3, 5, 7$)

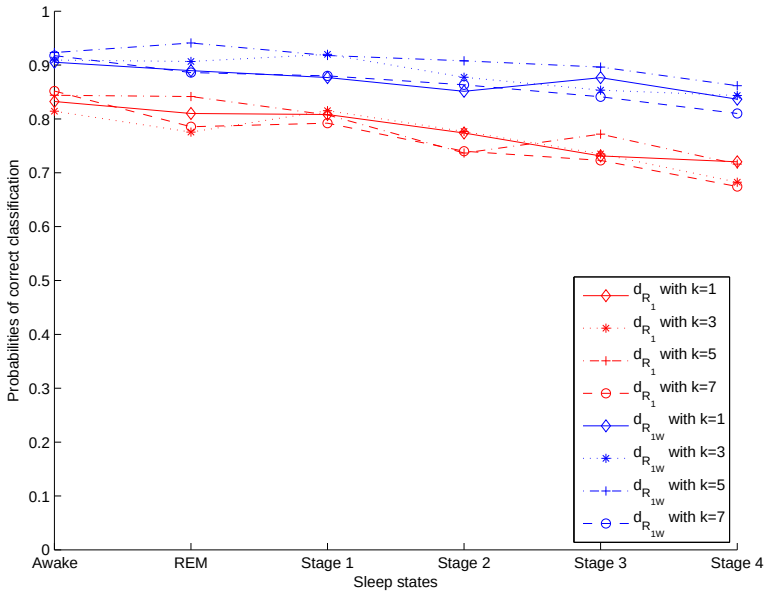


Figure 5.3: Classification results using d_{R_1} ($N_{\ell T} = 5, \quad k = 1, 3, 5, 7$)

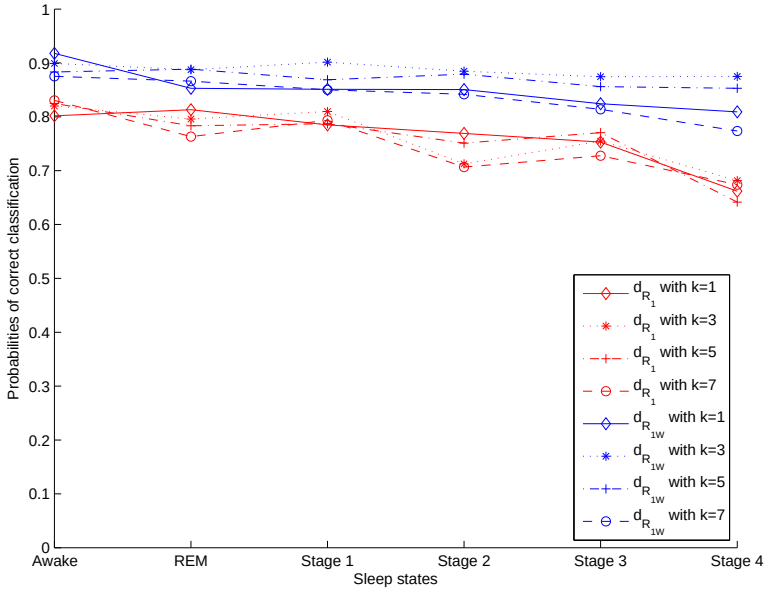


Figure 5.4: Classification results using d_{R_1} ($N_{\ell T} = 15, \quad k = 1, 3, 5, 7$)

larger $N_{\ell T}$ depletes the size of the training (library) set and the ratio of k/N in the nearest neighbor test is no longer small enough.

The performance using Riemannian distances d_{R_2} and d_{R_2W} are shown in Fig. 5.5, Fig. 5.6, and Fig. 5.7. We also tested the performance of the classification algorithm using Riemannian distances d_{R_3}. Since it has been shown in Appendix F that d_{R_3} is weight-invariant, the performance using d_{R_3W} is omitted and the performance of the classification using d_{R_3} is shown in Fig. 5.8, Fig. 5.9, and Fig. 5.10 For these two other Riemannian distances, similar observations as for d_{R_1} and d_{R_1W} are noted.

We note that when the validation is performed in the cases with $N_{\ell T} > 1$, the classification is carried out on each individual member of the test signals, i.e., the classification is carried out as if $N_{\ell T} = 1$. The only difference between the case of $N_{\ell T} > 1$ and $N_{\ell T} = 1$ is that the groups from which the test signals have been selected would have fewer members left as library reference. Therefore, from the above observations

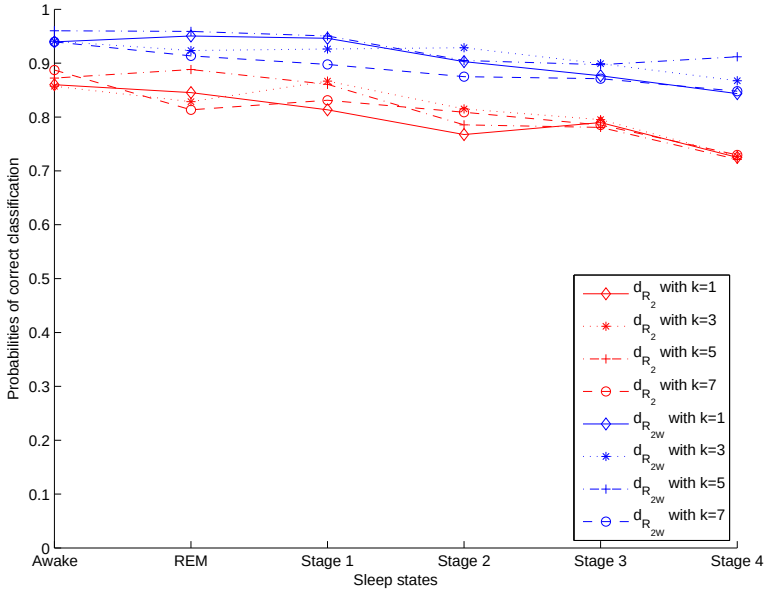


Figure 5.5: Classification results using d_{R_2} ($N_{\ell T} = 1$, $k = 1, 3, 5, 7$)

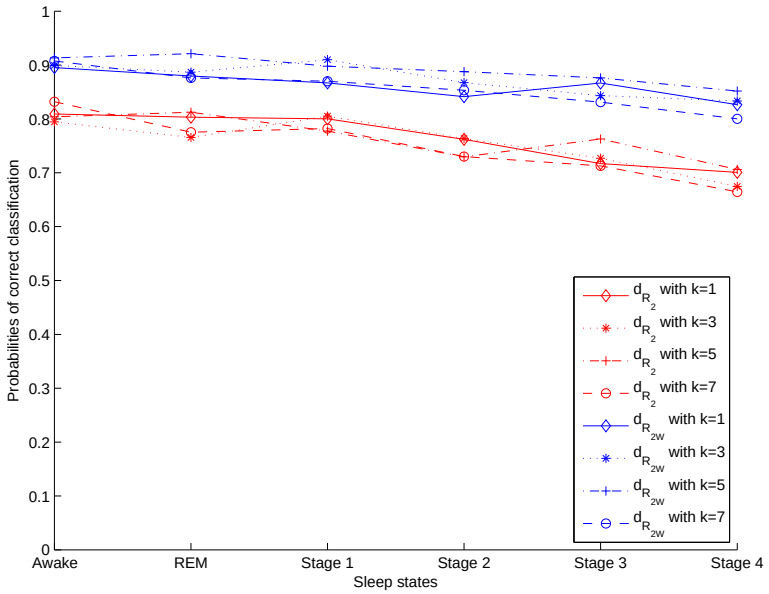


Figure 5.6: Classification results using d_{R_2} ($N_{\ell T} = 5, \quad k = 1, 3, 5, 7$)

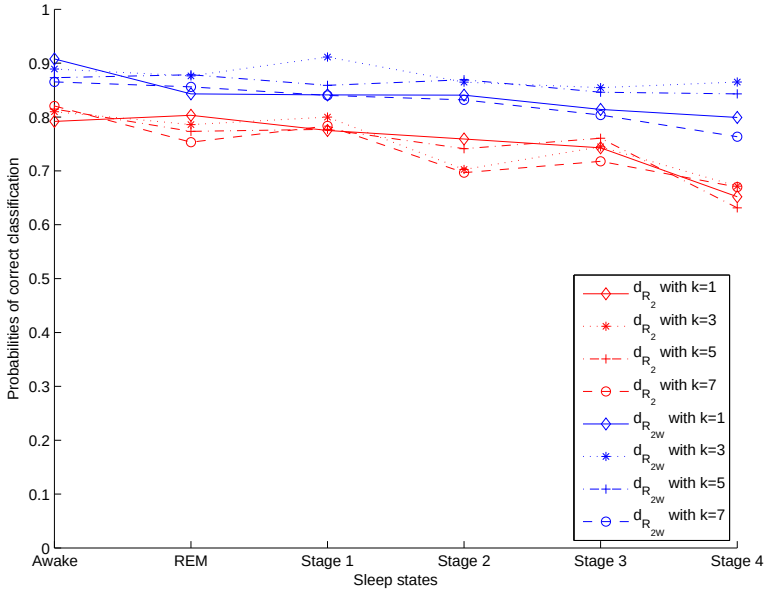


Figure 5.7: Classification results using d_{R_2} ($N_{\ell T} = 15, \quad k = 1, 3, 5, 7$)

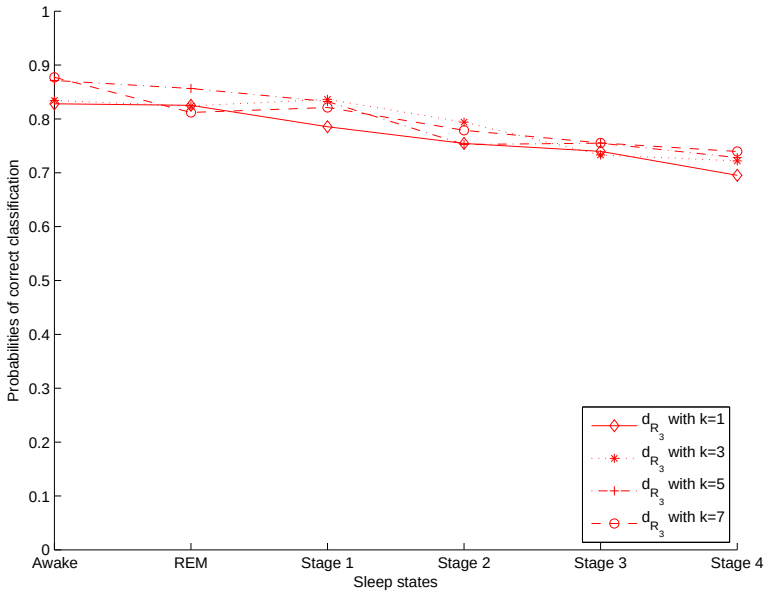


Figure 5.8: Classification results using d_{R_3} ($N_{\ell T} = 1$, $k = 1, 3, 5, 7$)

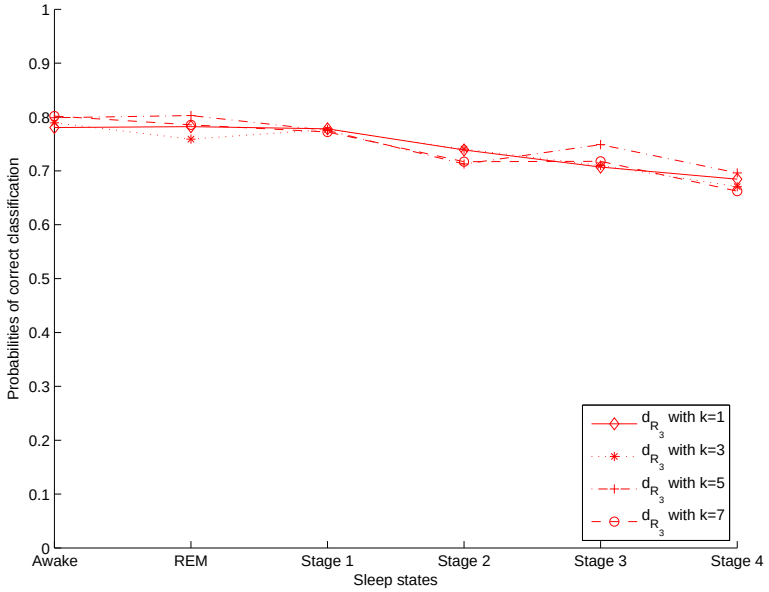


Figure 5.9: Classification results using d_{R_3} ($N_{\ell T} = 5, \quad k = 1, 3, 5, 7$)

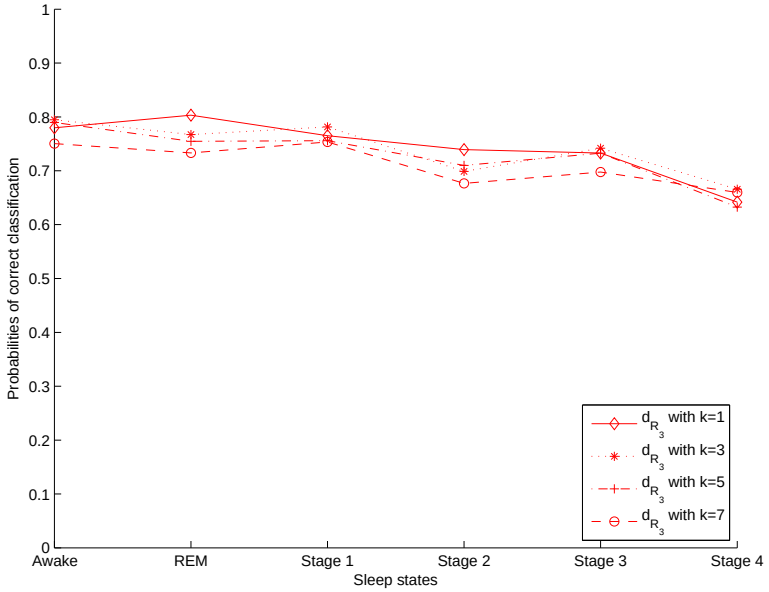


Figure 5.10: Classification results using d_{R_3} ($N_{\ell T} = 15, \quad k = 1, 3, 5, 7$)

on the performance of the classification algorithm using the different Riemannian distances, we can conclude that their performance are all similarly affected by the choice of the ratio of the number of test signals to the number of reference signals, as well as by the number of nearest neighbors. Increasing the number of test signals taken from a class depletes the number of reference signals in that class and the evaluation of similarity will be affected. As well, the increase of the number of nearest neighbors in the algorithm may violate the condition $k/N \to 0$ for good performance of the algorithm. Since the performance using the different Riemannian distances are all similarly affected, we can see that the performance of the algorithm under the effect of choices in $N_{\ell T}$ and in k does not depend on the definition of the distance, rather it depends on the relative size of $N_{\ell T}$ and k to the total library size. □

5.3.2 Example 5.2

In this example, we carry out a direct comparison between the performance of the Riemannian distances d_{R_1}, d_{R_2} and d_{R_3} and their weighted versions when applied to the EEG signal classification problem. Due to its weight-invariant nature (see Chapter 3), therefore, no weighting is needed for d_{R_3}. Figs. 5.11, Fig. 5.12 and Fig. 5.13 show the performance with different $N_{\ell T}$ and k. The test is repeated $Q = 75$ times. Also, since from the last example, the cases of $k = 3$ and 5 show good performance for the Riemannian distances, we maintain the use of these two choices of the number of nearest neighbors in our tests here. For the unweighted Riemannian distances, we can see that d_{R_1} and d_{R_2} generally have better performance than d_{R_3}. Comparing the results in these figures, we can see that for $N_{\ell T} = 1$, d_{R_1} and d_{R_2} have similar performance, both having accuracies in the mid-80% to high-80% while d_{R_3} is generally around 3 to 5% lower. As $N_{\ell T}$ increases, all the Riemannian distances yield deteriorated performance as observed in Example 4.1. Optimum weighting results in

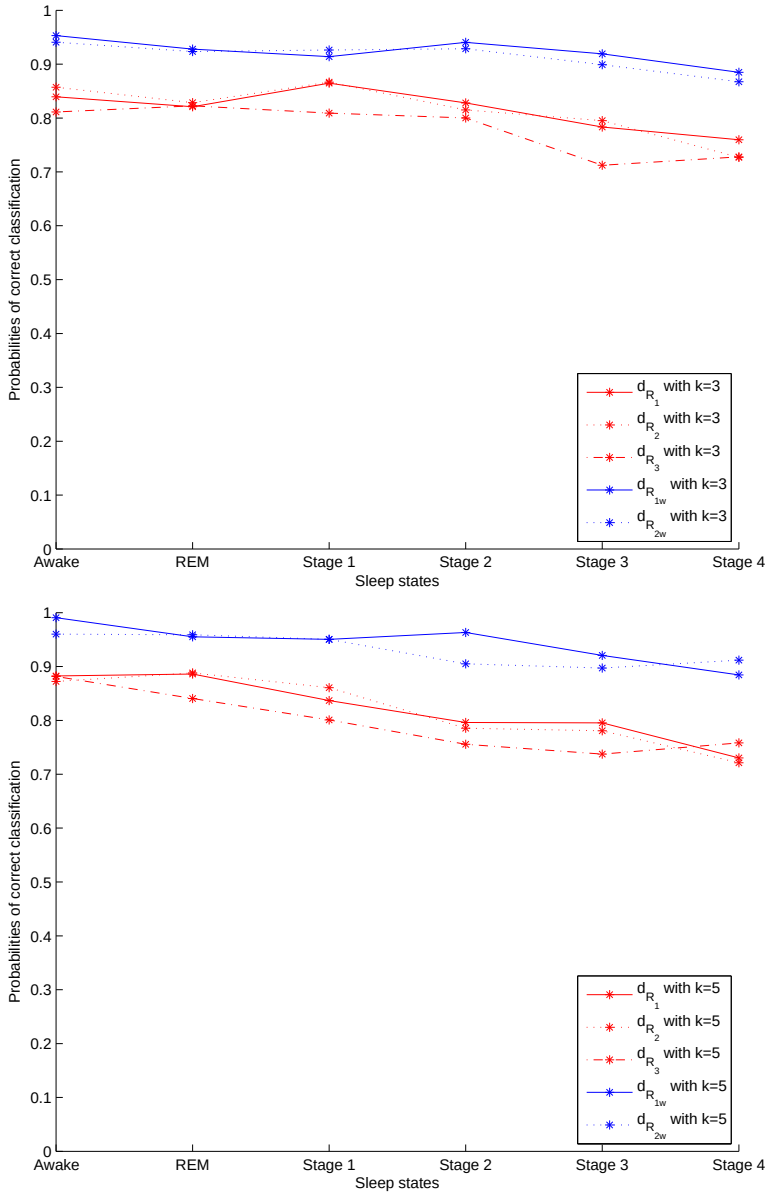


Figure 5.11: Performance of using Riemannian distances for $N_{\ell T} = 1$: a) k = 3, b) k = 5

