ADAPTIVE FEATURE SELECTION IN PATTERN RECOGNITION AND ULTRA-WIDEBAND RADAR SIGNAL ANALYSIS

Thesis by

Hao Jiang

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2008

(Defended February 13, 2008)

ACKNOWLEDGEMENTS

Foremost, I wish to express my sincerest gratitude to my advisor, Professor Joel W Burdick, for his tremendous support during my last six years at Caltech. The keen scientific insight of Professor Burdick has incessantly shaped my horizon on the research subjects. It is his wise elicitations that pave the way for the final materialization of every detail in this thesis. Off the research field, Professor Burdick also encourages and guides me to explore diverse interesting topics, and provides kind support to my daily life. I feel so happy to have lived my past years in this fertile and liberal environment. I cannot thank my advisor enough for all he has done for me.

Besides my advisor, I specially would like to thank Dr. Naoki Mitsumoto, who organized an absolutely enjoyable team research, and provided me valuable support and instructive suggestions all the way. The corporation with the talented team members in the radar project marks a happy memory for me. My special thanks go to Dr. Chris Assad, whose smart ideas resulted in an important radar signal preprocessing, and Dr. SangHyun Chang, from whom I gained understanding on multipath radar reflection. Additionally, I must thank Dr. Wissam Musallam and Dr. Eunjung Wang who kindly offered their experimental neural data, which granted me a precious algorithm test bed.

I also thank my friends for their selfless help throughout my time at Caltech. Finally, my particular gratitude is owed to my parents and my wife, Hongying Sha. Their unconditional support, encouragement and patience accompany me all the way from the past to the future.

ABSTRACT

Feature selection from measured data aims to extract informative features to reveal the statistic or stochastic mechanism underlying the complicated or high dimensional original data. In this thesis, the feature selection problem is probed under two situations, one is pattern recognition and the other is ultra-wideband radar signal analysis.

Classical pattern recognition methods select features by their ability to separate the multiple classes with certain gauge measure. The deficiency in this general strategy is its lack of adaptation in specific situations. This deficiency may be overcome by viewing the selected features as a function of not only the training samples but also the unlabeled test data. From this perspective, this thesis proposes an adaptive sequential feature selection algorithm which utilizes an information-theoretic measure to reduce the classification task complexity sequentially, and finally outputs the probabilistic classification result and its variation level. To verify the potential advantage of this algorithm, this thesis applies it to one important problem of neural prosthesis, which concerns decoding a finite number of classes, intended reach directions, from recordings of neural activities in the Parietal Reach Region of one rhesus monkey. Experimental results show that the classification scheme of combining the adaptive sequential feature selection algorithm and the information fusion method outperforms some classical pattern recognition rules, such as the nearest neighbor rule and support vector machine, in decoding performance.

The second scenario in this thesis targets developing a human presence and motion pattern detector through ultra-wideband radar signal analysis. To augment the detection robustness,

both static and dynamic features should be utilized. The static features reflect the information of target geometry and its variability, while the dynamic features extract the temporal structure among radar scans. The problem of static feature selection is explored in this thesis, which utilizes the Procrustes shape analysis to generate the representative template for the target images, and makes statistical inference in the tangent space through the Hotelling one sample T^2 test. After that, the waveform shape variation structure is decomposed in the tangent space through the principal component analysis. The selected principal components not only accentuate the prominent dynamics of the target motion, but also generate another informative classification feature.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS AND TABLES

CHAPTER 1

INTRODUCTION

1.1 Problem Statement

Feature selection from measured data is a general problem which has numerous applications in diverse fields of science and engineering. Nowadays, as large and complex data sets become typical and usual, data gathering efforts are increasingly oriented towards extracting informative features that reveal the statistic or stochastic mechanism which generates the data. In this thesis, the feature selection problem is probed under two scenarios, which are pattern recognition and ultra-wideband radar signal analysis, respectively. In the following, the backgrounds for these two scenarios are reviewed.

1.1.1 Adaptive Sequential Feature Selection

Classical pattern recognition methods decompose the task of classifier design into four stages: data gathering, feature generation, feature selection, and classification/decision rule learning. Generally, the feature generation step transforms the training data, and the feature selection step selects a subset of features from the transformed coefficients, thus reducing the computational complexity of classification rule learning and retaining the discrimination ability as much as possible. Features are often selected by their ability to

separate the multiple classes with a certain gauge measure [Theodoridis and Koutroumbas, 2006]. An optimal or suboptimal subset of features with the highest separability score is usually preferred. The selected feature set spans a subspace onto which the training data and test data are projected, and inside which the classification rule is trained and a final categorical recognition for the test data is made. There are deficiencies to this general approach in specific situations. Firstly, a feature may not necessarily be a good feature even when its separation score is high, but it would very likely be selected after the feature selection stage. For example, color is an excellent feature to distinguish diverse objects, but this feature becomes helpless in a dark environment. Especially this leads to a problem in one important application, neural signal decoding, in which case a single neuron or feature is very weak in terms of classification power. This means their separability scores are similar, therefore the difference between selected features and discarded ones may not be statistically significant. Secondly, classical feature selection depends on the set of training data. As a result, the feature selection process and the following decision rule learning process must start from scratch, if the pool of candidate patterns or the set of training samples changes. This adds a significant computational burden for some applications, such as neural signal decoding. The data generation mechanism for the neural signal is highly variable. To practically deal with the non-stationarity of the neural data, some laboratory simulations of neural prosthesis in non-human primates use a "moving window" where only the most recent N, e.g. $N = 160$, trials are used for the classifier training [Musallam et al., 2004]. To overcome these deficiencies, this thesis proposes a new adaptive sequential feature selection algorithm which simultaneously outputs a final probabilistic classification

result and its variation level. The point is that the selected feature can be a function of not only the training samples, but also of the unlabeled test data. The fact that features are selected adaptively balances the computation burden between the offline learning part and the online part, so that this new algorithm is relatively insensitive to re-instantiation of the data set in the moving window style and re-configuration of the parameters.

The field of brain computer interface (BCI) aims to utilize the understanding of brain functionality and operating mechanisms to enable people to control external devices merely by thought. One important application of BCI techniques is to construct neural prosthetic systems which tap into the thoughts of millions of paralytic patients who are deprived of any motor abilities, but not cognitive functions. While the concept of translating neural activity from the brain into external control signals has existed for decades, substantial progress towards realization of such systems has been made only relatively recently [Musallam et al., 2004; Santhanam et al., 2006; Shenoy et al., 2003; Schwartz and Moran, 2000; Wessberg et al., 2000; Isaacs et al., 2000; Donoghue, 2002; Nicolelis, 2001, 2002]. The construction of such a neuro-prosthetic system is not a small feat. The design and building of such an automatic mechanism involve disciplines ranging from neurobiology to diverse engineering branches. This thesis contributes a new technique to decode a finite number of classes, or intended “reach directions”, from recordings obtained from an electrode array implanted in the subject’s brain. Conceivably, one could use any known classification scheme to decode the neural signal. However, BCIs have several characteristics that challenge the classical classification schemes. Firstly, the real classification process should be performed in real time, e.g. within a few hundred

milliseconds so that the computer interface does not introduce delays between thought and response. Secondly, neural signals are variable and non-stationary, hence the training set data will have time-dependent statistics. Thirdly, the classes have considerable overlaps when conditioned on any selected feature. Lastly, although recordings from multiple neurons are typically available when using implanted electrode arrays, not all of the recorded neurons are well tuned to the task, and some of them may degrade their reliability as exterior conditions change, such as electrode and tissue migration. So how to utilize the information inside all the neurons efficiently and robustly is an important issue. This thesis shows that the application of the proposed adaptive sequential feature selection algorithm, together with an information fusion method, fulfills some of the challenges forementioned, and outperforms some classical pattern recognition rules in decoding performance.

1.1.2 Feature Selection in Ultra-wideband Radar Signal Analysis

As the explosive growth in the number of vehicles worldwide (800 million vehicles in global use), a large number of road accidents happen every year (1.2 million death a year. Among the fatal accidents, 65% of deaths involve pedestrians and 35% of pedestrian deaths are children) [Peden et al., 2004]. The issue of how to boost the vulnerable road user (VRU) safety has become critical for the automobile industry. The motivation behind this research topic is to augment VRU safety by developing the robust human presence and human behavior (walking, jogging, standing, et al.) classifier through the use of car loaded sensors. Ultra-wideband (UWB) impulse radar, which is characterized by its high range resolution, the ability of remote sensing, and the advantage of penetration into/around

obstacles, is one of the most promising sensors to accomplish this task. Although computer vision techniques can assist pedestrian detection, radar based techniques have some distinct advantages, such as detection power beyond stadia distance in foggy conditions or darkness, as well as the ability to see through obstacles or see around corners in advance. On the one hand, the ideal characteristics for UWB systems are its abundant information packing, precise positioning, and extremely low interference. On the other hand, in spite of recent experimental work, there is no satisfactory and systematized theory of UWB radar signal analysis available. The reason is that the process of signal transformation under the context of ultra-wide bandwidth is much more complex than its narrowband counterpart. Hence the well-known narrowband target recognition technique through Doppler shift effect [Richards, 2005] doesn't apply in UWB systems. Novel methods must be developed for our applications.

To prominently distinguish people from other targets and classify people's behaviors, an automatic pedestrian detection system should incorporate as many informative clues as possible. Several prominent features serve this goal, and can be categorized as static features and dynamic features. The static features usually reflect the information of target geometry and its variation structure, while the dynamic features extract the temporal structure among a sequence of radar scans, such as how the radar waveforms evolve over time. Fusion of static and dynamic body biometrics will augment the performance of automatic human identification. This thesis researches how to extract a compact set of static features of the target to unravel the dominant statistical properties of the target images. Moreover, the projection of sequential target images onto the subspace spanned by

the selected features accentuates the prominent target motion patterns, such as the gait, and therefore provides a sound platform to explore the target dynamics. More concretely, a collection of radar data can be geometrically transformed to a cluster of points in a high dimensional space. But due to the redundancy in the data, the dominant variation structure of the data cluster resides in a much lower dimensional subspace. The main task of this feature selection problem is to generate the representative template for target reflected waveforms, locate the high information packing subspace, and explore their statistical and algebraic properties. The selected template or subspace should depend on the data set under study, because the radar waveform shapes and radar signal dynamics are highly distinct for different target geometry, orientation, or motion patterns.

1.2 Contributions of the Thesis

The thesis mainly studies the feature selection problem in two scenarios: pattern recognition and radar signal analysis. This study has been pursued through the following efforts:

(1) Proposing an adaptive sequential feature selection algorithm. The point is that the selected features can be the function of not only the training data, but also the test data, and an information-theoretic criterion is used to make the sequential decision. This feature selection algorithm also unifies the final classification decision, and estimates the variance level of its probabilistic classification result.