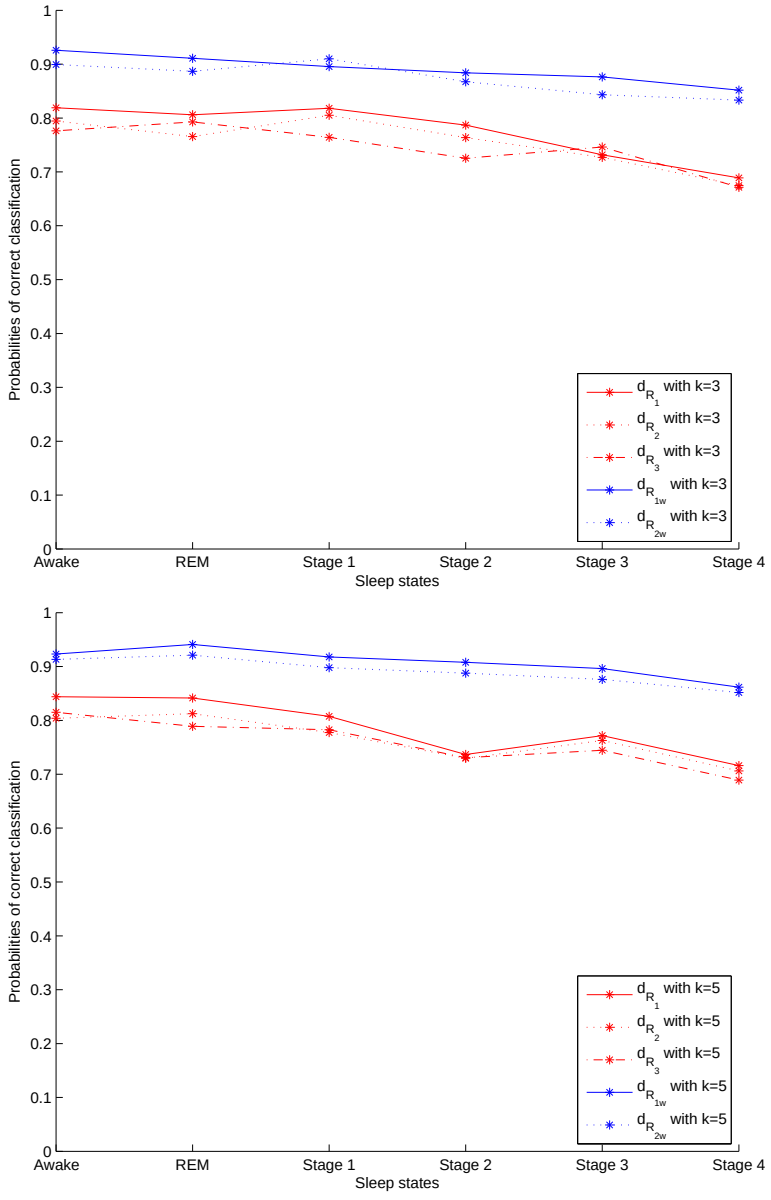


Figure 5.12: Performance of using Riemannian distances for $N_{\ell T} = 5$: a) k = 3, b) k = 5

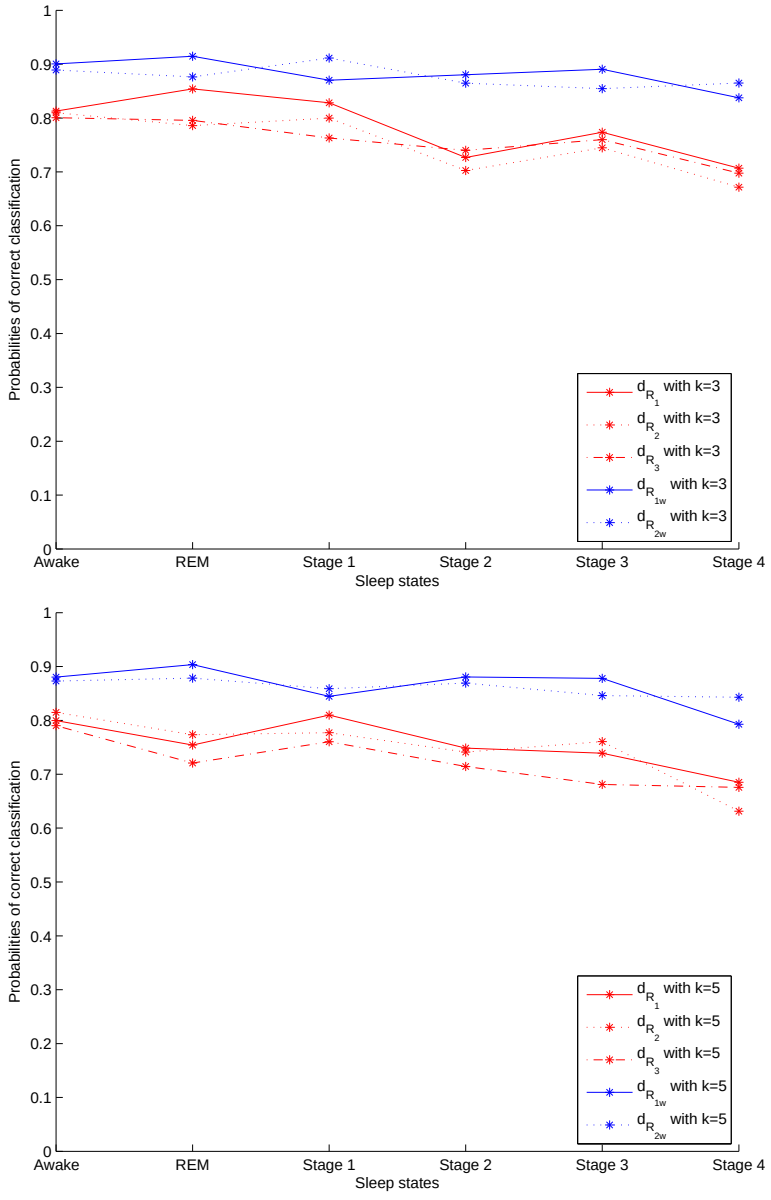


Figure 5.13: Performance of using Riemannian distances for $N_{\ell T} = 15$: a) k = 3, b) k =5

d_{R_1W} and d_{R_2W} having much enhanced performance, having accuracies both in the high 90% for $N_{\ell T} = 1$, and deteriorated to around 90% when $N_{\ell T}$ increases.

5.3.3 Example 5.3

We now compare the performance of the classification algorithm using the unweighted Riemannian distance measures to that using various other unweighted distances. For comparison, we have chosen the following distance measures:

a) K-L divergence d_{KL} – The K-L divergence has been introduced in Section 3.2.2. We have seen that it does not satisfy the triangular inequality and is therefore not a true distance. Furthermore, the "measure" is invariant to weighting. However, it uses the power spectral density matrices as the feature and has been applied to the classification of EEG signals. Here, we also include this measure and examine its effectiveness in our study of EEG signal classification.

b) Euclidean distance d_{E1} – The distance measure used here is the unweighted Euclidean distance between two vectors $\mathbf{v}_{\mathbf{P}_{\mathrm{a}}1}$ and $\mathbf{v}_{\mathbf{P}_{\mathrm{a}}2}$ generated as shown in Eq. (2.64) by vectorizing the two PSD matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ respectively.

c) Euclidean distance d_{E2} – The distance measure used here is the unweighted Euclidean distance between two vectors $\mathbf{v}_{\mathbf{P}_{\mathrm{L}}1}$ and $\mathbf{v}_{\mathbf{P}_{\mathrm{L}}2}$ generated as shown in Eq. (2.85) by utilizing the parameters of the Lie vectors of the two PSD matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ respectively.

The comparison of the performance of the classification algorithm employing the unweighted Riemannian distances with those using the above distances are shown in Fig. 5.14, Fig. 5.15, and Fig. 5.16 for various values of $N_{\ell T}$ and k. The test is repeated $Q = 75$ times. It can be observed that while the performance using the

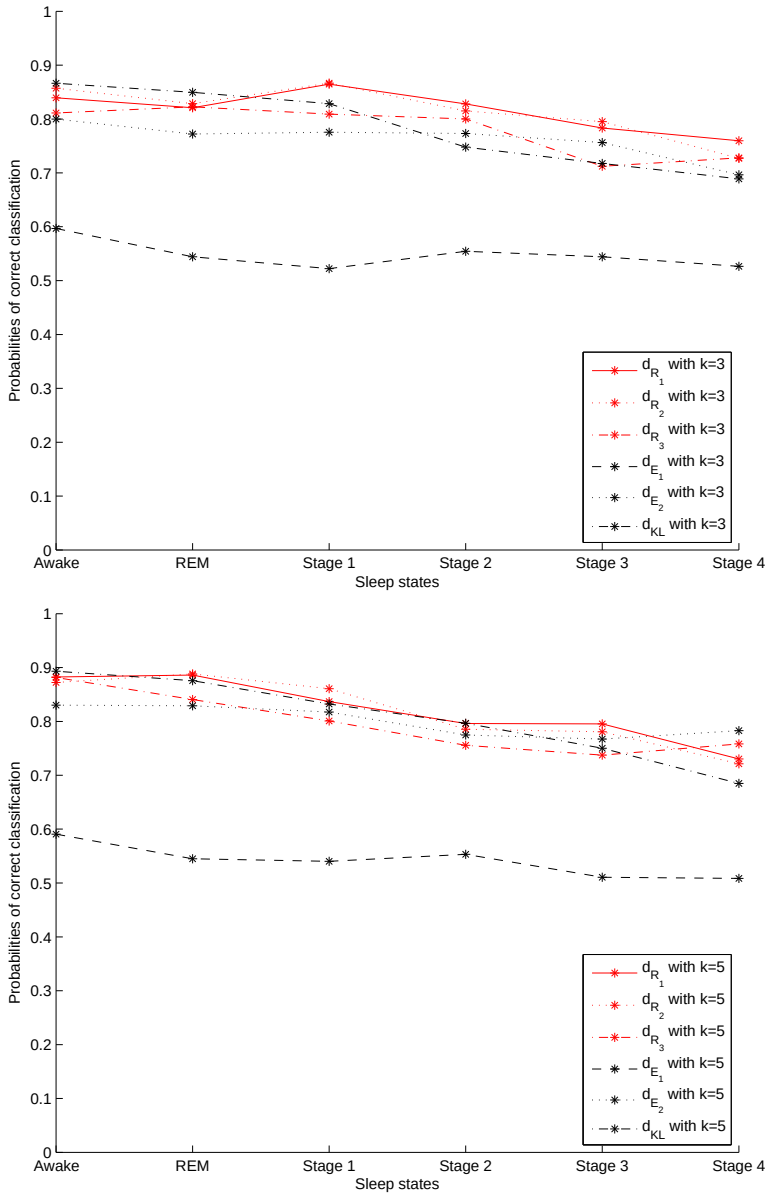


Figure 5.14: Performance using various unweighted distances for $N_{\ell T} = 1$: a) k = 3, b) k = 5

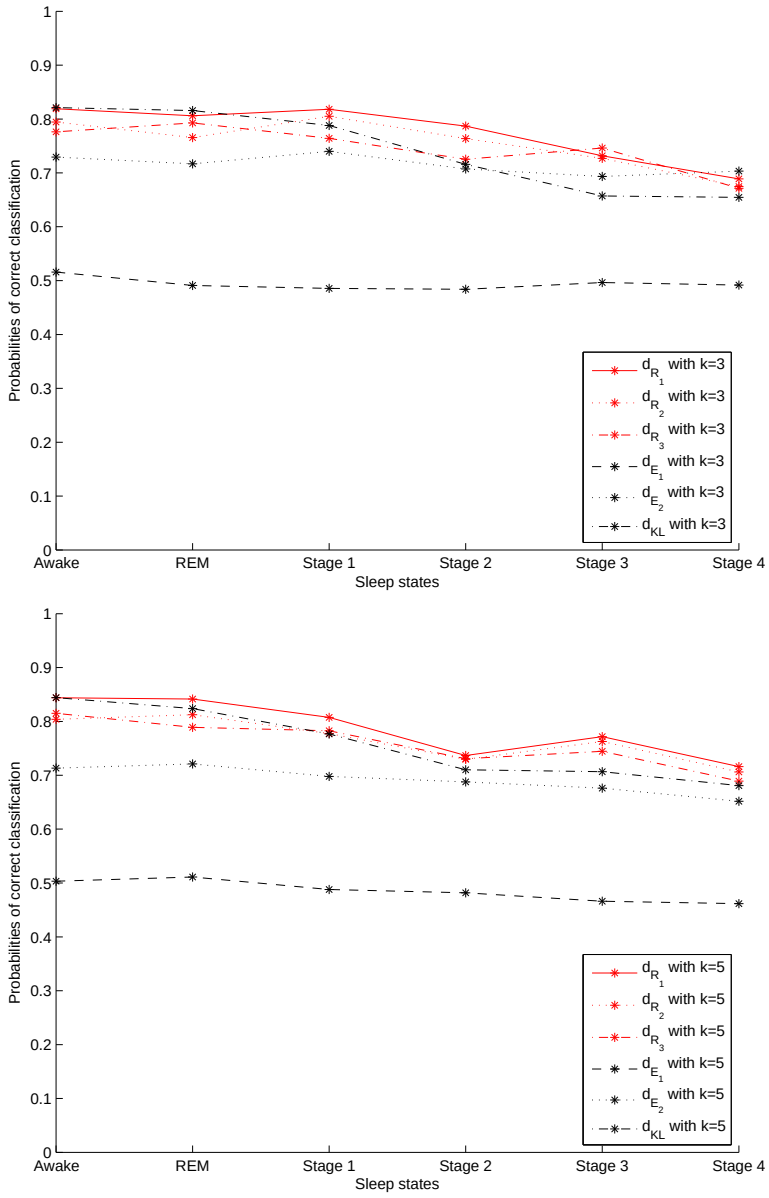


Figure 5.15: Performance using various unweighted distances for $N_{\ell T} = 5$: a) k = 3, b) k = 5

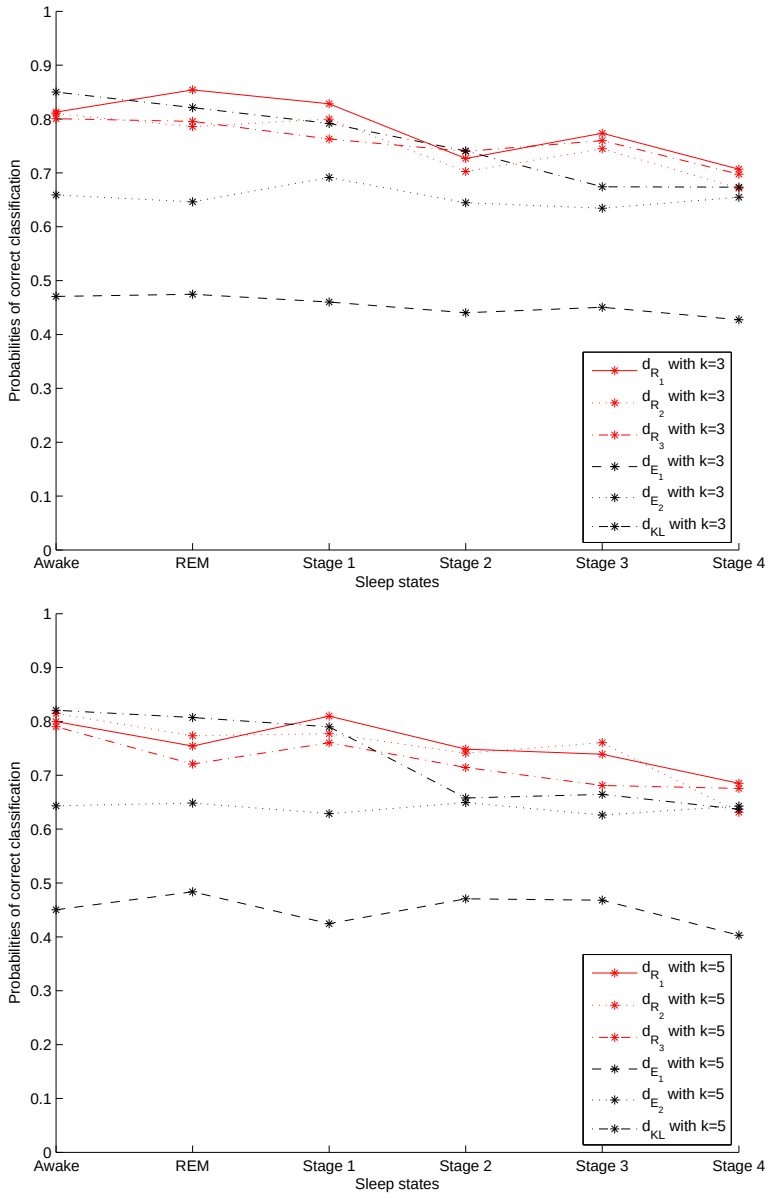


Figure 5.16: Performance using various unweighted distances for $N_{\ell T} = 15$: a) k = 3, b) k =5

other distance measures performs well and cluster around the low-80% mark, the unweighted Euclidean distance d_{E1} is very much worse, its accuracy being around the high-50%. The inferior performance of the algorithm using d_{E1} is apparent. d_{E1} measures the distance between two vectors formed by stacking up the elements of the PSD matrices. In other words, the information of the relationship between the elements is no longer available. In contrast, the other distance measures retain all the information of the PSD matrix, not only the element values, but also the structure and properties of the matrix.

As $N_{\ell T}$ increases and the number of reference signals in the library decreases, all the performance deteriorate as in Examples 1 and 2. However, among the better performance group of measures, d_{E2} appears to be more sensitive to the decrease of reference signals. Its performance accuracy drops to around 70% for $N_{\ell T} = 5$ and further to around 60% for $N_{\ell T} = 15$. Thus, the performance of the group using distances directly expressed as functions of the PSD matrices seem to be more robust against changes in statistical environments.

Apart from the information of the PSD matrix elements and structures, the Riemannian distances also explore the geometry of the manifold that is described by the PSD matrices. It is not surprising, therefore, to find the performance associated with these distance measures to be superior to those without utilizing the manifold geometry. However, we also observe that even though the K-L distance is not derived using the manifold geometry, its performance is by and large, comparable to the performance using the Riemannian distances. This should not be so surprising either since an examination of the K-L distance in Eq. (3.21) shows the similar employment of the traces of the matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ as in d_{R_1} and d_{R_2}.

In spite of this similarity of using the traces of the PSD matrices as distance measures, the K-L distance unfortunately cannot be weighted to enhance the measure

of similarity and dissimilarity of different signal classes, making it not so useful in this application.

5.3.4 Example 5.4

Optimally weighted distances, in general, have superior performance to the unweighted ones in EEG signal classification. This has been clearly observed in the cases of weighted and unweighted Riemannian distances in Examples 1 and 2. Here in this example, we compare the performance of the optimally weighted Riemannian distances d_{R_1W} and d_{R_2W} derived in this thesis with other optimally weighted distances for the purpose of EEG signal classification. Now, since optimum weighting in EEG classification essentially enhances the measure of similarity/dissimilarity between the groups of signals, we will examine the two approaches addressed in Chapter 4:

a) X-N-J optimum weighting of vectors [90] – This method of optimally weighting a signal vector has been introduced in Chapter 4. Here, we apply the X-N-J optimum weight obtained from Table 4.2 to vector representations of $\mathbf{P}$ as in Example 4.3, i.e.,

   (i) Euclidean distance between two vectors $\mathbf{v}_{\mathbf{P}_{\mathrm{a}}1}$ and $\mathbf{v}_{\mathbf{P}_{\mathrm{a}}2}$ generated as shown in Eq. (2.64) by vectorizing the two PSD matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ respectively.

   (ii) Euclidean distance between two Lie vectors $\mathbf{v}_{\mathbf{P}_{\mathrm{L}}1}$ and $\mathbf{v}_{\mathbf{P}_{\mathrm{L}}2}$ of the two PSD matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ respectively.

b) Weighted Riemannian distances – These are the same weighted Riemannian distances d_{R_1W} and d_{R_2W} which have been studied in Example 4.1.

Again, we apply the k-nearest neighbor algorithm through the Q-fold validation process to all the tests. Our experiments are carried out with $N_{\ell T} = 1, 5$, and 15 for all

the methods and we choose $Q = 75$. We only show the performance comparison for $k = 3$ and 5. Again, Figs 5.17, 5.18, and 5.19 show the comparison of performance of the various methods under different parameter values.

It can be observed from the figures that the performance of all the methods are greatly enhanced form that of their unweighted counter-parts in Example 4.3. For the two Euclidian distances d_{E_1} and d_{E_2}, their performance have been elevated respectively from high-50% to around 70% and from under 80% to around 85% for $N_{\ell T} = 1$. For the two Riemannian distances d_{R_1} and d_{R_2}, their performance have both been elevated from around 85% to over 95% for $N_{\ell T} = 1$. Judging from the results in Example 4.3 and 4.4, we can say that the performance using the X-N-J weighted Lie vector is only comparable to that of the unweighted Riemannian distances. On the other hand, the optimally weighted Riemannian distances yield a performance clearly superior to the optimally weighted d_{E_2W} by a margin of 8 to over 15% in accuracy of classification, and by an even greater margin when compared to d_{E_1W}. As the number of selected test signals $N_{\ell T}$ increases and the number of reference signals decreases, we can see that the performance of all the methods deteriorate, however, the margin of superior performance for the weighted Riemannian distance over the optimally weighted Euclidian distances still maintains. □

Many other examples with different parameters have also be tested and similar observations have been noted.

5.4 k-NN classification for large size data library

The above experiments have been carried out when the available data is limited and the reference library is relatively small. The effect of such limited data is quite apparent from the deterioration of performance of all the methods when the number

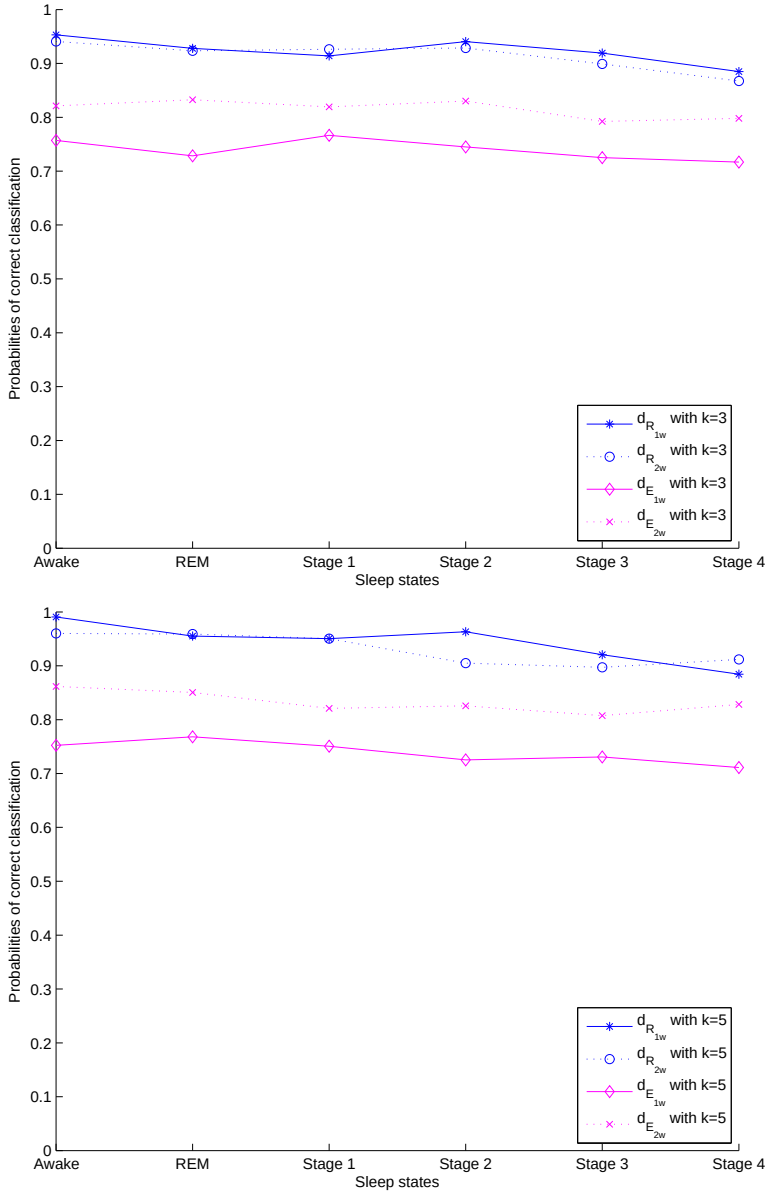


Figure 5.17: Performance using weighted distances for $N_{\ell T} = 1$: a) k = 3, b) k = 5

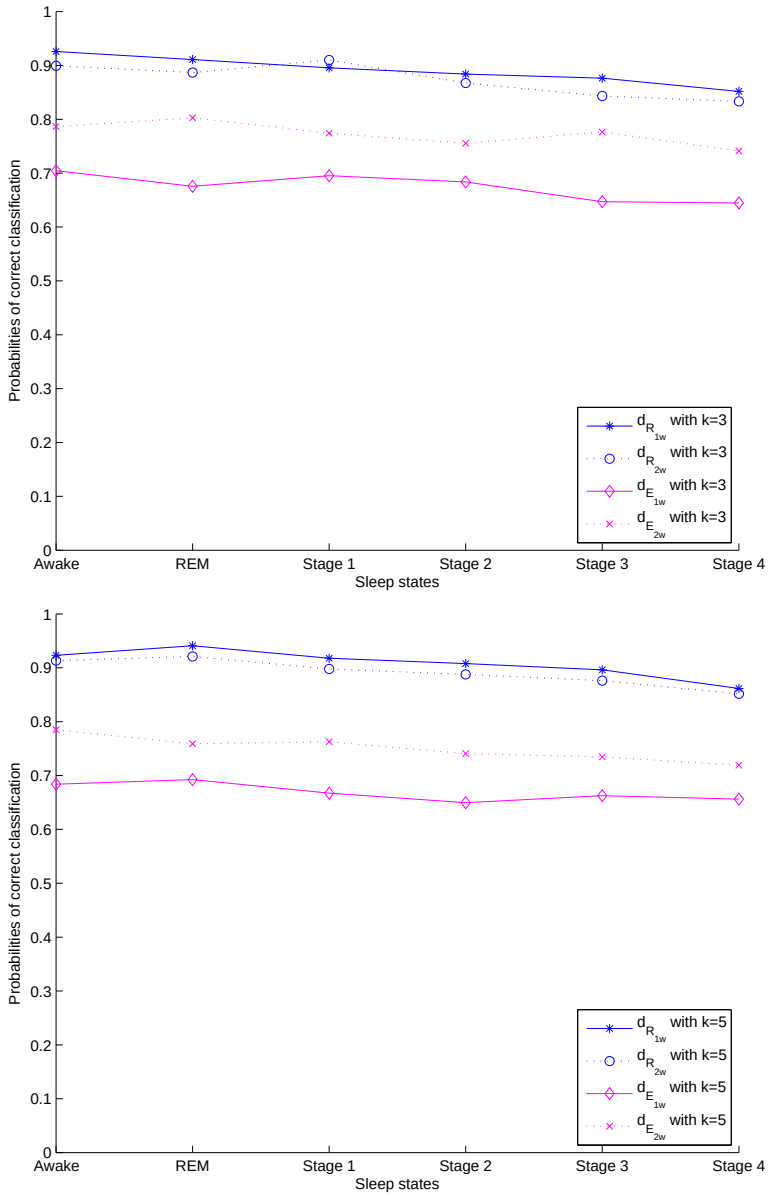


Figure 5.18: Performance using weighted distances for $N_{\ell T} = 5$: a) k = 3, b) k = 5

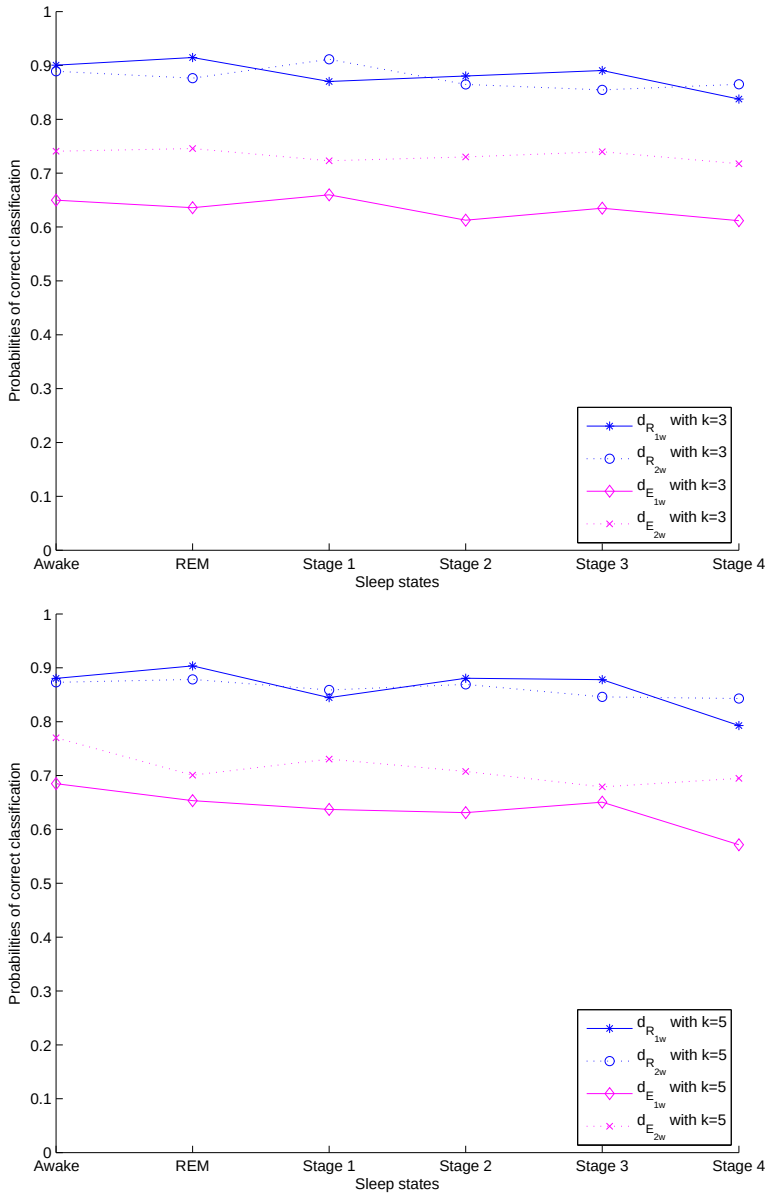


Figure 5.19: Performance using weighted distances for $N_{\ell T} = 15$: a) k = 3, b) k =5

of reference signals decreases and when the ratio of k/N increases. Under ideal circumstances, we should have a very much larger library of signals which would render the statistics of classification more stable. However, when the experiments are performed using a large amount of library data, then in order to make a decision, the k-nearest neighbor algorithm would have to use all of the library patterns of a class as the class representation so that the distances between the pattern to be tested and every pattern of the class has to be computed according to a dissimilarity measure. Although the k-nearest neighbor algorithm is simple and reliable, when the sample size of each class is large, the number of distances to be calculated is very large as well. This leads to a fundamental problem of how to reasonably represent each class of the data.

One way to overcome this difficulty is to divide the class of data into sub-classes and the mean of each sub-class is computed and used to represent the original class. In this way, the class can be reasonably characterized by its approximate data distribution. We call this the *multi-mean representation of classes.*

Since our training data are matrices on a manifold measured by a weighted Riemannian distance, the mean should take into account of the natural properties of the manifold. Hence, the usual definition of "mean" using Euclidean distance may not be appropriate. Instead, we may employ the *Karcher mean* [54] in our grouping of the power density matrices, i.e., we should solve the following optimization problem to find the Karcher mean of the sub-class $\mathcal{S}_{\ell j}$ of the power spectral density matrices belonging to $\mathcal{C}_\ell$:

$$\arg\min_{\mathbf{P}\in\mathcal{S}_{\ell j}} \sum_{\mathbf{P}_{ji}\in\mathcal{S}_{\ell j}} d^2_{GW}(\mathbf{P},\mathbf{P}_{ji}) \tag{5.1}$$

Solving this problem, however, is not easy in general.

Instead of trying to solve this problem exactly we choose the element $\mathbf{P}_{ji_0}$ in $\mathcal{S}_{\ell j}$

such that

$$i_0 = \arg\min \sum_{k \neq i_0,\ \mathbf{P}_{jk}, \mathbf{P}_{ji_0} \in \mathcal{S}_{\ell j}}^{k} d^2_{GW}(\mathbf{P}_{jk}, \mathbf{P}_{ji_0}) \tag{5.2}$$

and let $\bar{\mathbf{P}}_{\ell j} = \mathbf{P}_{ji_0}$ indicate the approximate Karcher mean of the sub-class $\mathcal{S}_{\ell j}$.

The algorithm of finding the Karcher means in class $\mathcal{C}_\ell$ consists of the following steps: First, we randomly choose the elements $\mathbf{P}_{j0}; j = 1, \cdots, J_\ell$ from the class $\mathcal{C}_\ell$ to form an initial set $\mathcal{G}_\ell^{(0)}$ of Karcher means of J_ℓ sub-classes of elements. Second, each of these initial sub-classes $\mathcal{S}_{\ell 1}^{(0)}, \cdots \mathcal{S}_{\ell J_\ell}^{(0)}$ is assigned elements closest to the jth Karcher mean $\mathbf{P}_{j0} \in \mathcal{G}_\ell^{(0)}$. Third, the approximate Karcher mean of each of the J_ℓ sub-classes are recalculated from the elements assigned to that sub-class. Fourth, the second and third steps are repeated until convergence occurs and the proper sub-classes $\mathcal{S}_{\ell 1}, \cdots \mathcal{S}_{\ell J_\ell}$ and the group $\mathcal{G}_\ell$ of Karcher means are established.

We are now ready to carry out the classification using the k-nearest neighbour rule. The procedure is given as follows:

1. For each of the classes $\mathcal{C}_\ell$ containing N_ℓ patterns of power spectral density (being functions of ω) of the same state of sleep, we randomly choose $N_{\ell T}$ patterns as the training set, while the rest are chosen as the testing set.

2. For all of the $N_{\ell T}$ patterns representing the training sets of all the L states of sleep, the weighting matrix $\mathbf{W}$ is first evaluated.

3. For each training set in class $\mathcal{C}_\ell$, we find the group $\mathcal{G}_\ell$ of J_ℓ approximate Karcher means using the weighted Riemannian distance.

4. For each pattern of the testing sets, we find the dissimilarity between its power spectral density $\mathbf{P}_0$ and the Karcher means in all the groups $\mathcal{G}_\ell, \ell = 1, \cdots, L,$.

5. Each $\mathbf{P}_0$ is then assigned to class $\mathcal{C}_{\ell_0}$ if

$$\ell_0 = \text{maj}(\ell_1, \cdots, \ell_k) \tag{5.3}$$

where $\ell_1, \cdots, \ell_k$ are the class labels of the k-nearest neighbors of $\mathbf{P}_0$ among the Karcher means, and $\text{maj}(\cdot)$ denotes the majority vote function, i.e., its value is the element which has occurred most in $\{\ell_1, \cdots, \ell_k\}$.

For the same data set as introduced in the previous Section, the rudimentary validation test results show that there are no significant deterioration in the the classification performance when 30 approximate Karcher means are calculated for each class as the class representatives. Therefore, this method has potential advantages in the case of large data size.

Chapter 6

Summary, Future Works, and Conclusions

6.1 Summary of thesis

In this thesis, we examine the problem of the classification of sleep states of a patient by analyzing his/her EEG signals. We focus on the geometrical aspects in signal analysis and emphasize on the improvement of signal classification by exploitation of the geometry of the signal space.

Following the practice of clinical experts whose judgements in sleep state classification are based essentially upon the power contents of the signals in the various frequency ranges, we propose to employ the power spectral density (PSD) matrix as the feature for the distinction between different classes of EEG signals. In so doing, we not only examine the power spectrum contents of the signal from each channel, but also utilize the the cross power spectra between signals collected from the different channels. To facilitate the classification, we argue that since the PSD matrices are positive definite and exhibit certain geometric properties in the signal feature space,

the use of the most widely accepted Euclidean distance between the signal features may not be the most appropriate for measuring their differences. Rather, we propose that the geometric properties of the feature space be exploited and appropriate metrics on the manifold of the PSD matrices be developed using Riemannian geometry. By characterizing EEG signals with their power spectral functions, the dissimilarity measure is defined based on the geodesic distances on the manifold of positive definite Hermitian matrices.

A general form of geodesic distance on the Riemannian manifold is then derived. With the help of fibre bundle theory and a particular choice of the Riemannian metric, we develop a closed form, d_{R_1}, of the geodesic between two points on the manifold. This new distance measure is then related to the Fisher-Rao and Fubini-Study distances in special cases. We then show that this geodesic distance can also be obtained by mimicking the Fréchet distance between two covariance matrices. In addition, by another choice of the Riemannian metric, we show that a conjectured distance d_{R_2} is also a geodesic distance on the manifold. We further provide a new proof following our own geometric interpretations for the geodesic distance d_{R_3} which has been in existence in the literature.

For the newly derived Riemannian distances, we also propose a weighting method to facilitate the enhancement of certain parts of the features. To obtain a suitable weighting so that the new metrics can be applied effectively to EEG signal classification, we argue that the weighting should render the distances of similar features minimized while the distances for dissimilar features maximized. Pursuing along this line of thought, we develop a general formulation of the optimization problem of the weighting matrix. Focusing on the special case of this generalized problem formulation, closed forms of the weighting matrix for the Riemannian distances have been obtained by solving an approximate convex optimization problem.

Using the k-nearest neighbor decision, we test the effectiveness of these new metrics by applying them to a collection of recorded EEG signals for sleep pattern classification. The results are compared to those obtained by using other metrics, and it is observed that the weighted Riemannian distances d_{R1W} and d_{R2W} yield an accuracy of approximately 10% higher than methods using other metrics.

6.2 Further elaboration of work and future research

6.2.1 Elaboration of research results

To the best knowledge of the author, this thesis is a first attempt to exploit the geometric properties of the signal feature manifold for improving the decision of sleep state of a patient. On hindsight, there are parts of the research which could be improved. Some of these may be due to the limitation of time and man-power, others may be considered important but outside the focus of the thesis. These points, which have not been carried out as perfectly as could be and which may be elaborated further, are listed below:

E1. Artifacts removal: There are various sources from which artifacts arise in EEG recordings. These include line interferences, EOG (electro-oculogram) recording, ECG (electrocardiogram) recordings, etc. Artifacts, which are a kind of interference, increase the difficulty in the analysis of the EEG signals and the extraction of clinical information, and thereby deteriorate the sleep classification performance. In this thesis, we have used a popular but rudimentary method to remove the artifacts embedded in the EEG signals. Even though this may be outside the scope of the thesis, to improve on the performance of any automatic EEG signal classification, a more sophisticated signal processing algorithm has

to be developed so that the artifacts are automatically recognized and removed.

E2. Estimation of the PSD matrix: In this work, we assume that the EEG signals are samples of a wide sense stationary (WSS) process and estimate the power spectral density functions with the use of a multi-channel auto-regression (AR) model. Although this is an acceptable way to characterize observed time series, the estimation accuracy of power spectral density functions depends on the selected model. Therefore, even though this may lie outside the scope of the present thesis, it is desirable to further explore the estimation of power spectral density functions of multi-channel EEG signals so that the EEG samples can be more accurately characterized by their power spectral density functions.

E3. Data collection for reference library: Due to the limitation of man-power,[1] the EEG sleep signals collected and the number of patients from whom the signals are collected are quite limited. The shortage of "clinical experts" to judge on the sleep state of the signals also raises doubt on the truth of the sleep state represented by the signal. The shortage of collected signals results in only a limited range of tests that can be carried out on the effectiveness of the k-nearest neighbor decision algorithm. (k is only limited to 3 and 5 in our tests). For a more thorough evaluation on the performance using the new metrics and on the effectiveness of the decision algorithm(s), a greater amount of reliable data have to be collected and tested.