(2) Applying the adaptive sequential feature selection algorithm to the problem of discrete neural signal decoding. The combination of the proposed algorithm and one information fusion method presents a new feature selection and classification scheme that is particularly well suited to neural decoding of a finite number of stimuli or command classes, as it fulfills many of the challenges that are specific to BCIs.

(3) Introducing the framework of Procrustes shape analysis and tangent space inference to ultra-wideband radar signal analysis. Procrustes shape analysis is utilized to generate the representative template of radar range profiles, and the statistical shape inference is made in the tangent space of the unit size shape sphere through the Hotelling T^2 test. This framework provides a promising platform for extracting the static and dynamic features of the target, which are informative clues for automatic target recognition.

(4) Analyzing the dominant data variation structure in the tangent space of the unit size shape sphere through principal component analysis. Moreover, by introducing principal component comparison, the similarity of two data variation structures can be quantified, hence providing another informative feature for target recognition. Shape variability analysis, together with Procrustes shape analysis, constitutes a more complete scheme to extract static features for the target.

(5) Proposing a blueprint classification scheme that utilizes all the features, static and dynamic, to make target identification more accurate and robust than using each type of features alone.

1.3 Structure of the Thesis

Chapter 2 proposes an adaptive sequential feature selection algorithm in detail. After that, two optimal properties of the algorithm are proved and its other practical characteristics are reviewed. To illustrate the usefulness of the proposed algorithm, the thesis compares it with one classical pattern recognition method, the nearest neighbor rule, through a simple simulation.

Chapter 3 firstly describes the application background of neural decoding of a finite number of stimuli and its experimental paradigm. After that, Chapter 3 applies the adaptive sequential feature selection algorithm and one information fusion method, the product rule, to the practice of decoding two samples of neural recordings from the Posterior Parietal Cortex of a rhesus monkey. One sample has 4 target directions, and the other one has 8. Experimental results show that for both samples, the proposed classification scheme outperforms some classical recognition methods, such as the nearest neighbor rule and the support vector machine.

Chapter 4 firstly motivates the research topic of target identification through UWB radar signal analysis, provides a brief overview of the UWB system, and proposes several useful preprocessing skills that make radar data more amiable to further analysis. After that, the Procrustes shape analysis is utilized to generate the representative template for the target, and the statistical inference is carried out in the tangent space by Hotelling one sample T^2 test. Furthermore, Chapter 4 uses principal component analysis to analyze the radar waveform shape variability in the tangent space, and measures the similarity between two data variation structures by comparing their corresponding principal components. Empirical tests show that the representative template and the data variation structure both are promising static features for target identification.

Chapter 5 concludes the whole thesis by briefly summarizing the main points of the thesis, pointing to several problems that should be explored further in order to make the adaptive sequential feature selection algorithm and the information fusion method extend more rigorously. Moreover, Chapter 5 sketches a new radar target identification scheme based on both the static and dynamic features, and nails down several challenges that must be cracked to make such a scheme more efficient and robust.

CHAPTER 2

ADAPTIVE SEQUENTIAL FEATURE SELECTION

This chapter proposes a new adaptive sequential feature selection algorithm. After that, two optimal properties of the proposed algorithm are proved and its other practical points are remarked upon. To illustrate the usefulness of the algorithm and prepare for its application in neural signal decoding, this chapter compares it with one classical pattern recognition method, the nearest neighbor rule, through a synthetic simulation. The next chapter explores in more detail the application of this algorithm to the problem of neural signal decoding.

2.1 Problem Statement

The general classification problem that is addressed in this chapter can be stated as follows. Let (X,Y) be a pair of random variables taking their respective values from $\Re^d$ and $\{0,1,\ldots,M-1\}$, where M is the number of classes or patterns. The set of pairs $D_n = \{(X^{(1)},Y^{(1)}),(X^{(2)},Y^{(2)}),\ldots,(X^{(n)},Y^{(n)})\}$ is called the training set with $(X^{(1)},Y^{(1)}),\ldots,(X^{(n)},Y^{(n)})$ being i.i.d. samples from a fixed but unknown distribution governing (X,Y). Let superscripts denote sample index and subscripts denote vector

components, then $X = [X_1, X_2, \ldots, X_d]^T$. Assume that the classes Y take a prior distribution $\{P(Y = j), j = 0,1,\ldots, M-1\}$. The goal is to find a mapping $g : \Re^d \to \{0,1,\ldots, M-1\}$ such that an arbitrary unlabeled test data $x = [x_1, x_2, \ldots, x_d]^T \in \Re^d$ can be classified into one of the M classes, while optimizing some criterion.

2.2 Background

To start, recall that the entropy of a discrete random variable, Z, is defined by

Definition 2.1 [p13, Cover and Thomas, 1991]

$$H(Z) = -\sum_{z \in \Im} P(Z = z) \log(P(Z = z)) \tag{2.1}$$

where $\Im$ is the alphabet set of Z, which has $\{P(Z = z), z \in \Im\}$ as its probability mass function (p.m.f.). This definition leads to the following result.

Proposition 2.1 [p27, Cover and Thomas, 1991]

$$0 \le H(Z) \le \log(M) \text{ if } |\Im| = M$$

where $|\Im|$ represents the cardinality (size) of set $\Im$. Left equality holds if and only if $P(Z = z)'s$ are all zero except one, right equality holds if and only if $P(Z = z) = 1/M$, $\forall z \in \Im$.

■

Proposition 2.1 implies that if a vector is 'sparse' or 'nonuniform', the entropy will be smaller than when it is more 'dense' or 'uniform'. Thus entropy can be a natural measure of the sparseness of a vector. Specifically, the entropy corresponding to the prior p.m.f. is

$$H_0 = -\sum_{i=0}^{M-1} P(Y=i)\log(P(Y=i)) \tag{2.2}$$

and the class entropy conditioned on the sample x_i can be calculated by:

$$H_i = -\sum_{j=0}^{M-1} P(Y=j|X_i=x_i)\log(P(Y=j|X_i=x_i)). \tag{2.3}$$

The quantity $H_0 - H_i$ measures how well the feature x_i (which is the i^{th} component of the unlabeled test data x) reduces the complexity of the classification task. Note that although H_i looks quite like a conditional entropy defined in information theory, it is not exactly of the same form. The information-theoretic conditional entropy is an expectation of H_i with respect to the distribution of X_i. Instead, by calculating the entropy conditioned on the specific observation x_i, $i=1,\ldots,d$, not only is the test data included into the feature selection process, but also the subsequent classification decision is unified in the same framework. The following is the scheme of this idea.

Without loss of generality, it is assumed that $P(Y=j)=1/M$, $j=1,\ldots,M$. This assumption implies an unbiased initial prediction of the classification result, hence $H_0=\log(M)$. From this assumption and (2.3), the class entropy conditioned on the feature x_i can be expressed as:

$$H_i = -\sum_{j=0}^{M-1}\left(\frac{P(X_i = x_i|Y = j)}{\sum_{j=0}^{M-1} P(X_i = x_i|Y = j)} \log(\frac{P(X_i = x_i|Y = j)}{\sum_{j=0}^{M-1} P(X_i = x_i|Y = j)}) \right). \quad (2.4)$$

Define

$$q(x_i) = H_0 - H_i \quad (2.5)$$

then the best feature in terms of reducing the class entropy can be selected as:

$$x_* = \arg\max_{x_i \in \{x_1,\ldots,x_d\}} q(x_i). \quad (2.6)$$

To implement the above scheme, the conditional p.d.f. $P(X_i = x_i|Y = j)$, $j = 0,\ldots,M-1$, must be estimated for each candidate feature x_i. To compensate for the fact that a suitable model of the probability density function may not be available - and even in some applications, such as neural decoding, the statistical mechanism of data generation is highly nonstationary - the conditional p.d.f. is estimated by a nonparametric method, kernel density estimation [Silverman, 1986; Scott, 1992], whose key properties are briefly reviewed below.

Definition 2.2 [Chapter 3, Silverman, 1986]: Assume that U is a random variable with probability density function $f(u)$, and $\{U^{(1)}, U^{(2)}, \ldots, U^{(n)}\}$ is a set of i.i.d. samples from the distribution of U, then the kernel density estimation for $f(u)$ is,

$$\hat{f}(u) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{u - U^{(i)}}{h}) = \frac{1}{n}\sum_{i=1}^{n} K_h(u - U^{(i)}) \quad (2.7)$$

where $K_h(t) = K(t/h)/h$, and the kernel function $K(\cdot)$ satisfies the following properties:

$$\int K(w)dw = 1, \quad \int wK(w)dw = 0, \quad \int w^2 K(w)dw = \sigma_K^2 > 0.$$

After some algebra and approximations, the following theorem is reached. The proof of this theorem can be found in [Chapter 3, Silverman, 1986; Appendix A].

Proposition 2.2 [Chapter 3, Silverman, 1986]: For a univariate kernel density estimator,

$$bias(u) = E\hat{f}(u) - f(u) = \frac{1}{2}\sigma_K^2 h^2 f''(u) + \mathrm{O}(h^4) \tag{2.8}$$

$$\mathrm{var}(u) = \mathrm{var}(\hat{f}(u)) = \frac{f(u)R(K)}{nh} - \frac{f^2(u)}{n} + O(\frac{h}{n}) \tag{2.9}$$

$$IMSE = ISB + IV = \frac{R(K)}{nh} + \frac{1}{4}\sigma_K^4 h^4 R(f'') \tag{2.10}$$

where $IMSE$, ISB, and IV are, respectively, abbreviations for integral of the mean square error, integral of square of the bias, and integral of the variance, and $R(f) = \int f^2(u)du$. Moreover, when h takes the value,

$$h^* = \left(\frac{R(K)}{\sigma_K^4 R(f'')}\right)^{1/5} n^{-1/5} \tag{2.11}$$

the minimum value of the integral of mean square error is

$$IMSE^* = \frac{5}{4}[\sigma_K R(K)]^{4/5} R(f'')^{1/5} n^{-4/5}. \tag{2.12}$$

■

Please note that the ratio of IV to ISB in the $IMSE^*$ is $4:1$ - that is, the ISB^* comprises only 20% of the $IMSE^*$. By the kernel density estimation just explained, one can derive an estimator of $P(X_i = x_i|Y = j)$, denoted as $\hat{P}(X_i = x_i|Y = j)$. Then an estimator for H_i is

$$\hat{H}_i = -\sum_{j=0}^{M-1} \frac{\hat{P}(X_i = x_i|Y = j)}{\sum \hat{P}(X_i = x_i|Y = j)} \log\left(\frac{\hat{P}(X_i = x_i|Y = j)}{\sum \hat{P}(X_i = x_i|Y = j)} \right) \tag{2.13}$$

and in the following algorithm statement, $\hat{H}_i$ is used to substitute for H_i to estimate $q(x_i)$ defined in (2.5).