6.2.2 Ideas for future research

Apart from improvements that can be done on the present research results, there are issues raised during the course of the research which are worth pursuing. These may

[1] XLTEK, the company which agreed to supply us with test data has changed ownership in 2007 and their division designing machines for sleep tests has ceased to exist.

be the foci for future research:

F1. The mapping π: In Chapter 3, we develop the two Riemannian distances d_{R_1} and d_{R_2} from two different mappings of π resulting in two different representation points $\tilde{\mathbf{P}}$ of the PSD matrix $\mathbf{P}$. The third Riemannian distance d_{R_3} is arrived at from yet another mapping π. While the performance of the classification algorithm using d_{R_1} and d_{R_2} are generally similar, the use of d_{R_3} yields inferior results. We can immediately raise the following question: "How does the choice of the mapping π affect the performance and what π will yield the best results?"

F2. Refining the weighting matrices:

(i) In the thesis, we have derived the optimum weighting matrix by considering the different classes of reference signals altogether and have applied the same weighting matrix to all the different classes. Suppose we derive a weighting matrix for each of the different classes of signals, would the classification results be improved?

(ii) In Chapter 4, we have formulated the problem of optimizing the weighting matrix in terms of a general $(K-\kappa)$th trace of a matrix, and we derive the optimum matrices for the Riemaannian distances using only the simplest trace, i.e., $K-\kappa=1$. Is there any advantage if we derive the weighting matrices using a different value of $(K-\kappa)$? How would the performance of the corresponding classification algorithm be affected by a different choice of $(K-\kappa)$?

F3. Application of different classifiers: In this thesis, we have employed the k-nearest neighbor algorithm for classification of the EEG signals. There are other widely

used classifiers such as neural network (NN), support vector machine (SVM), Gaussian mixture model, etc. However, the practice thus far, employs these classification algorithms based on Euclidean type of distance measures. With the fundamental idea of thinking feature space as manifold rather than Euclidean space, it deserves to explore the applications of using geodesic distances to these classifiers.

F4. Quantifying the performance of EEG classification: In this thesis, we evaluate the performance of the classification by experimentally testing the algorithms on our clinically collected EEG data. A more challenging task is: "Can we derive a theoretic evaluation of the performance by deriving an expression of the probability of error of the algorithm?"

F5. Application to other signal classification problems: The thesis opens up a new approach to the signal classification problem. It has shown that exploration of the geometry of the features space may lead to more reliable classification results. Many engineering problems involves signal classification similar to the one tackled in this thesis such as the testing of EEG signals for epilepsy, for brain damages, classification of ECG signals, or even signal classifications in radar and sonar systems. We can apply this concept to other signal classification problems each possibly having a different feature space.

6.3 Conclusion

In this thesis, we have proposed a new approach of exploiting the Riemannian geometry in EEG signals and applying it to the determination of sleep state for a patient.

The results obtained are very encouraging. This shows that such an approach deserves further exploration. In the previous section, we suggested a few ideas which have arisen during the course of research. These are by no means exhaustive. Until these areas are fully explored, the research on signal classification using Riemannian geometry is far from complete, by which time, other ideas will certainly arise, and the frontier of research on the subject will be pushed still further and our knowledge in this area will yet be more sophisticated and refined.

Appendix A

The Nuttall-Strand Algorithm

To describe the algorithm, we consider the forward and backward filters which are multichannel AR models of order q, i.e.,

$$\mathbf{e}_q(t) = \mathbf{s}(t) + \sum_{k=1}^{q} \mathbf{A}(k)\mathbf{s}(t-k) \tag{A.1}$$

and

$$\mathbf{b}_q(t) = \mathbf{s}(t) + \sum_{k=1}^{q} \mathbf{B}(k)\mathbf{s}(t+k), \tag{A.2}$$

respectively. The optimum forward and backward filters can be obtained by minimizing the expected mean-square values of $\mathbf{e}_q(t)$ and $\mathbf{b}_q(t)$. The minimum of $\mathbb{E}[\mathbf{e}_q^T(t)\mathbf{e}_q(t)]$ leads to the equation:

$$\mathbf{R}^{fw}\mathbf{F}(q) = \mathbf{V}^{fw} \tag{A.3}$$

where

$$\mathbf{F}(q) = [\mathbf{I}, \mathbf{A}^T(1), \cdots, \mathbf{A}^T(q)]^T, \tag{A.4}$$

$\mathbf{R}^{fw} = [\mathbf{R}_{ik}^{fw}]$, where $\mathbf{R}_{ik}^{fw} = \mathbf{R}_{k-i}^{fw}$, $i, k = 1, 2, \cdots, q$, and $\mathbf{V}^{fw} = [\mathbf{P}^{fw}, \mathbf{0}, \cdots, \mathbf{0}]^T$ with $\mathbf{P}^{fw} = \mathbb{E}[\mathbf{e}_q(t)\mathbf{e}_q^T(t)]$. Similarly, the minimum of $\mathbb{E}[\mathbf{b}_q^T(t)\mathbf{b}_q(t)]$ leads to the equation:

$$\mathbf{R}^{bw}\mathbf{B}(q) = \mathbf{V}^{bw}, \tag{A.5}$$

where

$$\mathbf{B}(q) = [\mathbf{I}, \mathbf{B}^T(1), \cdots, \mathbf{B}^T(q)]^T, \tag{A.6}$$

$\mathbf{R}^{bw} = [\mathbf{R}^{bw}_{ik}]$, where $\mathbf{R}^{bw}_{ik} = \mathbf{R}^{bw}_{k-i}$, $i, k = 1, 2, \cdots, q$, and $\mathbf{V}^{bw} = [\mathbf{P}^{bw}, \mathbf{0}, \cdots, \mathbf{0}]^T$ with $\mathbf{P}^{bw} = \mathbb{E}[\mathbf{b}_q(t)\mathbf{b}_q^T(t)]$. To solve the Equation (A.3) and (A.5), the forward and backward filters may be postulated as

$$\mathbf{F}(q) = \begin{bmatrix} \mathbf{F}(q-1) \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{B}^{bw}(q-1) \end{bmatrix} \mathbf{C}^{fw}(q) \tag{A.7}$$

and

$$\mathbf{B}(q) = \begin{bmatrix} \mathbf{F}(q-1) \\ \mathbf{0} \end{bmatrix} \mathbf{C}^{bw}(q) + \begin{bmatrix} \mathbf{0} \\ \mathbf{B}^{bw}(q-1) \end{bmatrix}, \tag{A.8}$$

where $\mathbf{B}^{bw}(q-1) = [\mathbf{B}^T(q-1), \cdots, \mathbf{B}^T(1), \mathbf{I}]^T$.

Let $\{\mathbf{s}_t : t = 1, \cdots, T\}$ be a sample of T consecutive observations of the EEG signal. Let

$$\mathbf{s}_k(q) = [\mathbf{s}^T_{k+q}, s^T_{k+q-1}, \cdots, \mathbf{s}^T_k]^T \tag{A.9}$$

for $k = 1, 2, \cdots, N-q$, $q = 0, 1, \cdots, T-1$. Let

$$\mathbf{e}_k(q) = [(\mathbf{F}(q-1))^T, \mathbf{0}]\mathbf{s}_k(q) \tag{A.10}$$

and

$$\mathbf{b}_k(q) = [\mathbf{0}, (\mathbf{B}^{bw}(q-1))^T]\mathbf{s}_k(q). \tag{A.11}$$

Then, the algorithm is as follows:

Algorithm A.1 *(Nuttall-Strand)*

(1) Initialize the residual power matrices $\mathbf{P}^{fw}(0)$ *and* $\mathbf{P}^{bw}(0)$*:*

$$\mathbf{P}^{fw}(0) = \mathbf{P}^{bw}(0) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{s}_t\mathbf{s}_t^T \tag{A.12}$$

(2) Calculate the forward and backward residuals for $k = 1, \cdots, T-q$*:*

- $q = 1$*:* $\mathbf{e}_k(q) = \mathbf{s}_{k+1}$, $\mathbf{b}_k(q) = \mathbf{s}_k$.
- $q > 1$*:* $\mathbf{e}_k(q) = \mathbf{e}_{k+1}(q-1) + (\mathbf{C}^{fw}(q-1))^T \mathbf{b}_{k+1}(q-1)$, $\mathbf{b}_k(q) = \mathbf{b}_k(q-1) + (\mathbf{C}^{bw}(q-1))^T \mathbf{e}_k(q-1)$.

(3) Calculate

$$\mathbf{E} = \frac{1}{T-q} \sum_{k=1}^{T-q} \mathbf{e}_k(q) \mathbf{e}_k^T(q) \tag{A.13}$$

$$\mathbf{G} = \frac{1}{T-q} \sum_{k=1}^{T-q} \mathbf{b}_k(q) \mathbf{e}_k^T(q) \tag{A.14}$$

$$\mathbf{B} = \frac{1}{T-q} \sum_{k=1}^{T-q} \mathbf{b}_k(q) \mathbf{b}_k^T(q). \tag{A.15}$$

(4) Solve $\mathbf{C}^{fw}(q)$ *from*

$$\mathbf{B}\mathbf{C}^{fw}(q) + \mathbf{P}^{bw}(q-1)\mathbf{C}^{fw}(q)(\mathbf{P}^{fw}(q-1))^{-1}\mathbf{E} = -2\mathbf{G}. \tag{A.16}$$

(5) Compute $\mathbf{C}^{bw}(q)$ *by*

$$\mathbf{C}^{bw}(q) = (\mathbf{P}^{fw}(q-1))^{-1}\mathbf{C}^T(q)\mathbf{P}^{fw}(q-1). \tag{A.17}$$

(6) Compute power matrices $\mathbf{P}^{fw}(q)$ *and* $\mathbf{P}^{bw}(q)$ *by*

$$\mathbf{P}^{fw}(q) = \mathbf{P}^{fw}(q-1) - (\mathbf{C}^{fw}(q))^T \mathbf{P}^{bw}(q-1)\mathbf{C}^{fw}(1). \tag{A.18}$$

and

$$\mathbf{P}^{bw}(q) = \mathbf{P}^{bw}(q-1) - (\mathbf{C}^{bw}(q))^T \mathbf{P}^{fw}(q-1)\mathbf{C}^{bw}(q). \tag{A.19}$$

(7) Update the filters coefficients using Equation (A.7) and (A.8).

(8) If $||\mathbf{P}^{fw}(q) - \mathbf{P}^{bw}(q)|| < \epsilon$*, then go to (2).*

(9) Calculate the power spectral density matrix

$$\mathbf{P}(\omega) = \mathbf{A}^{-1}(-\omega)\mathbf{P}^{fw}(q)\mathbf{A}^{-T}(\omega), \tag{A.20}$$

where $\mathbf{A}(\omega) = \mathbf{I} + \mathbf{A}(1)e^{-j\omega} + \cdots + \mathbf{A}(q)e^{-j\omega q}$.

The algorithms only involve manipulations of $M \times M$ rather than $MT \times MT$ matrices. For the detailed derivation of the algorithms and the implementation, see [68] [79].

Appendix B

Mathematical Background

In this appendix, we briefly introduce some related mathematical concepts which may be helpful for the understanding of the development of geodesic distances in this thesis. For more details, please refer to [40] [53].

B.1 Notations

For a manifold $\mathcal{M}$, its tangent space at $p \in \mathcal{M}$ is usually denoted as $T_p\mathcal{M}$. The coordinates at $p \in \mathcal{M}$ is denoted as $\mathbf{x}(p) = (x^1, x^2, \cdots, x^n)$. The Riemannian metric is denoted as g, which some authors refer to the line element ds^2. In this work, we use notations $\mathcal{T}_{\mathcal{M}}(p)$ and $\mathbf{x}(p) = (x_1, x_2, \cdots, x_n)$ to denote tangent space and coordinates, respectively. We use g to denote the inner product function defined on the tangent space as the Riemannian metric and ds^2 as the line element. The dual basis of $(x_1, x_2, \cdots, x_n)$ is denoted as $(dx_1, dx_2, \cdots, dx_n)$.

The exponential map in Lie group theory is denoted by exp. We adopt e as the exponential map.

B.2 Riemannian geometry - Riemannian distance

A *space* is a set of points that satisfy a set of postulates. For example, the Euclidean space $\mathbb{E}^n$ is the set of n-tuples, in which a notion of distance and angle is defined and it has no origin or special choice of coordinates. After imposing a coordinate system on $\mathbb{E}^n$ we identify it with $\mathbb{R}^n$, the vector space of n-tuples of numbers. Our intuitive understanding of space is the 3-dimensional Euclidean space $\mathbb{R}^3$. In $\mathbb{R}^3$, the *distance* between two points is defined as the length of the straight line connecting them.

Since a space may not always be Euclidean in the sense that the distance could be more reasonably defined rather than the length of the straight line connecting two points (imaging the surface of a sphere), the concept of *manifold*, needs to be introduced to study those curved spaces. A *manifold* is an abstract mathematical space in which every point has a neighborhood which resembles Euclidean space, but in which the global structure may be more complicated. In other words, manifolds allow more complicated structures can be expressed and understood in terms of the relatively well-understood properties of simpler spaces. Defining additional structures on manifolds can lead to different kind of manifolds such as topological manifolds, differentiable manifolds, Riemannian manifolds, symplectic manifolds, pseudo-Riemannian manifolds to name a few. In this work, we are interested in the distance measure on manifolds. Since we mainly consider the geometry of the real manifold of positive-definite Hermitian matrices, we focus on the Riemannian geometry, which is the study of differentiable manifold by endowing the manifold a Riemannian metric.

The key idea of investigation of surfaces presented by Gauss is that a point on a surface in ordinary Euclidean space is determined by two coordinates x_1 and x_2, and the arc element is expressed in terms of a given positive definite quadratic form in the differentials of these coordinates, i.e., $ds^2 = g_{11}(dx_1)^2 + 2g_{12}dx_1dx_2 + g_{22}(dx_2)^2$,

where g_{11}, g_{12}, and g_{22} are functions of the variables x_1 and x_2. In his famous lecture at Göttingen in 1854 with the title "On the hypotheses which underlie geometry", Riemann extended Gauss's idea and developed a metric geometry of a manifold of n dimensions, that is, a set of elements each of which is determined by n bits of numerical data, its coordinates $x_1, x_2, \cdots, x_n$. Riemann's idea is linked to the mode of determination of the distance between two infinitely close elements (the arc element) given by

$$ds = F(x_1, \cdots, x_n, dx_1, \cdots, dx_n) \tag{B.1}$$

where the function $F(\mathbf{x}, \mathbf{y})$ is linearly homogeneous in $\mathbf{y}$, i.e., $F(\mathbf{x}, \alpha\mathbf{y}) = \alpha F(\mathbf{x}, \mathbf{y})$, where α is a constant. An important special case is when

$$ds^2 = F^2(\mathbf{x}, d\mathbf{x}) = \sum_{i,j=1}^{n} g_{ij}(\mathbf{x}) dx_i dx_j \tag{B.2}$$

It is important to note that this is not just an extension of Gauss's formula to an n-dimensional manifold. Rather, it introduces the completely new idea of determining the metric on a manifold by specifying it in an infinitely small potion of that manifold.

B.2.1 Differentiable manifold

Roughly speaking, a differentiable manifold is a topological space with a differentiable structure.

1. **Topological space:** A topological space is a set $\mathcal{M}$ together with a collection of subsets of $\mathcal{M}$, $\mathcal{T}$, satisfying
   - the empty set $\emptyset$ and $\mathcal{M}$ are in $\mathcal{T}$.
   - the union of any collection of sets in $\mathcal{T}$ is also in $\mathcal{T}$.
   - the intersection of any finite collection of sets in $\mathcal{T}$ is also in $\mathcal{T}$.

The collection $\mathcal{T}$ is called a topology on $\mathcal{M}$.

2. **Hausdorff space:** A space $\mathcal{M}$ is said to be a Hausdorff space if for any $x, y \in mathcalM$ with $x \neq y$, then there exist neighborhoods U and V of x and y respectively such that $U \cap V = \emptyset$, i.e., U and V are disjoint.

3. **Continuous, Homeomorphism, and Diffeomorphism:** A map $f : \mathcal{N} \rightarrow \mathcal{M}$ between two topological spaces is said to be continuous if $f^{-1}(U)$ is open for any open subset U of $\mathcal{M}$. f is called a homeomorphism if it is a bijection and both f and f^{-1} are continuous. f is called a diffeomorphism if, in addition, f and f^{-1} are differentiable.

4. **Dimension of a manifold:** Let $\mathcal{M}$ be a Hausdorff space. If for any $\mathbf{p} \in \mathcal{M}$, there exists a neighborhood U of $\mathbf{p}$ such that U is homeomorphic to an open set in $\mathbb{R}^n$, then $\mathcal{M}$ is called an n-dimensional topological manifold.

5. **Coordinate charts and Atlas:** Let $\mathcal{I}$ be an index set, whose elements is used to index homeomorphisms. Let $U \subset \mathcal{M}$. If $\mathbf{x}_\alpha : U \rightarrow \mathbf{x}_\alpha(U)$, where $\alpha \in \mathcal{I}$ and $\mathbf{x}_\alpha(U)$ is an open set in $\mathbb{R}^n$, is a homeomorphism, then $(U, \mathbf{x}_\alpha)$ is a coordinate chart of $\mathcal{M}$. The coordinates of $\mathbf{p} \in U$ is the coordinates of $\mathbf{x} = \mathbf{x}_\alpha(\mathbf{p}) \in \mathbb{R}^n$, i.e.,

$$x = (x_1, \cdots, x_n) = ((\mathbf{x}_\alpha(\mathbf{p}))_1, \cdots, (\mathbf{x}_\alpha(\mathbf{p}))_n) \tag{B.3}$$

which called the local coordinates of the point $\mathbf{p} \in U$. The collection of all the coordinate charts forms an atlas $\mathcal{A}$, i.e., $\mathcal{A} = \{(U_\alpha, \mathbf{x}_\alpha) | \alpha \in \mathcal{I}\}$.

6. **Differential structure:** Suppose $(U, \mathbf{x}_\alpha)$ and $(V, \mathbf{x}_\beta)$ are two coordinate charts of $\mathcal{M}$. If $U \cap V \neq \emptyset$, then $\mathbf{x}_\alpha(U \cap V)$ and $\mathbf{x}_\beta(U \cap V)$ are two nonempty open sets in $\mathbb{R}^n$, and the map

$$\mathbf{x}_\beta \circ \mathbf{x}_\alpha^{-1}|_{\mathbf{x}_\alpha(U \cap V)} : \mathbf{x}_\alpha(U \cap V) \rightarrow \mathbf{x}_\beta(U \cap V) \tag{B.4}$$

defines a homeomorphism between these two open sets, with inverse given by

$$\mathbf{x}_\alpha \circ \mathbf{x}_\beta^{-1}|_{\mathbf{x}_\beta(U\cap V)} : \mathbf{x}_\beta(U \cap V) \to \mathbf{x}_\alpha(U \cap V) \tag{B.5}$$

These are both maps between open sets in a Euclidean space. Expressed in coordinates, $\mathbf{x}_\alpha \circ \mathbf{x}_\beta^{-1}$ and $\mathbf{x}_\beta \circ \mathbf{x}_\alpha^{-1}$ each represents n-real valued functions on an open set of a Euclidean space (see Figure B.1).

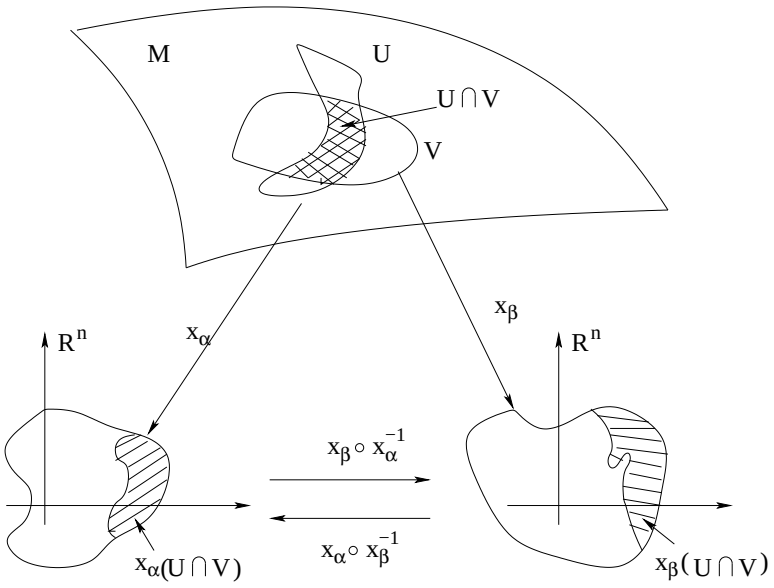


Figure B.1: Differentiable manifold

The manifold $\mathcal{M}$ is called a differentiable manifold if for all $\alpha, \beta \in \mathcal{I}$ the corresponding transition maps

$$\mathbf{x}_\beta \circ \mathbf{x}_\alpha^{-1}|_{\mathbf{x}_\alpha(U_\alpha\cap U_\beta)} : \mathbf{x}_\alpha(U_\alpha \cap U_\beta) \to \mathbb{R}^n \tag{B.6}$$

are differentiable.

B.2.2 Riemannian manifold

A Riamannian manifold $(\mathcal{M}, g)$ is a differential manifold $\mathcal{M}$ in which each tangent space is equipped with an inner product g in a manner which varies smoothly from point to point. In other worlds, a Riemannian manifold is a differentiable manifold in which the tangent space at each point is a finite-dimensional Hilbert space (inner product space). The dimension of the manifold is the dimension of the tangent space at each point on the manifold. A Riemannian metric on $\mathcal{M}$ allows one to measure lengths of smooth paths in $\mathcal{M}$ and hence to define a distance function by taking the infimum of the lengths of smooth paths between two points. This makes $\mathcal{M}$ a metric space.

1. **Vector field and Tangent space:** A vector field $\mathbf{V}$ on a given manifold $\mathcal{M}$ is an assignment to every point $\mathbf{p} \in \mathcal{M}$ a tangent vector to $\mathcal{M}$ at $\mathbf{p}$. That is, for each $\mathbf{p} \in \mathcal{M}$, we have a tangent vector $\mathbf{v} = \mathbf{V}(p)$ in the space $\mathcal{T}_{\mathcal{M}}(\mathbf{p})$, which is the tangent space of $\mathcal{M}$ at $\mathbf{p}$. Figure B.2 shows the tangent space of $\mathcal{M}$ at $\mathbf{p}$. The tangent bundle $\mathcal{TM}$ is the disjoint union of the tangent spaces of the points of $\mathcal{M}$, i.e., $\mathcal{TM} = \bigcup_{\mathbf{p}\in\mathcal{M}} \mathcal{T}_{\mathcal{M}}(p)$. The collection of all the vector fields on $\mathcal{M}$ is denoted by $\Gamma(\mathcal{TM})$.

2. **Riemannian metric:** Let $\mathcal{M}$ be a differentiable manifold of dimension n. A Riemannian metric on $\mathcal{M}$ is a family of inner products

$$g : \mathcal{T}_{\mathcal{M}}(\mathbf{p}) \times \mathcal{T}_{\mathcal{M}}(\mathbf{p}) \to \mathbb{R}, \quad \mathbf{p} \in \mathcal{M} \tag{B.7}$$

such that, for all differentiable vector fields $\mathbf{X}, \mathbf{Y} \in \Gamma(\mathcal{TM})$, the application

$$\mathcal{M} \to \mathbb{R}, \quad \mathbf{p} \mapsto g(\mathbf{X}(\mathbf{p}), \mathbf{Y}(\mathbf{p})) \tag{B.8}$$

is differentiable.

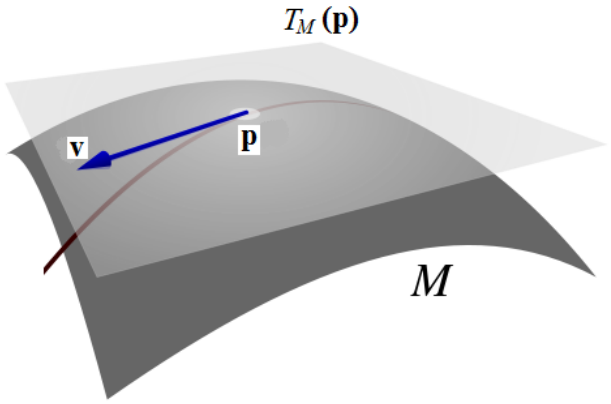


Figure B.2: Tangent space $\mathcal{T}_{\mathcal{M}}(\mathbf{p})$ and a tangent vector $\mathbf{v} \in \mathcal{T}_{\mathcal{M}}(\mathbf{p})$ along a curve through $\mathbf{p} \in \mathcal{M}$

In fact, Eq. (B.7) defines a metric tensor on $\mathcal{M}$ such that $g(\mathbf{v}, \mathbf{w})$, $\mathbf{v}, \mathbf{w} \in \mathcal{T}_{\mathcal{M}}(\mathbf{p})$, produces a real number (scalar) in a way that generalizes the inner product of vectors in Euclidean space.

Any differentiable manifold can be endowed with a Riemannian metric [53] . Therefore, there are infinitely many Riemannian metric on $\mathcal{M}$. A question raised is that given a compact differentiable manifold, does it carry a best, or a family of best, Riemannian structure? The most natural definition of a best metric is the least curved one (with the smallest curvature). However, it is generally difficult to find the Riemannian metrics best adapted to the given manifold structure.

3. **Differential maps:** Let $(\mathcal{M}, g_{\mathcal{M}})$ and $(\mathcal{N}, g_{\mathcal{N}})$ be two Riemannian manifolds. If $\phi : \mathcal{M} \to \mathcal{N}$ is a differentiable map from the manifold $\mathcal{M}$ to the manifold $\mathcal{N}$,

the induced differential map $\boldsymbol{\phi}_* : \mathcal{T}_{\mathcal{M}}(\mathbf{p}) \to \mathcal{T}_{\mathcal{N}}(\phi(\mathbf{p}))$ is called a push-forward map if

$$\phi_*(\mathbf{v})(f) = \mathbf{v}(f \circ \phi), \tag{B.9}$$

for $\mathbf{v} \in \mathcal{T}_{\mathcal{M}}(\mathbf{p})$ and $f \in C^\infty(N, \mathbb{R})$, where $C^\infty(\mathcal{N}, \mathbb{R})$ denotes the class of smooth functions from $\mathcal{N}$ to $\mathbb{R}$.

Roughly speaking, the push-forward transforms the velocity vectors of a curve $\boldsymbol{\gamma} : [0, 1] \to \mathcal{M}$ to the velocity vectors of the transformed curve $\boldsymbol{\phi}(\boldsymbol{\gamma})$ in $\mathcal{N}$.

A metric $\phi^* g_{\mathcal{N}}$ on $\mathcal{M}$ is called the pull-back metric if

$$(\phi^* g_{\mathcal{N}})(\mathbf{x}, \mathbf{y}) = g_{\mathcal{N}}(\phi_* \mathbf{x}, \phi_* \mathbf{y}), \tag{B.10}$$

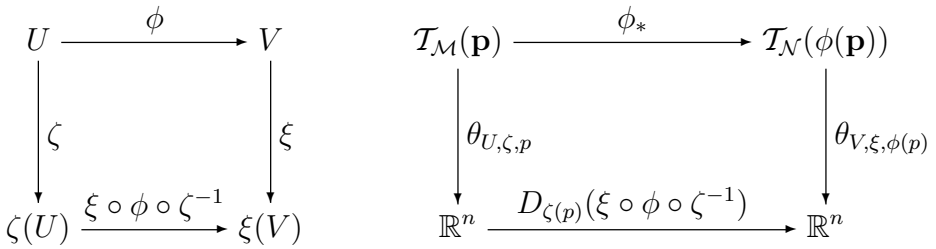
where $\mathbf{x}, \mathbf{y} \in \mathcal{T}_{\mathcal{M}}(\mathbf{p})$.

The map $\phi : \mathcal{M} \to \mathcal{N}$ is said to be an isometry if the following holds

$$g_{\mathcal{M}}(\mathbf{x}, \mathbf{y}) = g_{\mathcal{N}}(\phi_*(\mathbf{x}), \phi_*(\mathbf{y})), \tag{B.11}$$

for all $\mathbf{x}, \mathbf{y} \in T\mathcal{M}$.

Let (U, ζ) and (V, ξ) be charts for $\mathcal{M}$ and $\mathcal{N}$ around p and $\phi(p)$. The meaning of the map between manifolds and the induced differential map can be seen from the following diagrams [40].



Roughly speaking, by introducing local coordinates the operations on manifolds can be done equivalently on the Euclidean spaces.

The map $\phi : \mathcal{M} \to \mathcal{N}$ is called a submersion at $\mathbf{p} \in \mathcal{M}$ if the induced tangential map $\phi_* : \mathcal{T}_{\mathbf{p}}(\mathcal{M}) \to \mathcal{T}_{\mathcal{N}}(\phi(\mathbf{p}))$ is a surjective linear map.

Let $\phi : \mathcal{M} \to \mathcal{N}$ be a submersion. Let $\mathcal{T}_{\mathbf{p}}(\mathcal{M}) = \mathcal{V}_{\mathbf{p}} \oplus \mathcal{U}_{\mathbf{p}}$, where $\mathcal{V}_{\mathbf{p}}$ and $\mathcal{U}_{\mathbf{p}}$ are the vertical and horizontal subspaces of $\mathcal{T}_{\mathbf{p}}(\mathcal{M})$ at $\mathbf{p} \in \mathcal{M}$ respectively. Then, the map $\boldsymbol{\phi}$ from $\mathcal{M}$ to $\mathcal{N}$ is a Riemannian submersion if

- ϕ is a smooth submersion
- For any $\mathbf{p} \in \mathcal{M}$, $\mathcal{U}_{\mathbf{p}}$ and $\mathcal{T}_{\phi(\mathbf{p})}(\mathcal{N})$ are isometric

4. **Line element:** Let $(U, (x_1, \cdots, x_n))$ be a chart of the manifold around $\mathbf{p} \in \mathcal{M}$. Then the coordinate vector fields $\left(\frac{\partial}{\partial x_1}, \cdots, \frac{\partial}{\partial x_n}\right)$ form a basis for the tangent space $\mathcal{T}_{\mathcal{M}}(\mathbf{p})$. Let $(dx_1, \cdots, dx_n)$ be the basis of the dual space of $\mathcal{T}_{\mathcal{M}}(\mathbf{p})$, i.e., the cotangent space $\mathcal{T}^*_{\mathcal{M}}(\mathbf{p})$. Then we have

$$dx_j\Big(\frac{\partial}{\partial x_i}\Big) = \delta_i^j = \begin{cases} 1 & i = j \\ 0 & j \neq j \end{cases} \tag{B.12}$$

Let

$$g_{ij} := g\left(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j}\right) = \Big\langle \frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j} \Big\rangle \tag{B.13}$$

where $\langle \cdot, \cdot \rangle$ denotes inner product. Let $\boldsymbol{\gamma}(t) : [a, b] \to \mathcal{M}$ be a parameterized curve on $\mathcal{M}$, then we have $\dot{\boldsymbol{\gamma}}(t) \in \mathcal{T}_{\mathcal{M}}(\boldsymbol{\gamma}(t))$. Suppose $\mathbf{p} = \boldsymbol{\gamma}(c)$, $a \leq c \leq b$, and let $\mathbf{v} = \dot{\boldsymbol{\gamma}}(c)$, then we have

$$\mathbf{v} = \sum_i \alpha_i \frac{\partial}{\partial x_i} = \sum_i dx_i(\mathbf{v}) \frac{\partial}{\partial x_i} \tag{B.14}$$

by Eq. (B.13). Thus,

$$ds^2 = g(\mathbf{v}, \mathbf{v}) = \langle \mathbf{v}, \mathbf{v} \rangle = g\left(\sum_i dx_i(\mathbf{v}) \frac{\partial}{\partial x_i}, \sum_j dx_j(\mathbf{v}) \frac{\partial}{\partial x_j} \right) = \sum_{i,j} g_{ij} dx_i \otimes dx_j \tag{B.15}$$

where $\otimes$ denotes tensor product. In other worlds, the metric tensor g_{ij} defines the differential metrical distance along any smooth curve in terms of $(dx_1, \cdots, dx_n)$ according to

$$ds^2 := \sum_{i,j} g_{ij} dx_i \otimes dx_j, \tag{B.16}$$

Let

$$dx_i dx_j = \frac{1}{2}(dx_i \otimes dx_j + dx_j \otimes dx_i), \tag{B.17}$$

then, Equation (B.16) can be shorten as

$$ds^2 := \sum_{i,j} g_{ij} dx_i dx_j. \tag{B.18}$$

Eq. (B.18) is called the first fundamental form or element of arc length.

5. **Distance:** A connected Riemannian manifold carries the structure of a metric space. Let $\boldsymbol{\gamma} : [0, 1] \to \mathcal{M}$ be a parameterized curve in $\mathcal{M}$, which is differentiable with velocity vector $\dot{\boldsymbol{\gamma}} = \frac{\gamma}{dt}$. The length of the curve $\boldsymbol{\gamma}$ is defined as

$$\ell(\boldsymbol{\gamma}) = \int_{[0,1]} \sqrt{g(\dot{\boldsymbol{\gamma}}(t), \dot{\boldsymbol{\gamma}}(t))} dt. \tag{B.19}$$

The intrinsic distance $d : \mathcal{M} \times \mathcal{M} \to [0, \infty)$ is defined by

$$d(p, q) = \inf_{\gamma} \ell(\boldsymbol{\gamma}), \tag{B.20}$$

where $\boldsymbol{\gamma}$ runs over all differentiable curves connecting $p \in \mathcal{M}$ and $q \in \mathcal{M}$.

6. **Geodesics:**

   (1) **Connection:** Let $\boldsymbol{\gamma} : [0, 1] \to \mathcal{M}$ be a smooth curve. A smooth vector field along $\boldsymbol{\gamma}$ is a family $\{\mathbf{v}_t, t \in [0, 1]\}$ of tangent vectors $\mathbf{v}_t \in \mathcal{T}_{\mathcal{M}}(\boldsymbol{\gamma}(t))$ such that if $(U, (x_1, \cdots, x_n))$ is a chart near $\boldsymbol{\gamma}(t_0)$ and $\mathbf{v}_t = \sum_{i=1}^{n} v_i(t) \frac{\partial}{\partial x_i}|_{\boldsymbol{\gamma}(t)}$, for t in an interval around t_0, then $v_i(t)$ are smooth functions.

A linear connection in $\mathcal{M}$ is an operator ∇, which defines a vector field $\nabla_{\mathbf{X}}\mathbf{Y}$ for any two smooth vector fields $\mathbf{X}, \mathbf{Y}$ on $\mathcal{M}$ such that

* $\nabla_{\mathbf{X}}\mathbf{Y}$ is smooth.
* $\nabla_{\alpha\mathbf{X}_1+\beta\mathbf{X}_2}\mathbf{Y} = \alpha\nabla_{\mathbf{X}_1}\mathbf{Y} + \beta\nabla_{\mathbf{X}_2}\mathbf{Y}$, $\alpha, \beta \in \mathbb{R}$).
* $\nabla_{\mathbf{X}}(\mathbf{Y}_1 + \mathbf{Y}_2) = \nabla_{\mathbf{X}}\mathbf{Y}_1 + \nabla_{\mathbf{X}}\mathbf{Y}_2$.
* $\nabla_{\mathbf{X}}(f\mathbf{Y}) = f\nabla_{\mathbf{X}}\mathbf{Y} + \mathbf{X}(f)\cdot\mathbf{Y}$, $f \in C^{\infty}(\mathcal{M}, \mathbb{R})$, where $\mathbf{X}(f)$ denotes the directional directive of α in the direction of $\mathbf{X}$.

(2) **Covariant derivation:** The operator defined in the set of vector fields along a curve $\boldsymbol{\gamma}$

$$\frac{D\mathbf{V}}{dt} = \dot{\mathbf{V}} = \nabla_{\frac{\gamma}{dt}}(\mathbf{V}) \tag{B.21}$$

is called covariant derivation along $\boldsymbol{\gamma}$.

(3) **Parallel transportation:** A vector field $\mathbf{V}$ along $\boldsymbol{\gamma}$ is called parallel if $\frac{D\mathbf{V}}{dt} = 0$.

A connection provides a way of "connecting" the tangent space at one point by the tangent space at another point on a given manifold $\mathcal{M}$. Given a smooth curve $\boldsymbol{\gamma} : [0, 1] \to \mathcal{M}$, $\boldsymbol{\gamma}(0) = \mathbf{p}$ and $\boldsymbol{\gamma}(1) = \mathbf{q}$, a tangent vector $\mathbf{v} \in \mathcal{T}_{\mathcal{M}}(\mathbf{p})$ can be parallel transported to a vector in $\mathcal{T}_{\mathcal{M}}(\mathbf{q})$ along $\boldsymbol{\gamma}$ via a parallel transportation.

(4) **Geodesics:** A C^2 curve $\boldsymbol{\gamma}$ in a Riemannian manifold $\mathcal{M}$ is called a geodesic if the equation

$$\nabla_{\dot{\boldsymbol{\gamma}}(t)}\dot{\boldsymbol{\gamma}}(t) = 0 \tag{B.22}$$

is satisfied. This property reflects a property of straight lines in Euclidean geometry. Let $(U, x_1, \cdots, x_n)$ be a chart of $\mathcal{M}$. Let $\boldsymbol{\gamma}(t) = (x_1(t), \cdots, x_n(t))$ be a given curve on $\mathcal{M}$. Then its tangent field $\dot{\boldsymbol{\gamma}}(t) = \sum_{i=1}^{n} \dot{x}_i(t)\partial_i$, where $\partial_i = \frac{\partial}{\partial x_i}$. Since $\nabla_{\partial_j}\partial_i = \sum_{k=1}^{n} \Gamma_{ij}^k \partial_k$, where $\Gamma_{ij}^k : \mathcal{M} \to$

$\mathbb{R}$ are called the Christoffel symbols of the connection under the local coordinates, the geodesic equation turns into

$$\begin{aligned}\nabla_{\dot{\boldsymbol{\gamma}}(t)}\dot{\boldsymbol{\gamma}}(t) &= \nabla_{\dot{\boldsymbol{\gamma}}(t)}\Big(\sum_{i=1}^{n}\dot{x}_i(t)\partial_i\Big)\\ &= \sum_{i=1}^{n}\ddot{x}_i(t)\partial_i + \sum_{i,j=1}^{n}\dot{x}_j(t)\dot{x}_i(t)\nabla_{\partial_j}\partial_i\\ &= \sum_{k=1}^{n}\Big(\ddot{x}_k(t) + \sum_{i,j=1}^{n}\Gamma_{ij}^{k}(t)\dot{x}_i(t)\dot{x}_j(t)\Big)\partial_k = 0. \end{aligned} \tag{B.23}$$

Therefore, the curve $\boldsymbol{\gamma}$ will be a geodesic if and only if

$$\ddot{x}_k(t) + \Gamma_{ij}^{k}\dot{x}_i(t)\dot{x}_j(t) = 0, \quad k = 1, \cdots, n, \tag{B.24}$$

which is equivalent to the following system of first order ordinary differential equations:

$$\begin{cases} \dot{x}_k(t) = & -y_k(t) \\ \dot{y}_k(t) = & \sum_{i,j=1}^{n}\Gamma_{ij}^{k}(t)y_i(t)y_j(t), \quad 1 \le k \le n. \end{cases} \tag{B.25}$$

By the fundamental theorem of ordinary differential equations (existence and uniqueness theorem) we have that for a given $\mathbf{p}_0 \in \mathcal{M}$ and $\mathbf{u}_0 \in \mathcal{T}_{\mathcal{M}}(\mathbf{p}_0)$, there exists an $\epsilon > 0$ and a neighborhood $O(\mathbf{u}_0)$ of $\mathbf{u}_0$ in $\mathcal{TM}$ such that for any $\mathbf{u} \in O(\mathbf{u}_0)$ we have a unique geodesic $\boldsymbol{\gamma}(t)$ defined for $|t| \le \epsilon$ (ϵ is a sufficient small positive real number) which satisfies the initial conditions $\boldsymbol{\gamma}(0) = \mathbf{p}_0$, $\dot{\boldsymbol{\gamma}}(0) = \mathbf{u}_0$.