2.3 Adaptive Sequential Feature Selection Algorithm

A new feature selection and classification scheme is proposed as follows. Suppose Ω_k is a non-empty subset of $\{0,1,\ldots,M-1\}$, representing the subset of patterns, or classes, that is considered in the k^{th} $(k = 1,2,\ldots)$ stage of the algorithm. Let $\hat{P}(X_i = x_i|Y = j)$ denote the kernel density estimation of the conditional p.d.f., $P(X_i = x_i|Y = j)$, $i = 1,\ldots,d$, $\forall j \in \Omega_k$, from the training set, D_n. Let the d-dimensional test sample be $x = [x_1, x_2, \ldots, x_d]^T$. For notational convenience, $|\Omega_k|$ denotes the cardinality of Ω_k, and the elements of Ω_k are ordered in some fashion, so that $\Omega_k(i)$ represents the i^{th} element of Ω_k. The basic structure of the adaptive sequential feature selection algorithm (ASFS) is summarized by the following procedure:

Step 1: Initially, $k = 1$, $\Omega_k = \{0,1,\ldots,M-1\}$ (the labels of all candidate classes).

Step 2: The quality of feature x_i is estimated as

$$\hat{q}(x_i) = \log(|\Omega_k|) - \hat{H}_i$$

where $\hat{H}_i$ is calculated as

$$\hat{H}_i = -\sum_{j\in\Omega_k} \frac{\hat{P}(X_i = x_i|Y = j)}{\sum \hat{P}(X_i = x_i|Y = j)} \log\left(\frac{\hat{P}(X_i = x_i|Y = j)}{\sum \hat{P}(X_i = x_i|Y = j)}\right). \tag{2.14}$$

Step 3: Choose the optimal feature of the k^{th} stage, x_*, as

$$x_* = \underset{x_i\in\{x_1,\ldots,x_d\}}{\arg\max}\ \hat{q}(x_i).$$

Step 4: Threshold the estimated posterior probability distribution, $\hat{P}(Y = j|X_* = x_*)$, $\forall j \in \Omega_k$, by the value T using the following procedure:

$\Omega' = \Phi$ (Φ being a null set)

for $i = 1 : |\Omega_k|$

if $\hat{P}(Y = \Omega_k(i)|X_* = x_*) \geq T$

$\Omega' = \Omega' \bigcup \{\Omega_k(i)\}$

end

end

Step 5:

if $|\Omega'| > 1$

$k = k + 1$

$\Omega_k = \Omega^{'}$

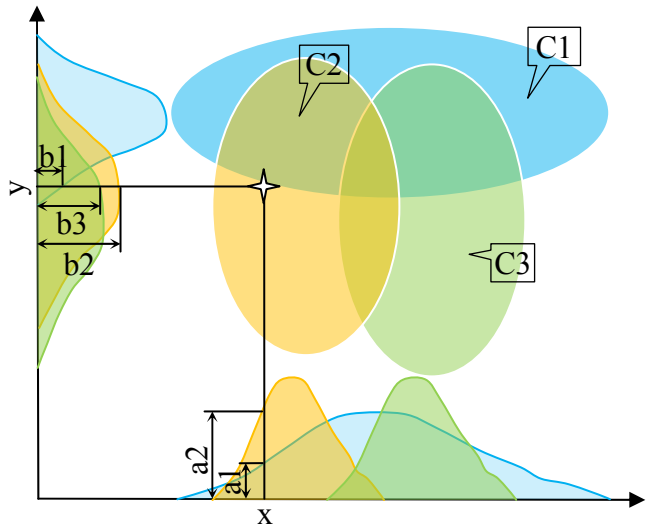
go to step2

else

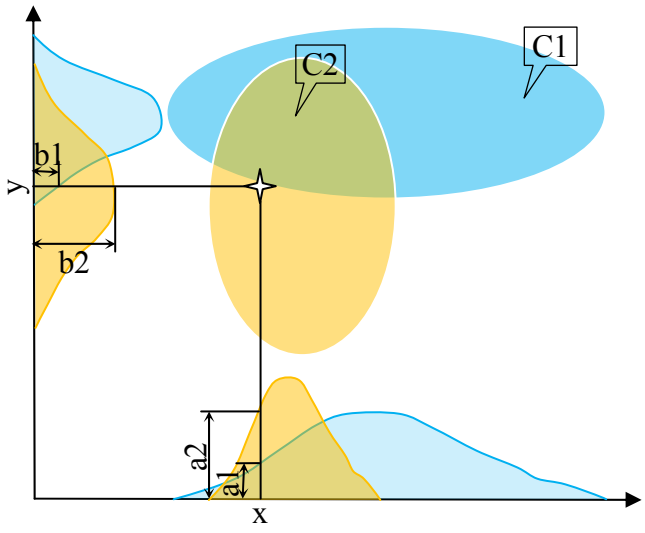
$\Omega^{'}(1)$ is the final retrieved class label.

■

The working process of the ASFS algorithm for a simple synthetic example is illustrated in Figure 2.1 and stated accordingly.



(a)



(b)

Figure 2.1 An illustrative show of the working process of the ASFS algorithm in its stage 1 (a), and stage 2 (b).

In this synthetic example, assume there are three classes (C_1, C_2, C_3), whose corresponding level sets of the conditional p.d.f.s are plotted as blue, orange, and green shaded ovals, respectively, in Figure 2.1(a). Also assume that a test data point belongs to class C_2, and lies at the position with coordinates (x, y), which is indicated by a star symbol. Denote $a_i = P(x|C_i)$ and $b_i = P(y|C_i)$, $i = 1,2,3$ and illustrate $\{a_1, a_2\}$ and $\{b_1, b_2, b_3\}$ in Figure 2.1(a). Please note that $a_3 = 0$ in this example. Without loss of generality, a uniform prior probability distribution is imposed on these three classes.

Stage 1: The class conditional entropy, H_i, is smaller for $\{a_1, a_2, a_3\}$ than for $\{b_1, b_2, b_3\}$, so the feature x is the best feature for stage 1. The posterior probability distribution conditioning on the feature x, $\{P(C_i|x), i = 1,2,3\}$, is thresholded by T, which is set to be the median of $\{P(C_1|x), P(C_2|x), P(C_3|x)\}$, then the class C_3 is discarded.

Stage 2: After stage 1, C_1 and C_2 remain. Figure 2.1(b) illustrates this intermediate result. Now H_i is smaller for $\{b_1, b_2\}$ than for $\{a_1, a_2\}$, so the feature y is the best feature for stage 2. The posterior probability distribution conditioning on the feature y, $\{P(C_i|y), i = 1,2\}$, is thresholded by T, which is set to be the median of $\{P(C_1|y), P(C_2|y)\}$, then C_1 is discarded and C_2 becomes the predicted class label.

2.4 Properties of the ASFS algorithm

It can be proved that the design of adaptive sequential feature selection leads to the following two properties.

Property 2.1: Let (X, Y) be a pair of random variables taking their respective values from $\Re^d$ and $\{0,1\}$, and denote $X = [X_1, X_2, \ldots, X_d]^T$. Denote the one dimensional posterior probability distribution functions as

$$\eta_i(t) = P(Y = 1|X_i = t),\ i = 1, \ldots, d \tag{2.15}$$

and assume they are known. Then the ASFS algorithm achieves an average classification error probability lower than or equal to the Bayes error probability of any single feature.

Proof:

Let $E^Z(\cdot)$ denote the expectation with respect to the distribution of the random variable (vector) Z, and $x=[x_1,x_2,\ldots,x_d]^T$ denote a specific test sample. Then the error probability for classifying x using the feature x_i is

$$\min(\eta_i(x_i),1-\eta_i(x_i)) \tag{2.16}$$

and the Bayes error probability of the feature X_i is

$$E^{X_i}\{\min(\eta_i(X_i),1-\eta_i(X_i))\}=E^X\{\min(\eta_i(X_i),1-\eta_i(X_i))\}. \tag{2.17}$$

By the ASFS algorithm, x_* is selected from $\{x_1,x_2,\ldots,x_d\}$ such that

$$H_*=\min_{i\in\{1,\ldots,d\}}(H_i). \tag{2.18}$$

In the case of binary classification, H_i is a concave function with respect to $\eta_i(x_i)$, so (2.18) implies that

$$\min(\eta_*(x_*),1-\eta_*(x_*))=\min_{i\in\{1,\ldots,d\}}(\min(\eta_i(x_i),1-\eta_i(x_i))). \tag{2.19}$$

Then, the error probability for classifying x by the ASFS algorithm is

$$\min(\eta_*(x_*),1-\eta_*(x_*)) \tag{2.20}$$

and the average classification error probability of the ASFS algorithm is

$$E^X(\min(\eta_*(X_*),1-\eta_*(X_*)))$$

$$=E^X(\min_{i\in\{1,\ldots,d\}}(\min(\eta_i(X_i),1-\eta_i(X_i))))$$

$$\leq E^X(\min(\eta_i(X_i),1-\eta_i(X_i))),\quad \forall i\in\{1,\ldots,d\}. \tag{2.21}$$

Because the right hand side of the inequality (2.21) is the Bayes error probability of the feauture X_i and this inequality holds for every $i\in\{1,\ldots,d\}$, the conclusion is reached.

■

Property 2.2: Let X, Y, $E^Z(\cdot)$, x and $\eta_i(t)$ have the same notation meanings as in Property 2.1, and assume $\eta_i(t)$, $i=1,\ldots,d$, are known. Also define

$$\eta(x)=P(Y=1|X=x) \tag{2.22}$$

and assume the class conditional independence condition holds, i.e.,

$$P(X_1=x_1,\cdots,X_d=x_d|Y=j)=\prod_{i=1}^{d}P(X_i=x_i|Y=j). \tag{2.23}$$

Then the ASFS algorithm achieves the Bayes error probability of X when $d=2$ and the prior probability distribution is uniform, i.e., $P(Y=0)=P(Y=1)=0.5$.

Proof:

Bayes error probability is achieved when the decision function is

$$g(x)=\begin{cases}1 & \eta(x)\geq 1/2\\ 0 & otherwise\end{cases}. \tag{2.24}$$

It suffices to show that the ASFS algorithm leads to the same decision function as (2.24).