In the presence of a metric, geodesics are defined to be locally the shortest path between points on the manifold [53]. In the presence of a connection, geodesics are defined to be curves whose tangent vectors remain parallel if they are transported along it. We are interested in the finding of explicit formula for geodesic distance between any two points in the manifold $\mathcal{M}$.

Although the geodesic equation under local coordinates is obtained, it is difficult to be solved due to the nonlinearity.

B.2.3 Lie group and Lie algebra

1. **Matrix exponential:**

$$e^{\mathbf{A}} = \mathbf{I} + \mathbf{A} + \frac{\mathbf{A}^2}{2!} + \frac{\mathbf{A}^3}{3!} + \cdots + \frac{\mathbf{A}^n}{n!} + \cdots \tag{B.26}$$

If

$$e^{\mathbf{A}} e^{\mathbf{B}} = e^{\mathbf{C}} \tag{B.27}$$

then

$$\mathbf{C} = \mathbf{A} + \mathbf{B} + \frac{1}{2}[\mathbf{A}, \mathbf{B}] + \frac{1}{12}([\mathbf{A}, [\mathbf{A}, \mathbf{B}]] + [\mathbf{B}, [\mathbf{B}, \mathbf{A}]]) + \cdots \tag{B.28}$$

where $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are $n \times n$ matrices and $[\mathbf{A}, \mathbf{B}] = \mathbf{AB} - \mathbf{BA}$ is the commutator bracket. The matrix exponential satisfies the following properties

   - $e^{\mathbf{0}} = \mathbf{I}$.
   - $e^{\mathbf{A}^H} = (e^{\mathbf{A}})^H$
   - $e^{\mathbf{GAG}^{-1}} = \mathbf{G} e^{\mathbf{A}} \mathbf{G}^{-1}$
   - $|e^{\mathbf{A}}| = e^{Tr\mathbf{A}}$
   - $(e^{\mathbf{A}})^{-1} = e^{-\mathbf{A}}$

- **Group:** A set G with the operation $\cdot$, denoted by $(G, \cdot)$, is a group if
   - Closure: $\forall a, b \in G$, $a \cdot b \in G$
   - Associativity: $\forall a, b, c \in G$, $(a \cdot b) \cdot c = a \cdot (b \cdot c)$
   - Identity element: $\exists e \in G$, $\forall a \in G$, $e \cdot a = a \cdot e = a$

- Inverse element: $\exists b \in G$, $a \cdot b = b \cdot a = e$ if $a \in G$

2. **Lie group:** A set G is a Lie group if
   - G is a differentiable manifold;
   - G is a group;
   - the map $(g, h) \mapsto gh^{-1}$ from the manifold $G \times G$ into G is differentiable.
3. **matrix group:** A matrix group is a group G consisting of invertible matrices over some field F with operations of matrix multiplication and inversion. For example, a linear algebra is a matrix group.
4. **Lie algebra:** Let F be a field (usually $\mathbb{R}$ or $\mathbb{C}$). A Lie algebra over F is an F-vector space $\mathfrak{g}$, together with a bilinear map, called the Lie bracket

$$\mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}, \quad (x, y) \mapsto [x, y] \tag{B.29}$$

which satisfies the following properties:

- alternating on $\mathfrak{g}$

$$[x, x] = 0 \ \ \forall x \in \mathfrak{g}; \tag{B.30}$$

- the Jacobi identity

$$[x, [y, z]] + [y, [z, x]] + [z, [x, y]] = 0 \ \ \forall x, y, z \in \mathfrak{g} \tag{B.31}$$

For example, in linear algebra $\mathcal{A}$, the Lie bracket is defined to be

$$[\mathbf{X}, \mathbf{Y}] = \mathbf{XY} - \mathbf{YX} \tag{B.32}$$

for two elements $\mathbf{X}, \mathbf{Y} \in \mathcal{A}$.

5. **Lie algebra of matrix group:** The Lie algebra of a matrix group G is the tangent space to G at the identity $\mathbf{I}$, i.e., $T_{\mathbf{I}}G = \{\dot{\boldsymbol{\gamma}}(0) : \boldsymbol{\gamma}(t) : (-\epsilon, \epsilon) \rightarrow G, \boldsymbol{\gamma}(0) = \mathbf{I}\}$. It is denoted by $\mathfrak{g} = T_{\mathbf{I}}G$.

6. **dimension of matrix group:** The dimension of a matrix group G means the dimension of its Lie algebra.

7. **Unitary group:** Let $\mathcal{M}_n(\mathbb{C})$ be the set of $n \times n$ complex matrices. The unitary group is defined as

$$U(n) = \{\mathbf{B} \in \mathcal{M}_n(\mathbb{C}) : \mathbf{B}^H\mathbf{B} = \mathbf{I}\} \tag{B.33}$$

8. **Lie algebra of a unitary group:** The Lie algebra $\mathfrak{u}(n)$ of $U(n)$ is skew-Hermitian, i.e., for any $\mathbf{S} \in \mathfrak{u}(n)$, $\mathbf{S}^H = -\mathbf{S}$. The dimension of $U(n)$ is n^2 [8].

B.2.4 Fiber bundles

1. **Submersion:** Let $\mathcal{M}$ and $\mathcal{N}$ be differentiable manifolds. A smooth map $\phi : \mathcal{M} \rightarrow \mathcal{N}$ is a submersion at $\mathbf{p} \in \mathcal{M}$ if its differential map $\phi_* : \mathcal{T}_{\mathcal{M}}(\mathbf{p}) \rightarrow \mathcal{T}_{\mathcal{N}}(\phi(\mathbf{p}))$ at $\mathbf{p}$ is surjective (onto).

2. **Riemannian submersion:** Let $(\mathcal{M}, g)$ and $(\mathcal{N}, h)$ be two Riemannian manifolds. A map $\phi : \mathcal{M} \rightarrow \mathcal{N}$ is a Riemannian submersion if:

   - ϕ is a smooth submersion.
   - the map $\phi_* : \mathcal{U}_{\mathcal{M}}(\mathbf{p}) \rightarrow \mathcal{T}_{\mathcal{N}}(\phi(\mathbf{p}))$ is an isometry, where $\mathcal{U}_{\mathcal{M}}(\mathbf{p}) = \mathcal{V}_{\mathcal{M}}(\mathbf{p})^{\perp}$ and $\mathcal{V}_{\mathcal{M}}(\mathbf{p}) = \ker(\phi_*)$ such that $T_{\mathcal{M}}(\mathbf{p}) = \mathcal{U}_{\mathcal{M}}(\mathbf{p}) \oplus \mathcal{V}_{\mathcal{M}}(\mathbf{p})$,

   $\mathcal{V}_{\mathcal{M}}(\mathbf{p})$ is called the vertical subspace of $\mathcal{T}_{\mathcal{M}}(\mathbf{p})$ and $\mathcal{U}_{\mathcal{M}}(\mathbf{p})$ is called the horizontal subsapce of $\mathcal{T}_{\mathcal{M}}(\mathbf{p})$. For example, let $\mathcal{M}$ and $\mathcal{N}$ be two Riemannian

manifolds. Then the projection map $pr_1 : \mathcal{M} \times \mathcal{N} \to \mathcal{M}$ is a Riemann submersion.

3. **Fiber bundle:** Let $\mathcal{E}, \mathcal{B}, \mathcal{F}$ be smooth manifolds and $\pi : \mathcal{E} \to \mathcal{B}$ be a smooth map. Let J be an index set. The triple $(\pi, \mathcal{E}, \mathcal{B})$ is a fiber bundle with fiber $\mathcal{F}$, basis $\mathcal{B}$, and total space $\mathcal{E}$ if:

   - the map π is surjective submersion.
   - there exists an open cover (O_j) $(j \in J)$ of $\mathcal{B}$ (i.e., $\mathcal{B} \subseteq \cup_{j \in J} O_j$), and diffeomorphisms
   $$h_j : \pi^{-1}(O_j) \to O_j \times \mathcal{F} \tag{B.34}$$
   such that $h_j(\pi^{-1}(x)) = \{x\} \times \mathcal{F}$ for $x \in O_j$ (this is called local triviality of the bundle).

4. **Principal G-bundle:** Let $\mathcal{E}$ and $\mathcal{B}$ be smooth manifolds and G be a Lie group acting on $\mathcal{E}$ such that $(\tilde{\mathbf{p}}, g) \in \mathcal{E} \times G$ is mapped to $\tilde{\mathbf{p}}g \in \mathcal{E}$, and $\tilde{\mathbf{p}}g \neq \tilde{\mathbf{p}}$ for $g \neq e$ (e is the identity element of G). Let $\pi : \mathcal{E} \to \mathcal{B}$ be a smooth surjective submersion such that the set $\{\tilde{\mathbf{p}}g : g \in G\}$ coincide with the fibers, i.e.,
   $$\{\tilde{p}g : g \in G\} = \pi^{-1}(\pi(\tilde{\mathbf{p}})) \quad \forall \tilde{\mathbf{p}} \in \mathcal{E} \tag{B.35}$$
   Then $\pi : \mathcal{E} \to \mathcal{B}$ is said to be a principal G-bundle.

   Remarks:

   - for $\mathbf{p} \in \mathcal{B}$, $\pi^{-1}(\mathbf{p})$ has a Lie group structure but not canonical since there is no preferred choice of an identity element;
   - Around each $\mathbf{p} \in \mathcal{B}$ there exists an open neighborhood O such that $\pi^{-1}(O)$ and $O \times G$ are diffeomorphic. This is called a trivialization;
   - The vertical subbundle of $\mathcal{E}$ has null projection to $T\mathcal{B}$.

The principal G-bundle is illustrated in Figure B.3.

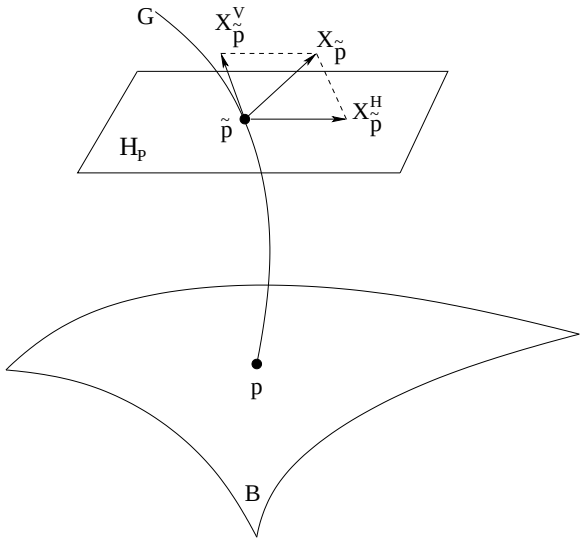


Figure B.3: Principal G-bundle

5. **Connections in principal G-bundles:** Let $\pi : \mathcal{E} \to \mathcal{B}$ be a principal G-bundle and let $\mathfrak{g}$ be the Lie algebra of G. Let $R_g : \mathcal{E} \to \mathcal{E}$ is the right multiplication defined by g, i.e., $\mathbf{p} \mapsto \mathbf{p} \cdot g$. A subset $\mathcal{H} \subset \mathcal{TE}$ in $\mathcal{E}$ is called a connection if the following holds

   - The induced differential map satisfies $(R_g)_* \mathcal{H}_{\mathbf{p}} = \mathcal{H}_{\mathbf{p} \cdot g}$ for every $\mathbf{p} \in \mathcal{E}$ and $g \in G$ (this is called G-invariant).
   - For every $\mathbf{p} \in \mathcal{E}$, $\mathcal{H}_{\mathbf{p}} \oplus \mathcal{V}_{\mathbf{p}} = \mathcal{T}_{\mathcal{E}}(\mathbf{p})$ (This is called direct decomposition).

6. **Horizontal lift:** A connection prescribes a manner for lifting curves from the base manifold $\mathcal{B}$ into the total space of $\mathcal{E}$ so that the tangents to the curve are horizontal. Specifically, suppose that $\boldsymbol{\gamma}(t) : [0, 1] \to \mathcal{B}$ is a smooth curve in $\mathcal{B}$ through the point $\mathbf{p} = \boldsymbol{\gamma}(0)$. Let $\tilde{\mathbf{p}} \in \mathcal{E}_{\mathbf{p}}$ be a point in the fibre over $\mathbf{p}$. A lift of

$\boldsymbol{\gamma}$ through $\tilde{\mathbf{p}}$ is a curve $\tilde{\boldsymbol{\gamma}}(t)$ in the total space $\mathcal{E}$ such that

$$\tilde{\boldsymbol{\gamma}}(0) = \tilde{\mathbf{p}} \tag{B.36}$$

and

$$\pi(\tilde{\boldsymbol{\gamma}}(t)) = \boldsymbol{\gamma}(t), \quad t \in [0,1] \tag{B.37}$$

A lift is horizontal if, in addition, every tangent of the curve lies in the horizontal subbundle of $\mathcal{TE}$, i.e.,

$$\dot{\tilde{\boldsymbol{\gamma}}}(t) \in \mathcal{H}_{\tilde{\boldsymbol{\gamma}}(t)}, \quad t \in [0,1]. \tag{B.38}$$

Let $\pi : \mathcal{E} \to \mathcal{M}$ is a Riemannian submersion. If $\tilde{\boldsymbol{\gamma}}$ is a geodesic of $\mathcal{E}$ such that $\dot{\tilde{\boldsymbol{\gamma}}}(0)$ is horizontal, then $\dot{\tilde{\boldsymbol{\gamma}}}(t)$ is horizontal for all t, and the curve $\pi \circ \tilde{\boldsymbol{\gamma}}$ os a geodesic of $\mathcal{M}$ of same length as $\boldsymbol{\gamma}$. The horizontal geodesics is shown in Fig. B.4.

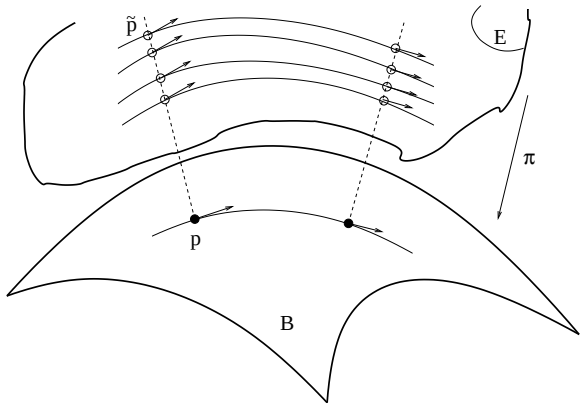


Figure B.4: Horizontal geodesics

If $\tilde{\mathbf{P}} \in \mathcal{E}$ and $\boldsymbol{\gamma}$ is a geodesic of $\mathcal{M}$ with $\pi(\tilde{\mathbf{P}}) = \boldsymbol{\gamma}(0)$, then there exists a unique local horizontal lift $\tilde{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$ such that $\tilde{\boldsymbol{\gamma}}(0) = \tilde{\mathbf{P}}$, and $\tilde{\boldsymbol{\gamma}}$ is also a geodesic of $\mathcal{E}$. This is illustrated in Fig B.5.

If $\mathcal{E}$ is complete, so is $\mathcal{M}$.

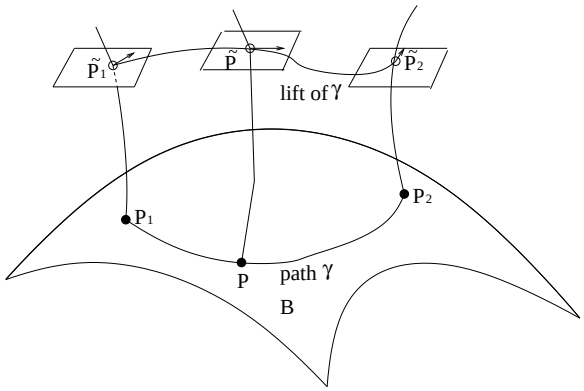


Figure B.5: Illustration of horizontal lift

Appendix C

Proof of Lemma 3.1

Proof: Let $(\mathcal{A}, \dagger)$ denote a set $\mathcal{A}$ associated with a map $^\dagger : \mathcal{A} \to \mathcal{A}$ which has the following properties

- For all $\mathbf{X}, \mathbf{Y} \in \mathcal{A}$

$$(\mathbf{X} + \mathbf{Y})^\dagger = \mathbf{X}^\dagger + \mathbf{Y}^\dagger \tag{C.1}$$

$$(\mathbf{XY})^\dagger = \mathbf{Y}^\dagger \mathbf{X}^\dagger \tag{C.2}$$

- For every $\lambda \in \mathbb{C}$ and every $\mathbf{X} \in \mathcal{A}$

$$(\lambda \mathbf{X})^\dagger = \lambda^* \mathbf{X}^\dagger \tag{C.3}$$

- For all $\mathbf{X} \in \mathcal{A}$

$$(\mathbf{X}^\dagger)^\dagger = \mathbf{X} \tag{C.4}$$

- For all $\mathbf{X} \in \mathcal{A}$

$$\|\mathbf{X}^\dagger \mathbf{X}\| = \|\mathbf{X}\| \|\mathbf{X}^\dagger\| \tag{C.5}$$

Let ϕ be a linear functional on $\mathcal{A}$. We say that ϕ is a positive linear functional on $\mathcal{A}$ if ϕ is such that $\phi(\mathbf{X}) \geq 0$ for every $\mathcal{A} \ni \mathbf{X} \succeq 0$ (i.e., $\mathbf{X}$ is nonnegative). We may endow

$\mathcal{A}$ with an inner product $\langle \mathbf{X}, \mathbf{Y} \rangle_\phi \triangleq \phi(\mathbf{X}^\dagger \mathbf{Y})$ (we use the subscript ϕ to indicate the inner product is defined with the use of a given positive linear functional ϕ) so that $\mathcal{A}$ is then a Hilbert space, which we denoted by $\mathcal{H} = (\mathcal{A}, \langle \cdot, \cdot, \rangle_\phi)$. For every $\mathbf{X} \in \mathcal{A}$ if we can define an operator $T_{\mathbf{X}}$ on $\mathcal{H}$, then it can be shown that there exits a vector $\tilde{\mathbf{X}} \in \mathcal{H}$ such that

$$\phi(\mathbf{X}) = \langle \tilde{\mathbf{X}}, T_{\mathbf{X}} \tilde{\mathbf{X}} \rangle_\phi \tag{C.6}$$

holds for any $\mathbf{X} \in \mathcal{A}$. This is called the Gelfand-Naimark-Segal (GNS) Construction [27].

Let $\mathcal{M}_M$ be the set of all the $M \times M$ complex matrices. Then, it can be verified that the matrix Hermitian H on $\mathcal{M}_M$ acts the same rule as $\dagger$ on $\mathcal{A}$. We define a positive linear functional on $\mathcal{M}_M$, $\phi : \mathcal{M}_M \to \mathbb{R}$ by

$$\phi(\mathbf{X}) = \mathrm{Tr}\mathbf{X}, \quad \mathbf{X} \in \mathcal{M}_M \tag{C.7}$$

The positivity of this functional follows from the fact that the trace of a nonnegative matrix is nonnegative. Let $\mathcal{H}_M = (\mathcal{M}_M, \langle \cdot, \cdot, \rangle_\phi)$ be the Hilbert space forme by $M \times M$ complex matrices with the inner product defined by $\langle \mathbf{X}, \mathbf{Y} \rangle_\phi \triangleq \phi(\mathbf{X}^H \mathbf{Y})$, $\mathbf{X}, \mathbf{Y} \in \mathcal{M}_M$. For every $\mathbf{X} \in \mathcal{M}_M$, we define an operator $T_{\mathbf{X}}$ on $\mathcal{H}_M$ by

$$T_{\mathbf{X}} \tilde{\mathbf{X}} = \mathbf{X}\tilde{\mathbf{X}}, \;\; \tilde{\mathbf{X}} \in \mathcal{H}_M \tag{C.8}$$

i.e., the operator is of the left multiplication by $\mathbf{X}$ on $\mathcal{H}_M$. Then, by the GNS construction, there is a vector $\tilde{\mathbf{X}} \in \mathcal{H}_M$ such that

$$\phi(\mathbf{X}) = \langle \tilde{\mathbf{X}}, T_{\mathbf{X}} \tilde{\mathbf{X}} \rangle_\phi = \langle \tilde{\mathbf{X}}, \mathbf{X}\tilde{\mathbf{X}} \rangle_\phi, \quad \forall \mathbf{X} \in \mathcal{M}_M. \tag{C.9}$$

Let $\mathcal{M} \subset \mathcal{M}_M$ be the manifold of all the positive definite Hermitian matrices. Since for any $\mathbf{P} \in \mathcal{M}$ and any $\mathbf{X} \in \mathcal{M}_M$, we have $\mathbf{PX} \in \mathcal{M}_M$. Thus, applying Eq. (C.9) to $\mathbf{PX}$ we must have

$$\phi(\mathbf{PX}) = \langle \tilde{\mathbf{X}}, T_{\mathbf{PX}} \tilde{\mathbf{X}} \rangle_\phi = \langle \tilde{\mathbf{X}}, \mathbf{PX}\tilde{\mathbf{X}} \rangle_\phi, \quad \forall \mathbf{X} \in \mathcal{M}_M. \tag{C.10}$$

By the definition of ϕ, we have

$$\mathrm{Tr}\mathbf{P}\mathbf{X} = \mathrm{Tr}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathrm{H}}\mathbf{P}\mathbf{X}, \ \ \forall \mathbf{X} \in \mathcal{M}_{\mathrm{M}} \tag{C.11}$$

holds. This is true if and only if $\mathbf{P} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^{H}\mathbf{P}$. Since $\mathbf{P}$ is nonsingular we have $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{H} = \mathbf{I}$.

For every $\mathbf{X} \in \mathcal{M}_M$ and a $\mathbf{P} \in \mathcal{M}$, we now define another operator on $\mathcal{H}_M$ by

$$T'_{\mathbf{X}}\tilde{\mathbf{X}} = \mathbf{P}T_{\mathbf{X}}\tilde{\mathbf{X}}, \ \ \tilde{\mathbf{X}} \in \mathcal{H}_M \tag{C.12}$$

where $T_{\mathbf{X}}$ is the operator defined in Eq. (C.8).

Then, applying the GNS construction again, there exists a vector $\tilde{\mathbf{X}}' \in \mathcal{H}_M$ such that

$$\phi(\mathbf{X}) = \langle \tilde{\mathbf{X}}', T'_{\mathbf{X}}\tilde{\mathbf{X}}' \rangle_\phi = \langle \tilde{\mathbf{X}}', \mathbf{P}T_{\mathbf{X}}\tilde{\mathbf{X}}' \rangle_\phi = \langle \tilde{\mathbf{X}}', \mathbf{P}\mathbf{X}\tilde{\mathbf{X}}' \rangle_\phi, \quad \forall \mathbf{X} \in \mathcal{M}_M. \tag{C.13}$$

Since for fixed $\mathbf{X} \in \mathcal{M}_M$ and $\mathbf{P} \in \mathcal{M}$ the Eq. (C.9) and Eq. (C.13) are the same, we must have

$$\langle \tilde{\mathbf{X}}', \mathbf{P}\mathbf{X}\tilde{\mathbf{X}}' \rangle_\phi = \langle \tilde{\mathbf{X}}, \mathbf{X}\tilde{\mathbf{X}} \rangle_\phi \tag{C.14}$$

This holds if and only if $\tilde{\mathbf{X}}'(\tilde{\mathbf{X}}')^H\mathbf{P} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^H = \mathbf{I}$. Thus, we have $\mathbf{P} = (\tilde{\mathbf{X}}'(\tilde{\mathbf{X}}')^H)^{-1}$. Let $\tilde{\mathbf{P}} = ((\tilde{\mathbf{X}}')^H)^{-1}$. Then we have

$$\mathbf{P} = \tilde{\mathbf{P}}\tilde{\mathbf{P}}^H \tag{C.15}$$

Therefore, we conclude that $\mathbf{P} = \tilde{\mathbf{P}}\tilde{\mathbf{P}}^H$ is a necessary and sufficient condition for the representative of $\mathbf{P} \in \mathcal{M}$ in the Hilbert space $\mathcal{H}_M$ as $\tilde{\mathbf{P}}$. We note that $\tilde{\mathbf{P}}$ is not unique. □

Appendix D

Proof of Lemma 3.2

Proof: Let $\mathbf{P} \in \mathcal{M}$ such that $\mathbf{P} = \tilde{\mathbf{P}}\tilde{\mathbf{P}}^H$, where $\tilde{\mathbf{P}} \in \tilde{\mathcal{H}}$. First, we show that the tangent space $\mathcal{T}_{\tilde{\mathcal{H}}}$ at $\tilde{\mathbf{P}} \in \tilde{\mathcal{H}}$ can be decomposed as $\mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}}) = \mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}}) \oplus \mathcal{V}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ such that $\mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}}) \perp \mathcal{V}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$.

Let $\mathcal{G} = \{\mathbf{G} : \mathbf{G}^H\mathbf{G} = \mathbf{I}\}$ be the unitary group acting on $\mathcal{H}$ by right multiplication such that $\tilde{\mathbf{P}}\mathbf{G} \neq \tilde{\mathbf{P}}$ for $\tilde{\mathbf{P}} \in \mathcal{H}$ if $\mathbf{G} \neq \mathbf{I}$. It is well-known that [8] the Lie algebra of $\mathcal{G}$ is $\mathfrak{g} = \mathcal{T}_{\mathcal{G}}(\mathbf{I}) = \{\mathbf{S} : \mathbf{S}^H = -\mathbf{S}\}$ (i.e.,the tangent space of $\mathcal{G}$ at its identity element $\mathbf{I}$) and $e^{r\mathbf{S}} \in \mathcal{G}$ for any $r \in \mathbb{R}$ and $\mathbf{S} \in \mathfrak{g}$.

Let $(-\epsilon, \epsilon) \subset \mathbb{R}$ be a small open interval in $\mathbb{R}$. For any $r \in (-\epsilon, \epsilon)$, let $\tilde{\mathbf{A}}(r) = \tilde{\mathbf{P}}e^{r\mathbf{S}}$. Then we have $\tilde{\mathbf{A}}(r)(\tilde{\mathbf{A}}(r))^H = \tilde{\mathbf{P}}\tilde{\mathbf{P}}^H \in \mathcal{M}$ and $\tilde{\mathbf{A}}(0) = \tilde{\mathbf{P}}$. Therefore, the mapping $\tilde{\mathbf{A}}(r) : (-\epsilon, \epsilon) \to \tilde{\mathcal{H}}$ defines a parameterized curve $\tilde{\mathbf{A}}(r)$ on $\tilde{\mathcal{H}}$, which is through $\tilde{\mathbf{P}}$ at $r = 0$. It is easy to verify that the tangent vector at $\tilde{\mathbf{P}}$ is $\frac{d}{dr}\mathbf{A}(r)_{r=0} = \tilde{\mathbf{P}}\mathbf{S} \in \mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$.

Now, let $\tilde{\mathbf{\Gamma}}(r) : (-\epsilon, \epsilon) \to \tilde{\mathcal{H}}$ be another different curve on $\tilde{\mathcal{H}}$ through $\tilde{\mathbf{P}}$ at $r = 0$. We denote the tangent vector of this curve at $\tilde{\mathbf{P}}$ by $\dot{\tilde{\mathbf{P}}} = \frac{d}{dr}\tilde{\mathbf{\Gamma}}(r)|_{r=0} \in \mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$.

Let $\tilde{\mathbf{P}}\mathbf{S}$ and $\dot{\tilde{\mathbf{P}}}$ be orthogonal. Then, we have that

$$\begin{aligned}\langle \tilde{\mathbf{P}}\mathbf{S}, \dot{\tilde{\mathbf{P}}} \rangle_{\mathcal{T}_{\mathcal{H}}(\tilde{\mathbf{P}})} &= \frac{1}{2}\mathrm{Tr}((\tilde{\mathbf{P}}\mathbf{S})^H \dot{\tilde{\mathbf{P}}} + \dot{\tilde{\mathbf{P}}}^H(\tilde{\mathbf{P}}\mathbf{S})) \\ &= \frac{1}{2}\mathrm{Tr}((\dot{\tilde{\mathbf{P}}}^H\tilde{\mathbf{P}} - \tilde{\mathbf{P}}^H\dot{\tilde{\mathbf{P}}})\mathbf{S}) \qquad \text{since } \mathbf{S}^H = -\mathbf{S} \\ &= 0 \qquad \text{orthogonality} \end{aligned} \tag{D.1}$$

holds for any $\mathbf{S} \in \mathfrak{g}$ if and only if

$$\tilde{\mathbf{P}}^H\dot{\tilde{\mathbf{P}}} = \dot{\tilde{\mathbf{P}}}^H\tilde{\mathbf{P}} \tag{D.2}$$

From Eq. (D.2), we have

$$\dot{\tilde{\mathbf{P}}} = (\tilde{\mathbf{P}}^H)^{-1}\dot{\tilde{\mathbf{P}}}^H\tilde{\mathbf{P}} = \mathbf{K}\tilde{\mathbf{P}}, \tag{D.3}$$

where $\mathbf{K} = (\tilde{\mathbf{P}}^H)^{-1}\dot{\tilde{\mathbf{P}}}^H$. Then, we can see

$$\mathbf{K}^H = \dot{\tilde{\mathbf{P}}}(\tilde{\mathbf{P}})^{-1} = [(\tilde{\mathbf{P}}^H)^{-1}\dot{\tilde{\mathbf{P}}}^H\tilde{\mathbf{P}}](\tilde{\mathbf{P}})^{-1} = (\tilde{\mathbf{P}}^H)^{-1}\dot{\tilde{\mathbf{P}}}^H = \mathbf{K} \tag{D.4}$$

Conversely, let $\dot{\tilde{\mathbf{P}}} = \mathbf{K}\tilde{\mathbf{P}}$ and $\mathbf{K}^H = \mathbf{K}$. Then we have

$$\tilde{\mathbf{P}}^H\dot{\tilde{\mathbf{P}}} = \tilde{\mathbf{P}}^H\mathbf{K}\tilde{\mathbf{P}} = (\mathbf{K}\tilde{\mathbf{P}})^H\tilde{\mathbf{P}} = \dot{\tilde{\mathbf{P}}}^H\tilde{\mathbf{P}} \tag{D.5}$$

which is exactly the Eq. (D.2). Since each $\dot{\tilde{\mathbf{P}}}$ is the tangent vector of a curve of type $\tilde{\mathbf{\Gamma}}(r)$ at $r = 0$ and the curve is different from the type of $\tilde{\mathbf{A}}(r)$, the Eq. (D.2) is a necessary and sufficient condition that

$$\mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}}) = \{\mathbf{K}\tilde{\mathbf{P}} : \mathbf{K} = \mathbf{K}^H\} \tag{D.6}$$

Let

$$\mathcal{V}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}}) = \{\tilde{\mathbf{P}}\mathbf{S} : \mathbf{S} \in \mathfrak{g}\} \tag{D.7}$$

Then, we have

$$\mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}}) = \mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}}) \oplus \mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}}) \tag{D.8}$$

with $\mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}}) \perp \mathcal{V}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$. We call $\mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ the *horizontal subspace* and $\mathcal{V}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ the *vertical subspace* of $\mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ at $\tilde{\mathbf{P}}$, respectively.

Now, we show the isometric between $\mathcal{T}_{\mathcal{M}}(\mathbf{P})$ and $\mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$. In other words, we need to show that

$$\langle \tilde{\mathbf{A}}, \tilde{\mathbf{B}} \rangle_{\mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})} = g_{\mathbf{P}}(\mathbf{A}, \mathbf{B}) \tag{D.9}$$

holds.

Let $\mathbf{\Gamma}(r) : (-\epsilon, \epsilon) \rightarrow \mathcal{M}$ be a curve on $\mathcal{M}$ such that $\mathbf{\Gamma}(r) = \tilde{\mathbf{\Gamma}}(r)\tilde{\mathbf{\Gamma}}(r)^H$ and $\mathbf{\Gamma}(0) = \mathbf{P}$ with $\mathbf{P} = \tilde{\mathbf{P}}\tilde{\mathbf{P}}^H$. Then, the tangent vector at $\mathbf{P}$ is

$$\dot{\mathbf{P}} = \frac{d}{dr}\mathbf{\Gamma}(r)|_{r=0} = \frac{d}{dr}(\tilde{\mathbf{\Gamma}}(r)\tilde{\mathbf{\Gamma}}(r)^H)|_{r=0} = \dot{\tilde{\mathbf{P}}}\tilde{\mathbf{P}}^H + \tilde{\mathbf{P}}\dot{\tilde{\mathbf{P}}}^H \tag{D.10}$$

Different curves through $\mathbf{P}$ will have different tangent vectors $\dot{\mathbf{P}}$, which altogether form the tangent space $\mathcal{T}_{\mathcal{M}}(\mathbf{P})$ of $\mathcal{M}$ at $\mathbf{P}$. Since $\dot{\mathbf{P}} \in \mathcal{T}_{\mathcal{M}}(\mathbf{P})$ and $\dot{\tilde{\mathbf{P}}} \in \mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$, we have, from Eq. (D.10), for any two tangent vectors $\mathbf{A}, \mathbf{B} \in \mathcal{T}_{\mathcal{M}}(\mathbf{P})$, there exist $\tilde{\mathbf{A}}, \tilde{\mathbf{B}} \in \mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ such that $\mathbf{A} = \tilde{\mathbf{A}}\tilde{\mathbf{P}}^H + \tilde{\mathbf{P}}\tilde{\mathbf{A}}^H$ and $\mathbf{B} = \tilde{\mathbf{B}}\tilde{\mathbf{P}}^H + \tilde{\mathbf{P}}\tilde{\mathbf{B}}^H$. Now, $\mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}}) \subseteq \mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$, thus we can write the inner product between $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$ as

$$\begin{aligned}
\langle \tilde{\mathbf{A}}, \tilde{\mathbf{B}} \rangle_{\mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})} &= \frac{1}{2} Tr(\tilde{\mathbf{A}}^H\tilde{\mathbf{B}} + \tilde{\mathbf{B}}^H\tilde{\mathbf{A}}) \\
&= \frac{1}{2}\mathrm{Tr}(\tilde{\mathbf{A}}^H\mathbf{K}\tilde{\mathbf{P}} + \tilde{\mathbf{P}}^H\mathbf{K}\tilde{\mathbf{A}}) \qquad \text{by Eq. (D.3)} \\
&= \frac{1}{2}\mathrm{Tr}((\tilde{\mathbf{A}}\tilde{\mathbf{P}}^H + \tilde{\mathbf{P}}\tilde{\mathbf{A}}^H)\mathbf{K}) = \frac{1}{2}\mathrm{Tr}\mathbf{A}\mathbf{K}
\end{aligned} \tag{D.11}$$

and

$$\begin{aligned}
\mathbf{B} &= \tilde{\mathbf{B}}\tilde{\mathbf{P}}^H + \tilde{\mathbf{P}}\tilde{\mathbf{B}}^H \\
&= \mathbf{K}\tilde{\mathbf{P}}\tilde{\mathbf{P}}^H + \tilde{\mathbf{P}}\tilde{\mathbf{P}}^H\mathbf{K} \qquad \text{by Eq. (D.3)} \\
&= \mathbf{K}\mathbf{P} + \mathbf{P}\mathbf{K}
\end{aligned} \tag{D.12}$$

By comparing Eq. (D.11) with Eq. (3.45), we can see that $\langle \tilde{\mathbf{A}}, \tilde{\mathbf{B}} \rangle_{\mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})} = g_{\mathbf{P}}(\mathbf{A}, \mathbf{B})$. Therefore, $\mathcal{U}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ and $\mathcal{T}_{\mathcal{M}}(\mathbf{P})$ are isometric. □

Appendix E

Proof of Theorem 3.5

Proof. For the classical proof, please see [17]. Here we give a new proof by using the method we developed in this thesis.

Let $\mathcal{H}_H$ be formed by all the $M \times M$ Hermitian matrices with the inner product induced by restriction of the inner product endowed to the Hilbert space $\mathcal{H}_M$ and is a Hilbert space in its owe right. In other words, we can endow $\mathcal{H}_H$ with an inner product $\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathcal{H}_H} = \mathrm{Tr}\mathbf{X}\mathbf{Y}$ so that $\mathcal{H}_H$ is a Hilbert space, denoted by $(\mathcal{H}_H, \langle \cdot, \cdot \rangle_{\mathcal{H}_H})$. Let

$$\tilde{\mathcal{H}} = \{\tilde{\mathbf{X}} : \tilde{\mathbf{X}} = \log \mathbf{P}, \mathbf{P} \in \mathcal{M}\} \tag{E.1}$$

Obviously, we have $\tilde{\mathcal{H}} \subset \mathcal{H}_H$. For any $\tilde{\mathbf{P}} \in \tilde{\mathcal{H}}$, we may let $\mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}}) = \mathcal{T}_{\mathcal{H}_H}(\tilde{\mathbf{P}})$ with the inner product induced by the inner product endowed to the Hilbert space $\mathcal{H}_H$.

Let $\mathbf{\Gamma}(r) : (-\epsilon, \epsilon) \to \mathcal{M}$ be a curve on $\mathcal{M}$ such that $\mathbf{\Gamma}(0) = \mathbf{P}$. Let $\tilde{\mathbf{\Gamma}}(r) : (-\epsilon, \epsilon) \to \tilde{\mathcal{H}}$ be a curve on $\tilde{\mathcal{H}}$ such that $\tilde{\mathbf{\Gamma}}(0) = \tilde{\mathbf{P}}$ and $\tilde{\mathbf{P}} = \log \mathbf{P}$. We assume that

$$\tilde{\mathbf{\Gamma}}(r) = \log \mathbf{\Gamma}(r) \tag{E.2}$$

i.e.,

$$\mathbf{\Gamma}(r) = e^{\tilde{\mathbf{\Gamma}}(r)} \tag{E.3}$$

when $r \in (-\epsilon, \epsilon)$. Taking derivative on both sides we have

$$\dot{\mathbf{\Gamma}}(r) = \dot{\tilde{\mathbf{\Gamma}}}(r)\mathbf{\Gamma}(r) \tag{E.4}$$

Denoting by $\dot{\mathbf{P}} = \frac{d}{dr}\mathbf{\Gamma}(r)|_{r=0}$ and $\dot{\tilde{\mathbf{P}}} = \frac{d}{dr}\tilde{\mathbf{\Gamma}}(r)|_{r=0}$ we have

$$\dot{\mathbf{P}} = \dot{\tilde{\mathbf{P}}}\mathbf{P} \tag{E.5}$$

Therefore, for $\mathbf{A}, \mathbf{B} \in \mathcal{T}_{\mathcal{M}}(\mathbf{P})$ we have

$$\mathbf{A} = \tilde{\mathbf{A}}\mathbf{P} \tag{E.6}$$

and

$$\mathbf{B} = \tilde{\mathbf{B}}\mathbf{P} \tag{E.7}$$

where $\tilde{\mathbf{A}}, \tilde{\mathbf{B}} \in \mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$. Thus,

$$\tilde{\mathbf{A}} = \mathbf{A}\mathbf{P}^{-1} \tag{E.8}$$

and

$$\tilde{\mathbf{B}} = \mathbf{B}\mathbf{P}^{-1} \tag{E.9}$$

Therefore, we have

$$\langle \tilde{\mathbf{A}}, \tilde{\mathbf{B}} \rangle = \mathrm{Tr}\tilde{\mathbf{A}}\tilde{\mathbf{B}} = \mathrm{Tr}\mathbf{A}\mathbf{P}^{-1}\mathbf{B}\mathbf{P}^{-1} \tag{E.10}$$

which is exactly the same as Eq. (3.103). Therefore, we have show that $\mathcal{T}_{\mathcal{M}}(\mathbf{P})$ and $\mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ are isometric. Since for any $\tilde{\mathbf{P}}$, $\mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}})$ and $\mathcal{T}_{\mathcal{H}_H}(\tilde{\mathbf{P}})$ have the same metric, the shortest curve connecting two points in $\tilde{\mathcal{H}}$ must be a straight line $\mathcal{H}_H$. This implies that $d_{R_3}(\mathbf{P}_1, \mathbf{P}_2) = \| \log \mathbf{P}_1 - \log \mathbf{P}_2 \|_2$.