When $d=2$, x_* is selected by the ASFS algorithm such that

$$H_*=\min_{i\in\{1,2\}}(H_i). \tag{2.25}$$

In the case of binary classification, H_i is a concave function with respect to $\eta_i(x_i)$, so (2.25) implies that

$$\min(\eta_*(x_*),1-\eta_*(x_*)) = \min_{i\in\{1,2\}} (\min(\eta_i(x_i),1-\eta_i(x_i))) . \tag{2.26}$$

Because of the class conditional independence condition and uniform prior distribution,

$$\frac{\eta(x)}{(1-\eta(x))} = \frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \frac{P(X=x|Y=1)}{P(X=x|Y=0)} = \frac{P(X_1=x_1|Y=1)P(X_2=x_2|Y=1)}{P(X_1=x_1|Y=0)P(X_2=x_2|Y=0)}$$

$$= \frac{P(Y=1|X_1=x_1)P(Y=1|X_2=x_2)}{P(Y=0|X_1=x_1)P(Y=0|X_2=x_2)} = \frac{\eta_1(x_1)\eta_2(x_2)}{(1-\eta_1(x_1))(1-\eta_2(x_2))} . \tag{2.27}$$

From (2.24) and (2.27)

$$g(x)=1 \;\Leftrightarrow\; \eta(x)/(1-\eta(x)) \geq 1 \;\Leftrightarrow\; \eta_1(x_1)\eta_2(x_2) \geq (1-\eta_1(x_1))(1-\eta_2(x_2)) . \tag{2.28}$$

The last inequality in (2.28) holds if and only if

$$\eta_1(x_1) \geq \eta_2(x_2) \geq 0.5 \tag{2.29}$$

$$\text{or } \eta_2(x_2) \geq \eta_1(x_1) \geq 0.5 \tag{2.30}$$

$$\text{or } \eta_1(x_1) \geq (1-\eta_2(x_2)) \geq 0.5 \tag{2.31}$$

$$\text{or } \eta_2(x_2) \geq (1-\eta_1(x_1)) \geq 0.5 . \tag{2.32}$$

From (2.25), the optimal feature, x_*, is x_1 for (2.29) and (2.31) and x_2 for (2.30) and (2.32), and the classification result of the ASFS algorithm is $Y=1$ for all 4 cases. By the same argument, it can be shown that the ASFS algorithm retrieves $Y=0$ whenever $g(x)=0$. Therefore the ASFS algorithm has the same decision function as the Bayes rule (2.24), hence achieving the Bayes error probability.

■

The above provides the proofs of the properties of the ASFS algorithm in the binary classification or low dimensional feature space, while the following gives other important

remarks on the implementation and output of the ASFS algorithm in the general situation, multiple classes and arbitrary dimensional feature space.

(1) The output of the ASFS algorithm provides not only a label of the retrieved class, but also the estimation of the posterior probability distribution and its variance level, which is carried out in a sequential fashion. More concretely, suppose $\Omega_k \subset \{0,1,\ldots,M-1\}$ represents the subset of classes that is considered in the k^{th} stage of the ASFS algorithm, $x_*^{(k)} \in \{x_1,\ldots,x_d\}$ is the optimal feature in the k^{th} stage. Let $\hat{f}_{*j}^{(k)}$ denote the kernel density estimation of $P(X_*^{(k)} = x_*^{(k)} | Y = j)$, $\forall j \in \Omega_k$, and $\hat{p}_{j*}^{(k)}$ denote the estimation of the posterior probability distribution conditioning on Ω_k, i.e.,

$$\hat{p}_{j*}^{(k)} = \hat{P}(Y = j \Big| X_*^{(k)} = x_*^{(k)}, \Omega_k) = \frac{\hat{f}_{*j}^{(k)}}{\sum_{j \in \Omega_k} \hat{f}_{*j}^{(k)}}, \quad \forall j \in \Omega_k . \tag{2.33}$$

Also, $\hat{p}_j^{(k)}$ is used to represent the estimation of the posterior probability distribution after the first k stages, i.e.,

$$\hat{p}_j^{(k)} = \hat{P}(Y = j | S_1,\ldots,S_k), \quad \forall j \in \{0,\ldots M-1\} \tag{2.34}$$

where S_i means the event of the i^{th} stage been carried out, and $\hat{v}_j^{(k)}$ is used to represent the estimation of the variance of $\hat{p}_j^{(k)}$, i.e.,

$$\hat{v}_j^{(k)} = \hat{\text{var}}(\hat{p}_j^{(k)}), \quad \forall j \in \{0,\ldots M-1\}. \tag{2.35}$$

Then the sequential update of $\{\hat{p}_j^{(k)}\}$ is,

$$\hat{p}_j^{(0)} = 1/M\,, \quad \forall j \in \{0,\ldots M-1\} \tag{2.36}$$

$$\hat{p}_j^{(k)} = \hat{p}_j^{(k-1)}\,, \quad k \geq 1 \quad j \notin \Omega_k \tag{2.37}$$

$$\hat{p}_j^{(k)} = \hat{p}_{j*}^{(k)} \sum_{j \in \Omega_k} \hat{p}_{j*}^{(k-1)}\,, \quad k \geq 1 \quad j \in \Omega_k\,. \tag{2.38}$$

From (2.36) through (2.38), the sequential update of $\{\hat{v}_j^{(k)}\}$ is,

$$\hat{v}_j^{(0)} = 0\,, \quad \forall j \in \{0,\ldots M-1\} \tag{2.39}$$

$$\hat{v}_j^{(k)} = \hat{v}_j^{(k-1)}\,, \quad k \geq 1 \quad j \notin \Omega_k \tag{2.40}$$

$$\hat{v}_j^{(k)} \approx \left(\sum_{j \in \Omega_k} \hat{p}_{j*}^{(k-1)} \right)^2 \hat{\text{var}}\left(\hat{p}_{j*}^{(k)}\right) + \left(\hat{p}_{j*}^{(k)}\right)^2 \sum_{j \in \Omega_k} \hat{\text{var}}\left(\hat{p}_{j*}^{(k-1)}\right),\ k \geq 1 \quad j \in \Omega_k\,. \tag{2.41}$$

From (2.33),

$$\hat{p}_{j*}^{(k)} = \frac{\hat{f}_{*j}^{(k)}}{\sum_{j \in \Omega_k} \hat{f}_{*j}^{(k)}}$$

$\hat{\text{var}}\left(\hat{f}_{*j}^{(k)}\right)$ is quantified through Proposition 2.2, so by applying the multivariate Taylor expansion [p153, Rice, 1995] to $\hat{p}_{j*}^{(k)}$, $\hat{\text{var}}\left(\hat{p}_{j*}^{(k)}\right)$ can also be obtained, hence making the sequential update in (2.41) applicable.

(2) Due to the combinatorial aspect of the problem, there is no efficient criterion for setting the threshold value, T, in order to achieve a global optimality, such as the output posterior probability distribution having the smallest entropy. The synthetic simulation and the neural signal decoding practice in this thesis suggest that T can be set by the

greedy rule or the median rule. More concretely, assume Ω_k represents the set of classes that is considered in the k^{th} stage of the ASFS algorithm, x_* is the optimal feature in the k^{th} stage, and $\Omega_{k+1} \subset \Omega_k$ represents the set of classes that remains after thresholding $\{\hat{P}(Y = j | X_* = x_*), j \in \Omega_k\}$ by T. The greedy rule is to set T to be the smallest value thresholded by which $\hat{H}_*$ is not the minimum one of $\{\hat{H}_i, i = 1, \ldots, d\}$, which are calculated over Ω_{k+1}. The median rule is to set T to be the median value of $\{\hat{P}(Y = j | X_* = x_*), j \in \Omega_k\}$, hence only a half number of patterns remain for the successive stage. There is only tiny performance difference for these two setting rules in the application of neural signal decoding. One reason is that the pool of candidate classes is relatively small $(M \leq 8)$, and another reason is that the early termination of the ASFS algorithm, explained in below, often occurs.

(3) In some applications, such as the neural signal decoding researched in Chapter 3, the training sets of individual patterns are highly overlapped in the feature space. So for some choices of the test data, x, no feature can decrease the entropy prominently. In detail, assume x_* is the optimal feature in the k^{th} stage of the ASFS algorithm, then it may happen (often in the neural signal decoding) that $\log(|\Omega_k|) - \hat{H}_*$ is close to zero. This case implies that the estimated conditional probability density values over Ω_k, $\{\hat{P}(X_* = x_* | Y = j), j \in \Omega_k\}$, are nearly uniformly distributed. Because the estimated values, $\{\hat{P}(X_* = x_* | Y = j), j \in \Omega_k\}$, have their associated variance (2.9), the null

hypothesis,

$$H_0 : E\{\hat{P}(X_* = x_* | Y = \Omega_k(1))\} = \cdots = E\{\hat{P}(X_* = x_* | Y = \Omega_k(|\Omega_k|))\} \tag{2.42}$$

where $E\{\cdot\}$, Y, and $|\Omega_k|$ denote the expectation operator, the class label, and the cardinality of Ω_k, respectively, cannot be rejected. Therefore, filtering the posterior probability distribution, $\{\hat{P}(Y = j | X_* = x_*), j \in \Omega_k\}$, by some threshold, T, lacks the statistical significance, To deal with this situation, in the implementation of the ASFS algorithm to the neural signal decoding, a threshold, t, is set so that if

$$\log(|\Omega_k|) - \hat{H}_* < t \tag{2.43}$$

the algorithm will terminate and output the sequentially updated estimation of the posterior probability distribution and its variance level up to the $(k-1)^{th}$ stage. This early termination criterion lets a classifier make the decision for each test data, x, according to its confidence level.

(4) Although ASFS selects only a single feature at each stage, it can be easily generalized to multiple feature selection. For example, when a large size training sample set or a good parametric model for the generating mechanism of the training data is available, estimating two or even higher dimensional p.d.f.s becomes reasonable and reliable.

(5) Better effectiveness of the ASFS algorithm results when the feature set is approximately class conditionally independent (2.23), so the preprocessing steps, such as principal component analysis, independent component analysis, et al., will help improve

the algorithm's performance.

2.5 Simulation

To illustrate the effectiveness of the ASFS algorithm, it is compared with the classical 1-nearest neighbor (1-NN) rule, for which there exits theoretical results on its asymptotic classification performance. The general k-nearest neighbor rule (k-NN) for a binary classification problem, i.e., $M = 2$, is

$$\forall x \in \Re^d \quad g_n(x) = \begin{cases} 1 & \sum_{i=1}^{n} w_{ni} I_{\{Y_i=1\}} > \sum_{i=1}^{n} w_{ni} I_{\{Y_i=0\}} \\ 0 & otherwise \end{cases} \tag{2.44}$$

where I_A is an indicator function of the event A, $w_{ni} = 1/k$ if X_i is among the k nearest neighbors of x, and 0 otherwise. X_i is said to be the k^{th} nearest neighbor of x if the distance measure $\|x - X_i\|$ is the k^{th} smallest among $\|x - X_1\|, \ldots, \|x - X_n\|$. For a training set, D_n, the best one can expect from a decision function, $g : \Re^d \to \{0,1\}$, is to achieve the Bayes error probability, L^*, through the Bayes rule, $g^* : \Re^d \to \{0,1\}$. L^* and g^* are given as:

$$g^*(x) = \begin{cases} 1 & \eta(x) > 1/2 \\ 0 & otherwise \end{cases} \tag{2.45}$$

$$L^* = E\{\min(\eta(X), 1 - \eta(X))\} \tag{2.46}$$

where $\eta(x) = P(Y = 1 | X = x)$.

Generally, it is not possible to obtain a function that exactly achieves the Bayes error, but it is possible to construct a sequence of classification rules $\{g_{D_n}\}$, such that the error probability

$$L_n = P\{g_{D_n}(X) \neq Y | D_n\} \tag{2.47}$$

gets close to L^*. The following theorems summarize some classical results for the k-NN rule.

Proposition 2.3 [Chapter 3 and 5, Devroye et al., 1996]: Let

$$L_{NN} = \lim_{n\to\infty} E\{L_n\} \tag{2.48}$$

be the asymptotic classification error of the 1-NN rule, then for any distribution of the random variables (X,Y), L_{NN} satisfies:

$$L_{NN} = E\{2\eta(X))(1-\eta(X))\} \tag{2.49}$$

$$L^* \leq L_{NN} \leq 2L^*(1-L^*) \leq 2L^*. \tag{2.50}$$

■

Proposition 2.4 [Chapter 5, Devroye et al., 1996]: For any distribution of the random variables, (X,Y), and $k \geq 3$ being odd, the asymptotic classification error probability of k-NN rule satisfies:

$$L_{kNN} \leq L^* + \frac{1}{\sqrt{ke}} \tag{2.51}$$

$$L_{kNN} \leq L^* + \sqrt{\frac{2L_{NN}}{k}} \tag{2.52}$$

$$L_{kNN} \le L^*\left(1+\frac{\gamma}{\sqrt{k}}\left(1+O(k^{-1/6})\right)\right),\ \gamma \approx 0.34. \tag{2.53}$$

■

It is also shown [Chapter 5, Devroye et al., 1996] that when L^* is small, $L_{NN} \approx 2L^*$ and $L_{kNN} \approx L^*$ for $k \ge 3$ odd. But attention must also be paid to another situation where L^* becomes large. In many practical applications, such as neural signal decoding, the problem of pattern recognition is much more challenging than just separating two nearly well separated classes. Additionally, if L^* is very small, many classification techniques will perform very well. The only difference will be their convergence rates. The next chapter will show that the difference between two classification methods (ASFS versus k-NN), in the case of large L^*, can be quite noticeable.

To appreciate the benefits of the proposed approach, firstly consider the following simple synthetic example. Assume in a binary classification task, Y equals 0 or 1 with a uniform prior p.m.f., i.e., $P(Y=0)=P(Y=1)=0.5$. Conditioning on $Y=0$ or 1, $X \sim N(0, I_d)$ or $N(\frac{r1_d}{\sqrt{d}}, I_d)$, respectively. $N(\mu, \Sigma)$ represents a multivariate normal distribution with the mean, μ, and the covariance matrix, Σ. I_d is a $d \times d$ identity matrix, 1_d is a $d \times 1$ vector with all entries being 1, and r is a positive scalar. In other words, this task is to distinguish between two classes whose members are distributed as multivariate Gaussians with the same parameters except the means. Let n be the number of i.i.d. training samples for each class and m be the number of test samples. Figure 2.2 illustrates the empirical correct

classification rates from the ASFS algorithm and the 1-NN rule under different choices of r. In this simulation, $m = 2000$, and the parameter pair, (n,d), takes different values for each subplot.

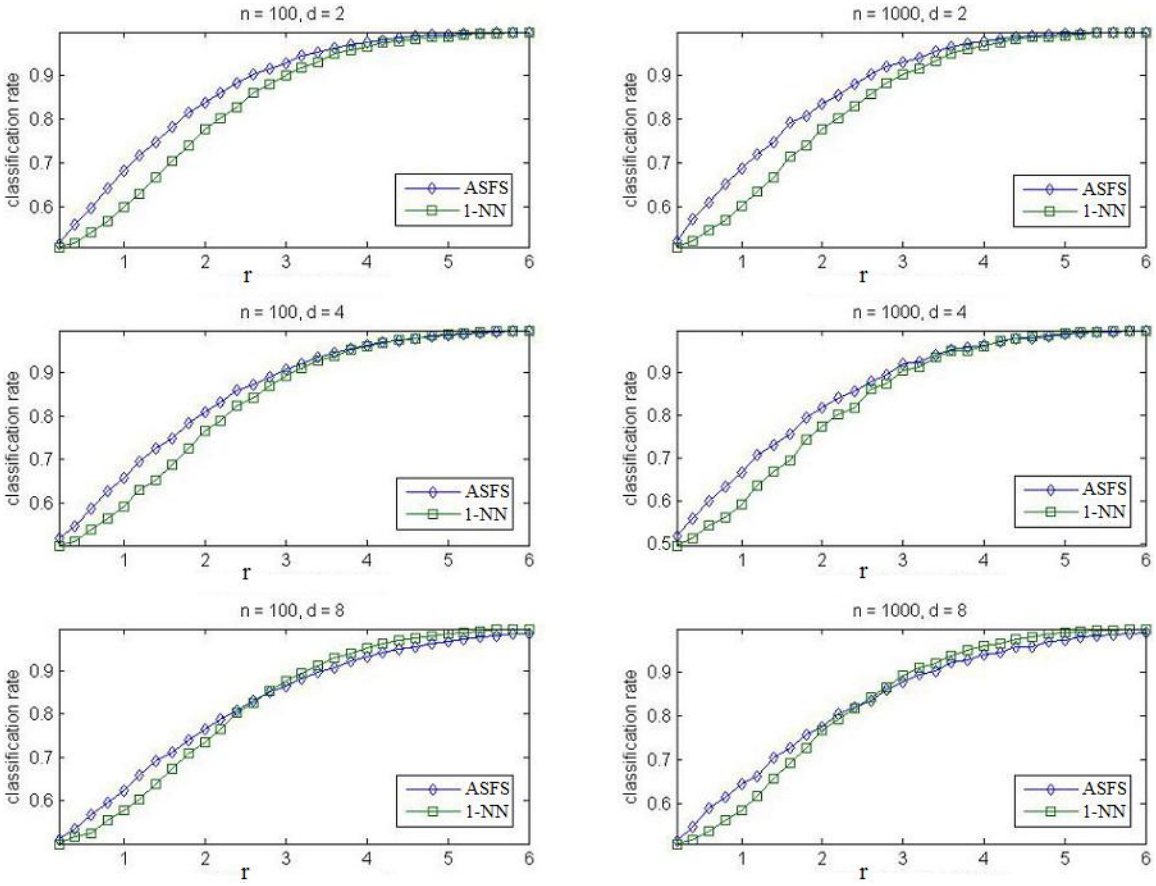


Figure 2.2 An empirical comparison of the ASFS algorithm and the 1-NN rule in a binary synthetic classification task.

Several observations are contained in Figure 2.2:

(1) The ASFS algorithm outperforms the 1-NN rule in all the 6 subplots when L^* is large. This observation is most relevant to Chapter 3 of this thesis, as empirical evidence shows that the neural signal classes are typically not well separated. In general, a fast convergence rate for small L^* does not guarantee the best performance for large L^*.

(2) When the classes are well separated (i.e., L^* is small or r is large), the 1-NN rule shows stable convergence rate, while the ASFS algorithm yields a lower convergence rate as d becomes larger. The stable convergence of 1-NN arises from the fact that when L^* is small, $L_{NN} \approx 2L^*$. Note that the ASFS method only selects one best feature of the test data to carry out the binary classification. When L^* is small, all the features are deemed to be good. So whichever feature is selected, the discrimination power of that feature will be weaker than that of the whole set of features, which is implicitly used in the distance measure of the 1-NN rule. When L^* is large, there is a higher probability that more features are corrupted by noise, thus utilizing all the features in the 1-NN rule will introduce more noise than discrimination information.

CHAPTER 3

NEURAL SIGNAL DECODING

3.1 Background

Neural prosthetic systems aim to translate neural activities from the brains of patients who are deprived of motor abilities but not cognitive functions, into external control signals. Substantial progress towards realization of such systems has been made only recently [Musallam et al., 2004; Santhanam et al., 2006; Shenoy et al., 2003; Schwartz and Moran, 2000; Wessberg et al., 2000; Isaacs et al., 2000; Donoghue, 2002; Nicolelis, 2001, 2002]. The design and construction of such devices involve challenges in diverse disciplines. This chapter concerns how to decode a finite number of classes, the intended "reach directions", from recordings of an electrode array implanted in a subject's brain. Especially, this chapter applies the ASFS algorithm, the k-NN rule, and the support vector machine technique, together with an information fusion rule, to decode neural data recorded from the Posterial Parietal Cortex (PPC) of a rhesus monkey, and compares their performance on the experimental data. While motor areas have mainly been used as a source of command signals for neural prosthetics [Schwartz and Moran, 2000; Nicolelis, 2002], a pre-motor area of PPC called the Parietal Reach Region (PRR) has also been shown to provide useful control signals [Musallam et al., 2004]. It is believed that reaching plans are formed in the PRR preceding an actual reach [Meeker et al.,

2001]. The advantage of using higher-level cognitive brain areas is that they are more anatomically removed from regions that are typically damaged in paralyzed patients. Furthermore, the plasticity of PRR enables the prosthetic user to more readily adapt to the brain-machine interface.

Extracellular signals were recorded from a 96 wire micro-electrode array (MicroWire, Baltimore, Maryland) chronically implanted in the PRR area of a single rhesus monkey. The training and test data sets were obtained as follows. The monkey was trained to perform a center-out reaching task (see Figure 3.1). Each daily experimental session consisted of hundreds of trials, which are categorized into either the reach segment or the brain control segment. Each session started with a manual reach segment, during which the monkey performed about 30 memory guided reaches per reach direction. While fixating on a central lit green target, this task required the subject to reach to a flashed visual cue (consisting of a small lit circle in the subject's field of view) after a random delay of 1.2 to 1.8 seconds (the memory period). After a "go" signal (consisting of a change in the intensity of the central green target) the monkey physically reached for the location of the memorized target. Correct reaches to the flashed target location were rewarded with juice. The brain control segment began similarly to the reach trials, but the monkey wasn't allowed to move its limbs, only the monkey's movement intention was decoded from signals derived from the memory period neural data. A cursor was placed at the decoded reach location and the monkey was rewarded when the decoded and previously flashed target locations coincided. Electrode signals were recorded under two

conditions: one having 4 equally spaced reach directions (rightward, downward, leftward, upward), and the other having 8 (previous four plus northeastward, southeastward, southwestward, northwestward). Let P_4 denote the experimental data set recorded under the first condition, and P_8 the second. Both data sets include not only reach trials but also brain control trials.