On the other hand, let $\mathbf{W} = \mathbf{\Omega}\mathbf{\Omega}^H$ and $\mathbf{P}_W = \mathbf{\Omega}^H\mathbf{P}\mathbf{\Omega}$. Then, we have $\mathbf{A}_W = \mathbf{\Omega}^H\mathbf{A}\mathbf{\Omega}$ and $\mathbf{B}_W = \mathbf{\Omega}^H\mathbf{B}\mathbf{\Omega}$, where $\mathbf{A}_W, \mathbf{B}_W \in \mathcal{T}_{\tilde{\mathcal{H}}}(\tilde{\mathbf{P}}_W)$ and $\mathbf{A}, \mathbf{B} \in \mathcal{T}_{\tilde{\mathcal{H}}}(\mathbf{P})$. Therefore,

we have

$$\begin{aligned} g_{P_W}(\mathbf{A}_W, \mathbf{B}_W) &= \mathrm{Tr}\mathbf{A}_\mathrm{W}{\mathbf{P}_\mathrm{W}}^{-1}\mathbf{B}_\mathrm{W}{\mathbf{P}_\mathrm{W}}^{-1} \\ &= \mathrm{Tr}(\mathbf{\Omega}^\mathrm{H}\mathbf{A}\mathbf{\Omega})(\mathbf{\Omega}^\mathrm{H}\mathbf{P}\mathbf{\Omega})^{-1}(\mathbf{\Omega}^\mathrm{H}\mathbf{B}\mathbf{\Omega})(\mathbf{\Omega}^\mathrm{H}\mathbf{P}\mathbf{\Omega})^{-1} \\ &= \mathbf{\Omega}^H\mathbf{A}\mathbf{P}^{-1}\mathbf{B}\mathbf{P}^{-1}\mathbf{\Omega})^{-1} \\ &= \mathrm{Tr}\mathbf{A}\mathbf{P}^{-1}\mathbf{B}\mathbf{P}^{-1} = \mathrm{g_P}(\mathbf{A}, \mathbf{B}) \end{aligned} \tag{E.11}$$

This means that the Riemannian metric is weighting invariant. Thus, it implies that the Riemannian distances are weighting invariant, i.e., $d_{R3}(\mathbf{P}_{1W}, \mathbf{P}_{2W}) = d_{R3}(\mathbf{P}_1, \mathbf{P}_2)$. As a result of this, we have $d_{R3}(\mathbf{P}_1, \mathbf{P}_2) = d_{R3}(\mathbf{I}, {\mathbf{P}_1}^{-1/2}\mathbf{P}_2{\mathbf{P}_1}^{-1/2})$.

Therefore, we have

$$\begin{aligned} d_{R3}(\mathbf{P}_1, \mathbf{P}_2) &= \sqrt{\mathrm{Tr}\Big(\log\mathbf{P}_1 - \log\mathbf{P}_2\Big)^2} \\ &= \sqrt{\mathrm{Tr}\Big(\log\mathbf{I} - \log\mathbf{P}_1^{-1/2}\mathbf{P}_2\mathbf{P}_1^{-1/2}\Big)^2} \\ &= \sqrt{\mathrm{Tr}\Big(\log\mathbf{P}_1^{-1/2}\mathbf{P}_2\mathbf{P}_1^{-1/2}\Big)^2} \\ &= \sqrt{\sum_{i=1}^{n}\log^2\lambda_i(\mathbf{P}_1^{-1}\mathbf{P}_2)}, \end{aligned} \tag{E.12}$$

where λ_i are the eigenvalues of $\mathbf{P}_1^{-1}\mathbf{P}_2$. □

Bibliography

[1] U. R. Abeyratne, V. Swarnker, S. I. Rathnayake, and C. Hukins, "Sleep-stage and event dependency of brain asynchrony as manifested through surface EEG", *Proceedings of the 29th Annual International Conference of the IEEE EMBS, Lyon, France*, pp. 709-712, 2007.

[2] P. Achermann and A. A. Borbély, "Temporal evolution of coherence and power in the human sleep electroencephalogram", *Journal of Sleep Research*, vol. 7, suppl. 1, pp. 36-41, 1998.

[3] S. Amari, *Differential-Geometrical Methods in Statistics*, Lecture Notes in Statistics, Springer-Verlag, Berlin, 1985.

[4] P. Anderer, G. Gruber, S. Parapatics, and G. Dorffner, "Automatic sleep classification according to Rechtschaffen and Kales", *Proceedings of the 29th Annual International Conference of the IEEE EMBS, Lyon, France*, pp.3994-3997, 2007.

[5] C. W. Anderson, J. N. Knight, T. O'connor, M. J. Kirby, and A. Sokolov, "Geometric subspace methods and time-delay embedding for EEG artifacts removal and classification", *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* Vol. 14, No. 2, pp. 142-146, 2006.

[6] E. Aserinsky and N. Kleitman, "Regularly occurring periods of eye motility and concomitant phenomena during sleep", *Science*, vol 118:273274, 1953.

[7] M. H. Asyali, R. B. Berry, M. C. K. Khoo, and A. Altinok, "Determining a continuous marker for sleep depth", *Computers in Biology and Medicine,* vol. 37,no. 11, pp. 1600-1609, 2007.

[8] A. Baker, *Matrix Groups: An Introduction to Lie Group Theory*, Springer, 2003.

[9] D. Balakrishnan and S. Puthusserypady, "Multilayer perceptrons for the classification of brain computer interface data", *Proc. IEEE 31st Annual Northeast Bioengineering Conference*, 2005.

[10] V. Barnett and T. Lewis, *Outliers in Statistical Data,* 3rd Ed., Wiley, 1994.

[11] G. Becq, S. Charbonnier, F. Chapotot, A. Buguet, L. Bourdon and P. Baconnier, "Comparison between five classifiers for automatic scoring of human sleep recordings", *Studies in Computational Intelligence (SCI)*, vol. 4, pp.113-127, 2005.

[12] I. Bengtsson and K. Zyczkowski, *Geometry of Quantum States: An Introduction to Quantum Entanglement*, Cambridge University Press, 2008.

[13] M. Berger, *A Panoramic View of Riemannian Geometry,* Springer, 2003.

[14] D. S. Bernstein, *Matrix Mathematics: Theory, Facts, and Formulas with Application to Linear Systems*, Princeton University Press, 2005.

[15] C. Berthomier, X. Drouot, M. Herman-Stoïca, P. Berthomier, J. Prado, D. Bokar-Thire, O. Benoit, J. Mattout, and M-P, d'Ortho, "Automatic analysis of single-channel sleep EEG: validation in healthy individuals", *Sleep,* vol. 30, no. 11, pp. 1587-1595, 2007.

[16] A. Besilevsky, *Applied Matrix Algebra in the Statisical Sciences*, North-Holland, 1983.

[17] R. Bhatia, *Positive Definite Matrices,* Princeton University Press, 2006.

[18] B. Bianchi and R. Sorrentino, *Electronic Filter Simulation & Design*, McGraw-Hill Professional, 2007.

[19] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1996.

[20] J. F. Borisoff, S. G. Mason, A. Bashashati, and G.E. Birch, "Brain-computer interface design for asynchronous control applications: improvements to the LF-ASD asynchronous brain switch", *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp.985-992, 2004.

[21] D. Bures, "An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite w^*-algebras", *Trans. Amer. Math. Soc.* **135**, pp. 199-212, 1969.

[22] G. Buzsáki, *Rhythms of the Brain*, Oxford University Press, 2006.

[23] C. Cao, R. L. Tutwiler, and S. Slobounov, "Automatic classification of athletes with residual functional deficits following concussion by means of EEG signal using support vector machine", *IEEE Trans. Neural Systems and Rehabilitation Engineering*, vol. 16, no. 4, pp. 327-335, Aug. 2008.

[24] G-M Carlos and N. V. Angel, "Automatic removal of ocular artifacts from EEG data using adaptive filtering and independent component analysi", *17th European Signal Processing Conference,* pp.2317 - 2321, 2009.

[25] F. Cincotti, A. Scipione, A. Tiniperi, D. Mattia, M. G. Marcinani, J. R. Millán, S. Salinari, L. bianchi, and F. Babiloni, "Comparison of different feature classifiers for brain computer interface", *First Int. IEEE EMBS Conf. on Neural Engineering, Conf. Proc.* , pp. 645-647, 2003.

[26] S. R. Cloude and E. Pottier, "Concept of polarization entropy in optical scattering", *Optical Engineering,* Vol. 34, No.6, pp. 1599-1610, 1995.

[27] J. B. Conway, *A Course in Functional Analysis,* Springer, 2nd Ed., 1994.

[28] S. Crisler, M.J. Morrisey, A. M. Anch, and D. W. Barnett, "Sleep-stage scoring in the rat using a support vector machine", *Journal of Neuroscience Methods*, vol. 168, no. 2, pp. 524-535, 2008.

[29] J. Dattorro, *Convex Optimization and Euclidean Distance Geometry,* Meboo Publishing USA, 2010.

[30] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, 1996.

[31] P. A. M. Dirac, *Principles of Quantum Mechanics,* 4th Ed, Oxford, England: Oxford University Press, 1982.

[32] L. Doroshenkov, V. Konyshev, S. Selishchev, "Classification of human sleep stages based on EEG processing using Hidden Markov models", *Biomedical Engineering*, vol. 41, no. 1, pp.25-28, 2007.

[33] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Recognition*, New York: Wiley-Interscience, 2001.

[34] T. S. Elali, *Discrete Systems and Digital Signal Processing with MATLAB,* CRC press, 2005.

[35] H. Flanders, *Differential Forms with Applications to the Physical Sciences,* Academic Press, 1963.

[36] L. Franks, *Signal Theory*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1969.

[37] M. Fréchet, "Sur la Distance de Deux Lois de Probabilite", *C. R. Acad. Sci. Paris,* vol. 244, pp. 689692, 1957.

[38] K. Fukunaga, *Statistical Pattern Recognition*, New York: Academic, 1990.

[39] J. Gallier, *Geometric Methods and Applications - For Computer Science and Engineering*, Springer, 2001.

[40] S. Gallot, D. Hulin and J. Lafontaine, *Riemannian Geometry*, Springer 3rd Ed., 2004.

[41] G. N. Garcia, T. Ebrahimi and J-M Vesin, "Support vector EEG classification in the Fourier and time-frequency correlation domain", in *Conference Proc. 1st Int. IEEE EMBS Conf. on Neural Engineering*, 2003.

[42] D. Garrett, D. D. Peterson, C. W. Anderson and M. H. Thaut, "Comparison of linear, nonlinear, and feature selection methods for EEG signal classification", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, pp. 141–144, 2003.

[43] W. Gersch, "Nearest neighbor rule classification of stationary and nonstationary time series", in *App. Time Series Anal. II*, Academic Press, Inc., 1981.

[44] S. Ghosh-Dastidar, H. Adeli, and N. Dadmehr, "Principal component analysis-enhanced cosine radial basis function neural network for robust epilepsy and seizure detection", *IEEE Trans. Biomedical Engineering*, vol. 55, no. 2, pp. 512-518, Feb. 2008.

[45] A. Graham, *Kronecker Products and Matrix Calculus with Applications*, Ellis Horwood Ltd., 1981.

[46] G. Gratton, MG Coles, and E. Donchin, "A new method for off-line removal of ocular artifact", *Electroencephalography and Clinical Neurophysiology,* Vol. 55, No. 4, pp. 468-484, 1983.

[47] A. Gut, *Probability: A Graduate Course,* Springer, 2005.

[48] H. Hinrikus, A Suhhova, M Bachmann, K. Aadamsoo, Ũ . Võhma, J. Lass and V. Tuulik, "Electroencephalographic spectral asymmetry index for detection of depression", *Med. Biol. Eng. Comput.*, vol. 47, pp. 1291-1299, Dec. 2009.

[49] A. Hiraiwa, K. Shimohara and Y. Tokunaga, "EEG topography recognition by neural networks", *IEEE Eng. Med. Biol. Mag.*, vol. 9, pp. 39–42, 1990.

[50] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 1990.

[51] L.-K. Hua, *Introduction to Number Theory,* Springer, 1987.

[52] A. K. Jain, R. P. W. Duin and J. Mao, "Statistical pattern recognition: a review", *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 22, pp.4-37, 2000.

[53] J. Jost, *Riemannian Geometry and Geometric Analysis*, Springer, 2008.

[54] H. Karcher, "Riemannian center of mass and mollifier smoothing", *Communications on Pure and Applied Mathematics*, 30(5), pp. 509-541, 1977.

[55] S. Kay, *Modern Spectral Estimation: Theory and Application,* Prentice-Hall, 1988.

[56] Z. A. Keirn and J. I. Aunon, "A new mode of communication between man and his surrondings", *IEEE Transactions on Biomedical Engineering*, vol. 37, pp.1209-1214, 1990.

[57] D.G. Kendall, D. Barden, T.K. Carne, and H. Le, *Shape and Shape Theory*, Wiley, 1999.

[58] S. Kullback, *Information Theory and Statistics,* John Weily and Sons, NY, 1959.

[59] S. Kullback and R.A. Leibler, "On information and sufficiency", *Annals of Mathematical Statistics*, vol.22, no.1, pp.79-86, 1951.

[60] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, Academic Press, 1985.

[61] M. Lipschutz, *Differential Geometry*, Schaum's Outline Series, McGraw-Hill, 1969.

[62] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Amaldi, "A review of classification algorithms for EEG-based brain-computer interfaces", *J. Neural Eng.*, vol. 4, no. 2, pp. 1-13, 2007.

[63] H. Lütkepohl, *Handbook of Matrices*, John Wiley & Sons, Ltd, 1996.

[64] J. McEwen and G. B. Anderson, "Modelling the stationarity and Gaussianity of spontaneous electroencephalographic activity", *IEEE Transactions on Biomedical Engineering,* vol. 22, no. 5, pp.361-369, 1975.

[65] J. R. Millán, J. Mouriñ, F. Cincotti, F. Babiloni, M. Varsta and J. Heikkonen, "Local neural classifier for EEG-based recognition of mental tasks", *IEEE-INNS-ENNS Int. Joint Conf. on Neural Networks*, 2000.

[66] A.W. Naylor and G.R. Sell, *Linear Operator Theory un Engineering and Science*, Holt, Rinehart, and Winston, Inc., 1971.

[67] N. Nguyen and Y. Guo, "Metric learning: a support vector approach", in *Machine Learning and Knowledge Discovery in Databases,* Springer, Berlin, 2008.

[68] A. H. Nuttall, "Multivariate linear predictive spectral analysis employing withghted forward and backward averaging: a generalization of Burg's algorithm", in *Naval Underwater Systems Center Technical Report 5501*, New London, Conn., 1976.

[69] B. Obermeier, C. Guger, C. Neuper and G. Pfurtscheller, "Hidden Markov models for online classification of single trial EEG", *Pattern Recognit. Lett.*, pp.1299-1309, 2001.

[70] B. Obermeier, C. Neuper, C. Guger and G. Pfurtscheller, "Information transfer rate in a five-classes brain-computer interface", *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 9, pp.283-288, 2000.

[71] F. Oveisi and A. Erfanian, "A tree-structure mutual information-based feature extraction and its application to EEG-based Brain-Computer Interfacing", *Proceedings of the 29th Annual International Conference of the IEEE EMBS, Lyon, France*, pp. 5075-5078, 2007.

[72] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, Inc., 3rd Edition, 1991.

[73] A. Papoulis, *Signal Analysis*, McGraw-Hill, Inc., 1977.

[74] A. Rechtschaffen and A. Kales, "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects", *Public Health Service, U.S. Government Printing Office*, 1968.

[75] G. S. Sebestyen, *Decision-Making Processes in Pattern Recognition*, the Macmillan Company, New York, 1962.

[76] SY Shao, KQ Shen, CJ Ong EP W-S, and XP Li, "Automatic EEG artifact removal: a weighted support vector machine approach with error correction", *IEEE Transactions on Biomedical Engineering,* Vol. 56, No. 2, pp. 336 - 344, 2008.

[77] G.F. Simmons, *Introduction to Topology and Modern Analysis*, McGraw-Hill, Inc., 1963.

[78] M. Spivak, *A Comprehensive Introduction to Differential Geometry, Vol. 2,* Publish or Perish; 2nd edition, 1990.

[79] O. N. Strand, "Multichannel complex maximum entropy (autoregressive) spectral analysis", *IEEE Trans. Auto. control.*, vol. 22, no. 4, pp. 634-640, 1977.

[80] K. Tapp, *Matrix Groups for Undergraduates,* American Mathematical Society, 2005.

[81] J. Virkalla, J. Hasan, A. Värri, S-L Himanen, and K. Müller, "Automatic sleep stage classification using two-channel electro-oculography", *Journal of Neuroscience Methods,* vol. 166, no. 1, pp. 109-115, 2007.

[82] Y. Wang, Z. Zhang, Y. Li, X. Gao, S. Gao and F. Yang, "BCI competition 2003 - data set iv: an algorithm based on CSSD and FDA for classifying single-trial EEG", *IEEE Transactions on Biomedical Engineering*, vol. 51, pp.1081-1086, 2003.

[83] H. J. Weber and G. B., Arfken, *Mathematical methods for physicists*, Academic press, 6ed, 2005.

[84] A. Wennberg and L. H. Zetterberg, "Application of a computer-based model for EEG analysis", *Electroencephalography and Clinical Neurophysiology,* vol. 31, no. 5, pp. 457-468, 1971.

[85] A. N. Whitehead, *An Introduction to Mathematics*, London: Thornton Butterworth, 1911.

[86] JC Woestengurg, MN Verbaten, and JL Slangen, "The removal of the eye movement artifact from the EEG by regression analysis in the frequency domain", *Biological Physiology*, Vol. 16, pp. 127-147, 1982.

[87] K.M. Wong, *Lecture Notes on Signal Theory*, McMaster University, 2001.

[88] K.M. Wong, J.P. Reilly, Q. Wu, and S. Qiao, "Estimation of the direction of arrival of signals in unknown correlated noise Pt. I: The MAP approach and its implementation", *IEEE Transactions on Signal Processing*, vol.40, no.8, pp. 2007 2017, Aug. 1992.

[89] P. Xavier, K. Behbehani, D. Watenpaugh, and J. R. Burk, "Detecting encephalography variations due to sleep disordered breathing events", *Proceedings of the 29th Annual International Conference of the IEEE EMBS, Lyon, France*, pp. 6097-6100, 2007.

[90] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning with application to clustering with side-information", *Advances in NIPS*, vol. 15, pp. 505-512, 2003.

[91] A. M. Yaglom, *An Introductoin to the Theory of Stationary Random Functions,* Prentice-Hall, 1962.

[92] K-S Yoo, T. Basa, and W-H Lee, "Removal of eye blink artifacts from EEG signals based on cross-correlation", *International conference on convergence information technology,* Vol, No. 21-23, pp.2005-2014, 2007.