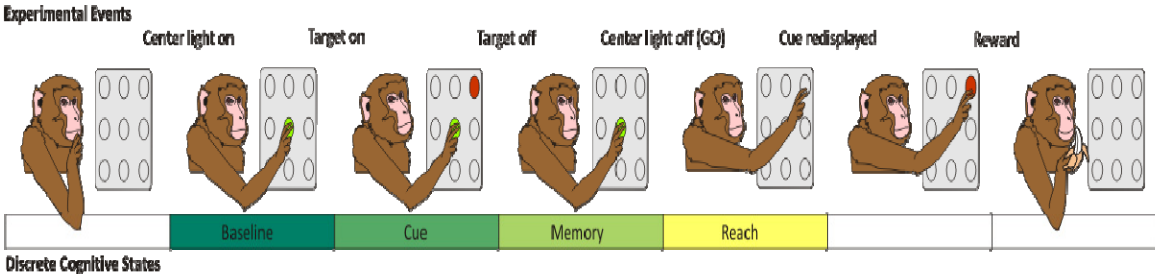


Figure 3.1 One experimental procedure of the center-out reach task for a rhesus monkey.

3.2 Neural Signal Decoding

To ensure that only the monkey's intentions were analyzed and decoded and not signals related to motor or visual events, only the memory period activities were used in this analysis. More precisely, assume the beginning of memory period in each trial marks an alignment origin, i.e., $t = 0$, then the recorded neural data in one trial takes a form of binary sequence $T = (\ldots, T_{-2}, T_{-1}, T_0, T_1, T_2, \ldots)$:

$$T_k = \begin{cases} 1 & \exists\, spike\, in \quad (k\Delta t, (k+1)\Delta t] \\ 0 & otherwise \end{cases}, \ \Delta t = 1 \text{ ms}. \tag{3.1}$$

A spiking data sub-sequence was extracted from the time interval $200 \sim 1200$ ms after

the cue for P_4, and similarly from the interval $100 \sim 1100$ ms for P_8. For the analysis given below, the spiking data was then binned into 4 subsegments of 250 ms duration each. The number of spikes within each subsegment was recorded as one entry of the vector, $S = [S_1, S_2, S_3, S_4]^T$. Furthermore, the binned data vector S was preprocessed by a multi-scale Haar wavelet transformation [Mallat, 1999], because the optimal bin width is still unknown and by the wavelet transformation, both short-term features and long-term features are generated. Moreover, the simple structure of the Haar functions give the wavelet coefficients intuitive biological interpretations [Cao, 2003], such as firing rates, bursting, and firing rate gradients. In detail, let W be the Haar wavelet transformation matrix and $X \in \Re^4$ be the vector of wavelet coefficients for S, then

$$X = WS = \begin{bmatrix} (S_2 - S_1)/\sqrt{2} \\ (S_4 - S_3)/\sqrt{2} \\ (S_3 + S_4 - S_1 - S_2)/2 \\ (S_1 + S_2 + S_3 + S_4)/2 \end{bmatrix}. \tag{3.2}$$

The vector, X, for each neuron serves as the input to the different algorithms that are implemented and compared in this chapter. Figure 3.2 shows the estimated p.d.f.s of 4 wavelet coefficients with the four different target directions ($D118$ - rightward, $D120$ - downward, $D122$ - leftward, $D124$ - upward) associated with P_4. Each subplot shows the p.d.f.s of one wavelet coefficient conditioned on four target directions. Note that the conditional p.d.f.s from different classes have very significant overlaps.

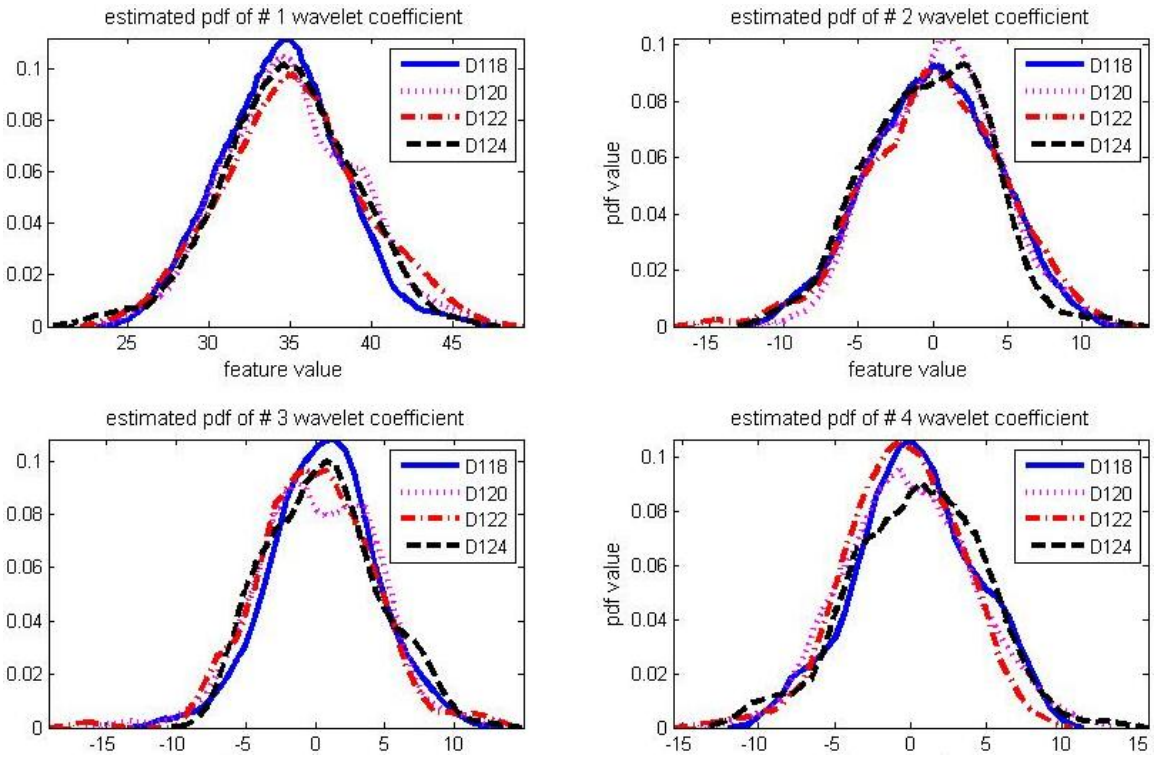


Figure 3.2 Estimated wavelet coefficients p.d.f.s conditioned on different directions from one typical neuron in P_4.

Although each neuron is a very weak classifier, one example being shown in Figure 3.2, a much better overall performance can be achieved by assembling the information of all neurons. There are two choices. One choice is input fusion, which is to concatenate the data from each neuron into an augmented vector. On the one hand, the Bayes error is a decreasing function with dimension of feature space [p29, Devroye et al., 1996]. On the other hand, as analyzed in [p315, Fukunaga, 1990], the bias between asymptotic and finite sample 1-NN classification error correlates with sample size and dimensionality of the feature space. Generally speaking, the bias increases as the dimensionality goes higher, and the bias drops off slowly as the sample size increases, particularly when the

dimensionality of the data is high. So when only a reasonably finite data set, say, 100 training samples per class, is available, it is possible that the bias increment will overwhelm the benefit of the decrement of L_{NN} (2.48) in a relatively high dimensional feature space. This phenomenon matches the results observed while applying the k-NN method to neural signal decoding.

Another more useful choice is output fusion, which is to let the decision results of individual classifiers vote. Unlike input fusion, output fusion is a very economical way to exploit the capabilities of multiple classifiers. For a good survey reference, please check into [Miller and Yan, 1999]. The specific output fusion methods implemented in neural signal decoding of this chapter are the product rule and the summation rule, whose justifications [Theodoridis and Koutroumbas, 2006] are described in the following paragraphs.

In a classification task of M classes, assume one is given R classifiers. For a test data sample, $x \in \Re^d$, each classifier produces its own estimate of the a posteriori probabilities, i.e., $\hat{P}_r(Y = j|X = x)$, $j = 0,\ldots,M-1$, $r = 1,\ldots,R$. The goal is to devise a method to yield an improved estimate of a final a posteriori probability $\hat{P}(Y = j|X = x)$ based on all the individual classifier estimates. Based on the Kullback-Leibler (KL for abbreviation) probability distance measure, one can choose $\hat{P}(Y = j|X = x)$ in order to minimize the average KL distance, i.e.,

$$\begin{aligned} &\min \quad D_{av} = \frac{1}{R}\sum_{r=1}^{R} D_r \\ &s.t. \quad \sum_{j=0}^{M-1} \hat{P}_r(Y=j|X=x) = 1 \quad \forall r = 1,...,R \end{aligned} \tag{3.3}$$

where D_r is a discrete KL distance measure

$$D_r = \sum_{j=0}^{M-1} \hat{P}(Y=j|X=x) \log \frac{\hat{P}(Y=j|X=x)}{\hat{P}_r(Y=j|X=x)}. \tag{3.4}$$

By utilizing Lagrange multipliers, the optimal probability distribution to solve (3.3) is obtained as,

$$\hat{P}(Y=j|X=x) = \frac{1}{C}\left(\prod_{r=1}^{R} \hat{P}_r(Y=j|X=x)\right)^{1/R} \tag{3.5}$$

where C is a class independent constant quantity. So the rule becomes equivalent to assigning the unknown feature vector x to the class maximizing the product, the so called the product rule, i.e.,

$$g(x) = \underset{j\in\{0,...,M-1\}}{\arg\max} \prod_{r=1}^{R} \hat{P}_r(Y=j|X=x). \tag{3.6}$$

The KL measure is not symmetric. If an alternative KL distance measure

$$D'_r = \sum_{j=0}^{M-1} \hat{P}_r(Y=j|X=x) \log \frac{\hat{P}_r(Y=j|X=x)}{\hat{P}(Y=j|X=x)} \tag{3.7}$$

is taken, then, minimizing $D'_{av} = \frac{1}{R}\sum_{r=1}^{R} D'_r$ subject to the same constraints in (3.3) leads to assigning the unlabeled test data, x, to the class that maximizes the summation, the so called the summation rule, i.e.,

$$g(x) = \underset{j\in\{0,\ldots,M-1\}}{\arg\max} \sum_{r=1}^{R} \hat{P}_r(Y = j \mid X = x) . \tag{3.8}$$

Note that the product rule and summation rule require that the estimates of the a posteriori probabilities from each classifier be independent, otherwise voting becomes biased. Fortunately, in the neural decoding application, the independence assumption is well approximated due to the significant distance ($500\,\mu m$) between adjacent recording electrodes relative to the minute neuronal size. Moreover, because each neuron calculates its output based on its own input, the product rule and the summation rule take another equivalent form. More concretely, assume $X_{(i)}$ represents the input feature vector of the i^{th} neuron, $i = 1,\ldots,R$, and X_c is the concatenation vector of all $X_{(i)}$, i.e., $X_c = [X_{(1)},\ldots,X_{(R)}]$, then

$$P_r(Y = j \mid X_c = x_c) = P_r(Y = j \mid X_{(r)} = x_{(r)}) , \quad \forall r = 1,\ldots,R . \tag{3.9}$$

The product rule becomes

$$g(x) = \underset{j\in\{0,\ldots,M-1\}}{\arg\max} \prod_{r=1}^{R} \hat{P}_r(Y = j \mid X_{(r)} = x_{(r)}) . \tag{3.10}$$

The summation rule becomes

$$g(x) = \underset{j\in\{0,\ldots,M-1\}}{\arg\max} \sum_{r=1}^{R} \hat{P}_r(Y = j \mid X_{(r)} = x_{(r)}) . \tag{3.11}$$

The probability, $\hat{P}_r(Y = j \mid X_{(r)} = x_{(r)}), j = 0,\ldots,M-1$, can be viewed as an adaptive critic for neuron r under the test data x. This critic evaluates how confidently a given

neuron can classify a specific input signal. In effect, the product rule (3.10) or the summation rule (3.11) allows neurons that generate more non-uniform posterior probability estimates to dominate the final probabilistic classification result.

3.3 Application Results

Below, Figures 3.3 and 3.4 show performance comparisons between the k-NN rules (for $k = 1,5,9$) and the ASFS classification method when applied to the P_4 (584 trials) and P_8 (1034 trials) data sets. The percentage of classification error is used as a metric for comparison of these neural decoding methods. In this comparison, the percent classification error, L_n, was estimated from the data set, D_n, by its leave-one-out estimator, $L_n^{(D)}$. Because $EL_n^{(D)} = EL_{n-1}$ [p407, Devroye et al., 1996], $L_n^{(D)}$ is a good estimator of L_n for large n. Each curve in Figures 3.3 and 3.4 represents this estimated decoding rate as a function of the number of utilized neurons, which are randomly chosen from the full set of available neurons. For each marked point on the curves of Figures 3.3 and 3.4, the estimated correct decoding rate comes from the average of 15 random samplings.

For the k-NN rules, both input fusion and output fusion have been used. Specifically, because the product rule cannot be applied to the output of the k-NN methods, the output fusion method implemented with the k-NN classifiers is the summation rule, i.e., the pattern receiving the maximum number of votes is chosen as the final decision. Figures

3.3 and 3.4 show that the combination of the ASFS algorithm and the output fusion method (the product rule specifically, the summation rule yields only slightly worse results) outperforms the combination of the k-NN rules and the input/output fusion methods in these data sets. Although the performance of the k-NN classifier also increases with k, it saturates quickly for large k. Please notice that the k-NN classification rule demonstrates a slow rate of performance increase with respect to the number of neurons utilized in the case of input fusion. This is indeed the phenomenon explained in [p315, Fukunaga, 1990]: with fixed number of training samples, the increment of bias gradually dominates the decrement of L_{kNN} (2.48) as the dimensionality of the feature vector goes higher.

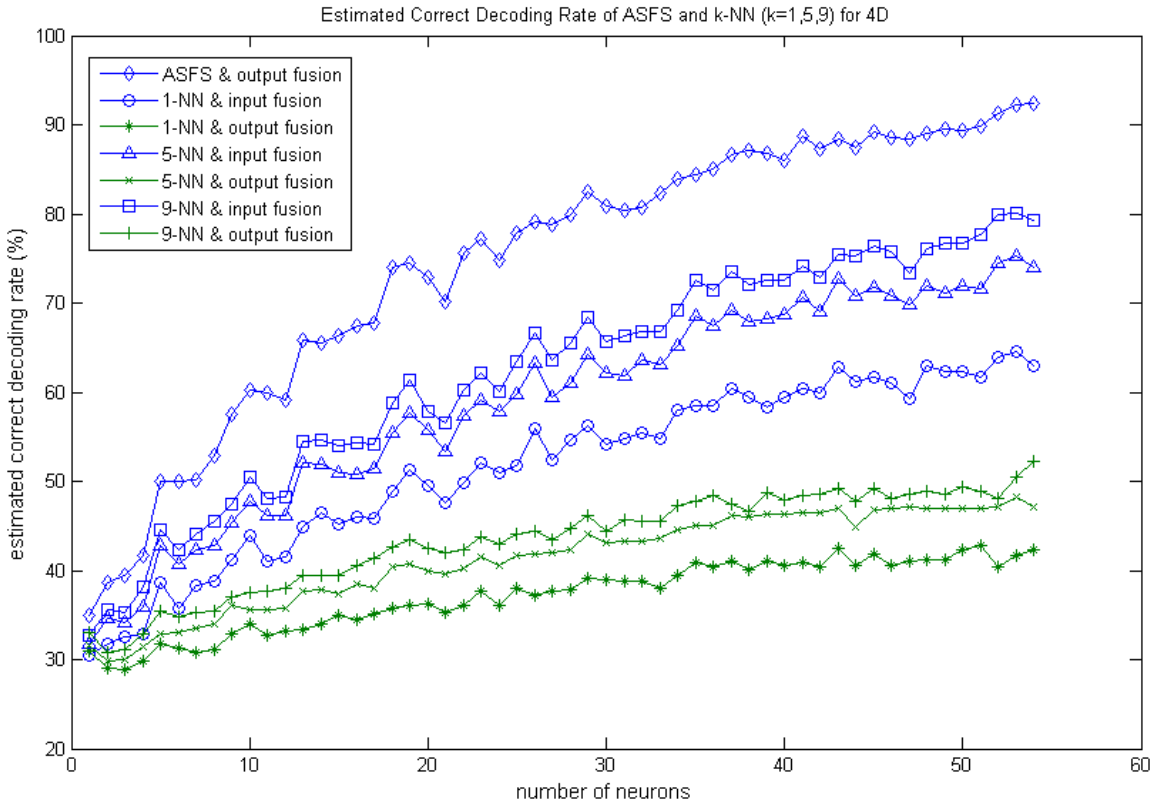


Figure 3.3 Experimental comparison of percent correct decoding rates of ASFS and k-NN ($k = 1,5,9$), together with input/output fusion methods, for P_4.

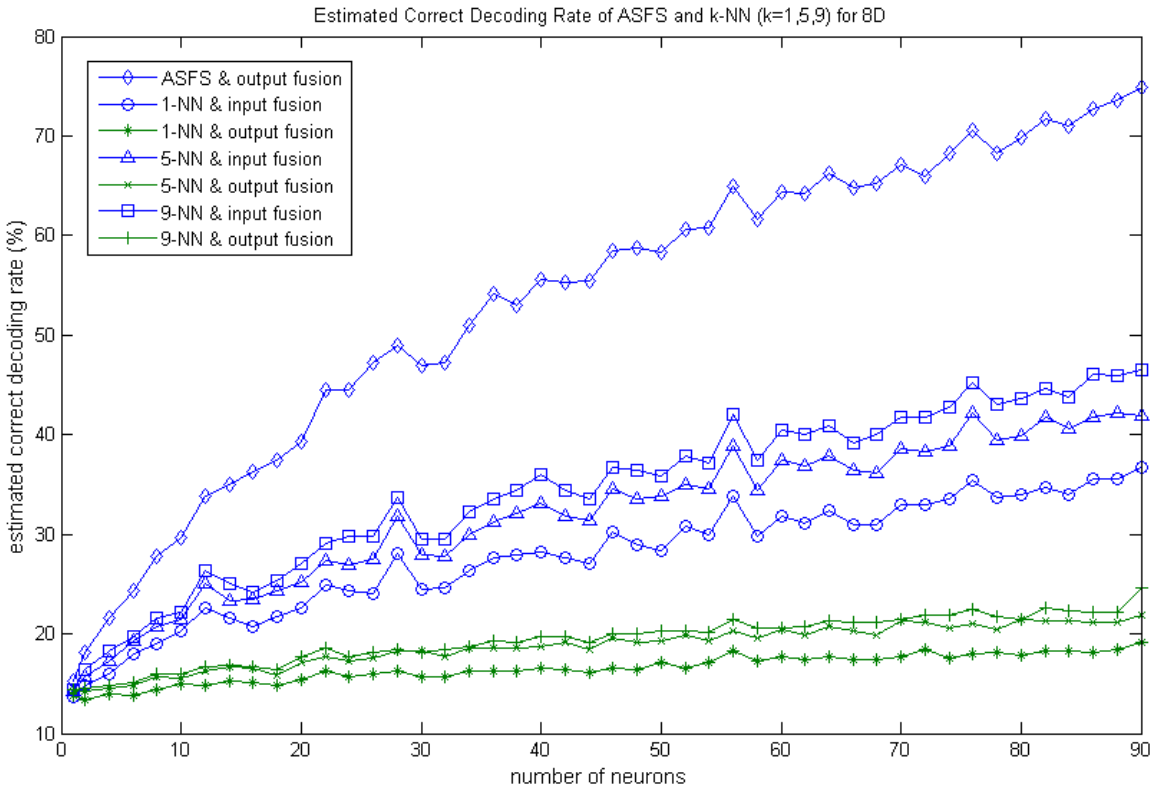


Figure 3.4 Experimental comparison of percent correct decoding rates of ASFS and k-NN ($k = 1,5,9$), together with input/output fusion methods, for P_8.

Next, another comparison of the ASFS algorithm with a popular classification method, support vector machine (SVM), is carried out. To implement the SVM classifier on the neural data sets, one SVM toolbox, LIBSVM, developed by Lin et al. [Chang and Lin, 2001] was used. LIBSVM is an integrated software for classification, regression, and distribution estimation. The classification methods supported by LIBSVM include C-

SVC and nu-SVC, and the former one was selected for these studies. More concretely, [Hsu et al., 2007] is a practical guidance provided by Lin's group to explain how to implement C-SVC to yield good performance, including data scaling, the use of an RBF kernel, parameter selection by cross-validation accuracy and grid search, etc. The implementation of LIBSVM (C-SVC especially) in these studies follows this practical guidance. Figures 3.5 and 3.6 show the comparison results between the ASFS algorithm and the output of the C-SVC classifier in LIBSVM. Each curve in Figures 3.5 and 3.6 represents the estimated percent correct decoding rate by 6-fold cross validation as a function of the number of utilized neurons, which are randomly chosen from the full set of available neurons. The leave-one-out estimation method was not used for this study because its use with the SVM classifier is computationally expensive. Each marked point on the curves of Figures 3.5 and 3.6 represents the mean correct decoding rate of 15 random samplings. A special characteristic of the C-SVC classifier in LIBSVM is that it can not only predict the class label of each test data, but also estimate the posterior probability of that test data belonging to each class. The estimate of the posterior probability distribution provides higher resolution information than a prediction of the class label only, therefore the output fusion based on the posterior probability estimate is superior to the output fusion based on the predicted label. Also, as mentioned in [Miller and Yan, 1999], and as is consistent with the experimental findings in these studies, the product rule usually yields a little better performance than the summation rule. So again the combination of the ASFS algorithm and the product rule is compared with the combination of the C-SVC classifier and the product rule. Figures 3.5 and 3.6 show that although the C-SVC classifier yields slightly better average performance when only a few

neurons are available, the combination of the ASFS algorithm and the product rule quickly and significantly exceeds the combination of the C-SVC classifier and the product rule when an increasing number of neurons are utilized.

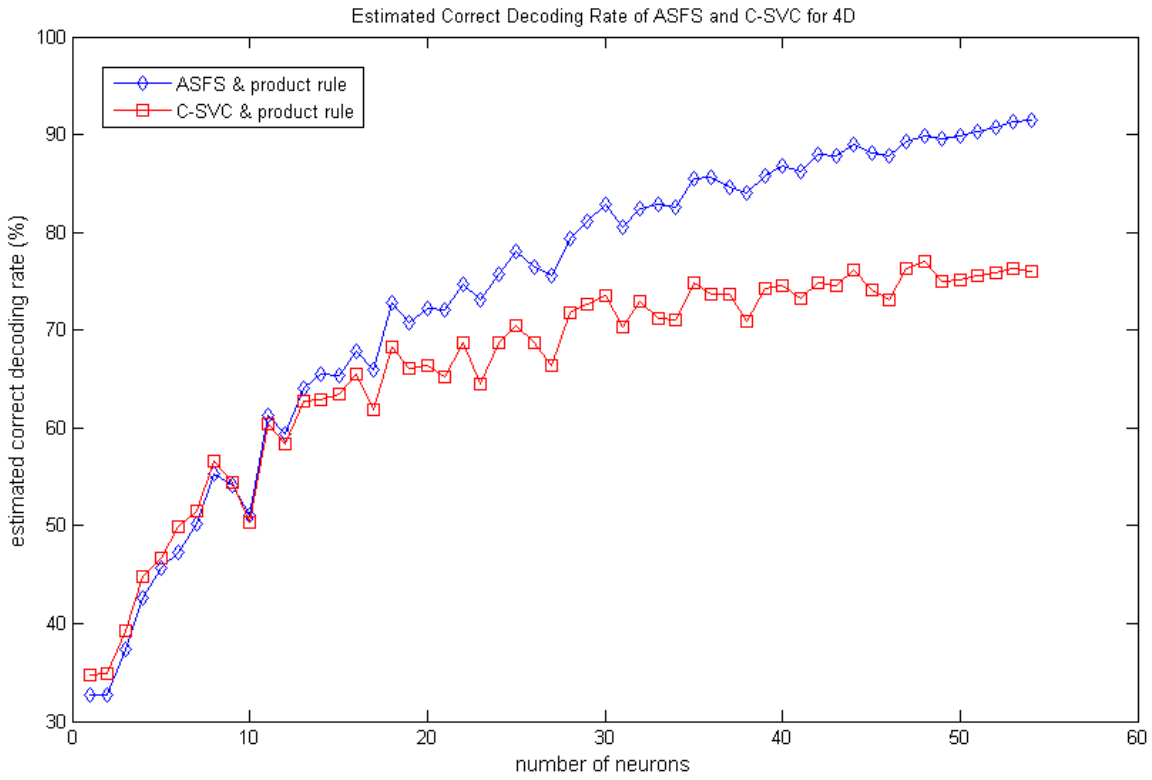


Figure 3.5 Experimental comparison of correct decoding rates of ASFS and C-SVC, together with the product rule, for P_4.

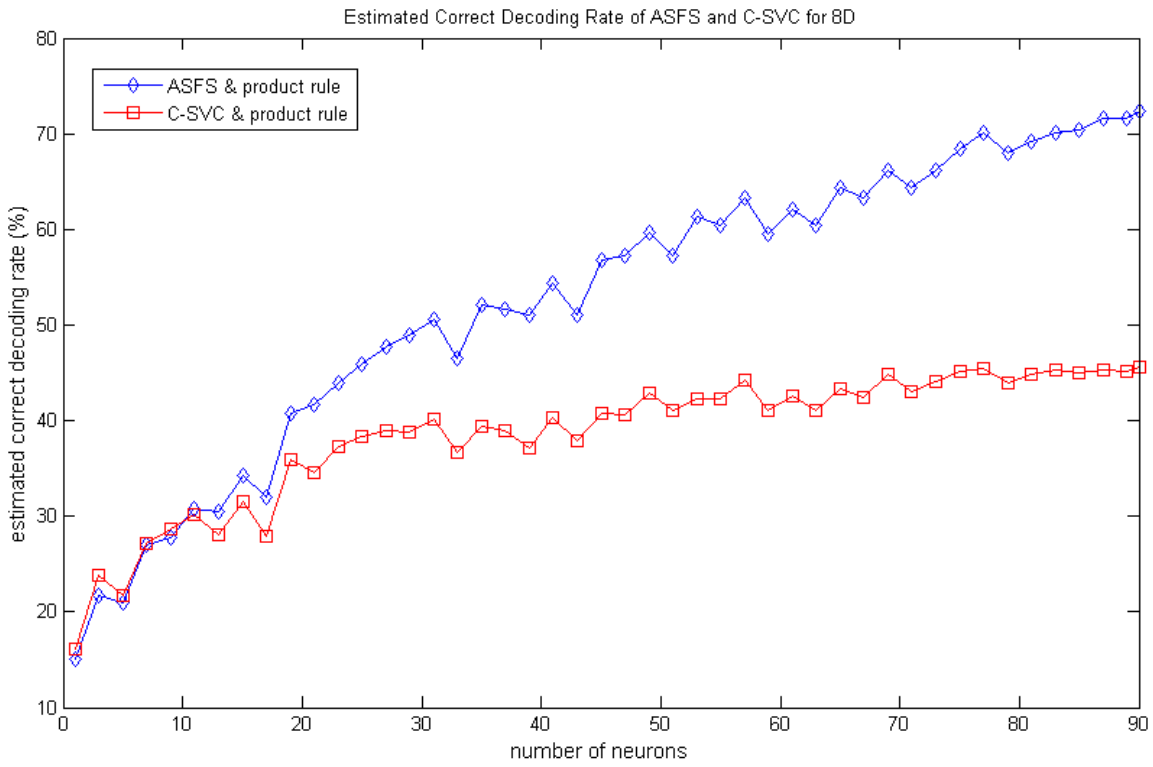


Figure 3.6 Experimental comparison of correct decoding rates of ASFS and C-SVC, together with the product rule, for P_8.

CHAPTER 4

FEATURE SELECTION IN ULTRA-WIDEBAND RADAR SIGNAL ANALYSIS

4.1 Research Motivation

As the explosive growth in the number of vehicles worldwide (800 million vehicles in global use), a large number of road accidents happen every year (1.2 million death a year. Among the fatal accidents, 65% of deaths involve pedestrians and 35% of pedestrian deaths are children) [Peden et al., 2004]. The issue of how to boost the vulnerable road users (VRU) safety has become critical for the automobile industry. The motivation behind this research topic is to augment VRU safety by developing a robust human presence detector and a human behavior (walking, jogging, standing, etc.) classifier through the use of car loaded sensors. Ultra-wideband (UWB) impulse radar, which is characterized by its high range resolution, the ability of remote sensing, and the advantage of penetration into/around obstacles, emerges as one of the most promising sensors to accomplish this task. Although computer vision techniques can assist pedestrian detection, radar based techniques have some distinct advantages, such as detection power beyond stadia distance in foggy conditions or darkness, as well as the ability to see through obstacles or see around corners in advance. To provide the

background information for the research topic in this chapter, UWB systems are briefly reviewed below.

4.2 Ultra-wideband Radar System Overview

The majority of traditional radio systems use a narrowband of signal frequencies modulating a sinusoidal carrier signal. The resonant properties of this system allow an easy frequency selection of necessary signals. Narrowband signals limit the information capability of radio systems, because the amount of the information transmitted in a unit of time is governed by Shannon's equation (Ghavami, Michael, and Kohno, 2004),

$$C = B\log(1+\frac{S}{N}) \tag{4.1}$$

where C is the maximum channel capacity in bits/sec, B is the bandwidth in Hz, and S, N are power in watts of signal and noise, respectively. The equation (4.1) also illustrates that the most efficient way to increase information capability of a radio system is through the use of ultra-wideband radiation. Ultra-wideband radiation transmits signals with -10 dB bandwidths that are at least 20% of its central frequency [Federal Communications Commission, 2002]. Indeed, with the advancement of UWB radar system development and the release of UWB application regulations by the Federal Communications Commission (FCC) in 2002, UWB technology has been a focus in many applications in consumer electronics and communications [Shen et al., 2006]. On the one hand, the ideal characteristics for UWB systems are abundant information packing, precise positioning, and extremely low interference. On the other hand, in spite of recent experimental work,

there is no satisfactory and systematized theory of UWB signal analysis available. The reason is that the process of signal transformation under the context of ultra-wide bandwidth is much more complex than its narrowband counterpart. Hence the well-known narrowband target recognition technique by the Doppler shift effect [Richards, 2005] doesn't apply to UWB systems. Novel methods must be developed for the concerned applications.

UWB radar signals are spread across a very wide range of frequencies. The typical continuous sinusoidal radio wave is replaced by a train of short pulses at millions of pulses per second. For the UWB impulse radar used in this research, Time Domain PulsON 210, Figure 4.1 shows the transmitted monocycle pulse waveform measured in the anechoic chamber of the USC UltRa laboratory together with its Fourier spectrum, and Figure 4.2 is a photograph of the device. When the transmit antenna propagates the generated pulse train, the receiver antenna measures power reflected off targets by scan sampling the return reflections over multiple pulses. Each pulse is sampled once at a particular delay time. Thousands of pulses are needed to construct a graph of reflected waveform with respect to sampling delay time. This graph can also be understood as reflected waveform with respect to range, because sampling delay time is strictly proportional to distance between target scattering points and the radar. Figure 4.3 is one graph of reflected waveform versus range. This graph contains all the target information gathered by the radar for a short time interval. In the remainder of this thesis, this kind of graph is also termed "one radar scan", "one radar waveform", "one target image", or "one target range profile".

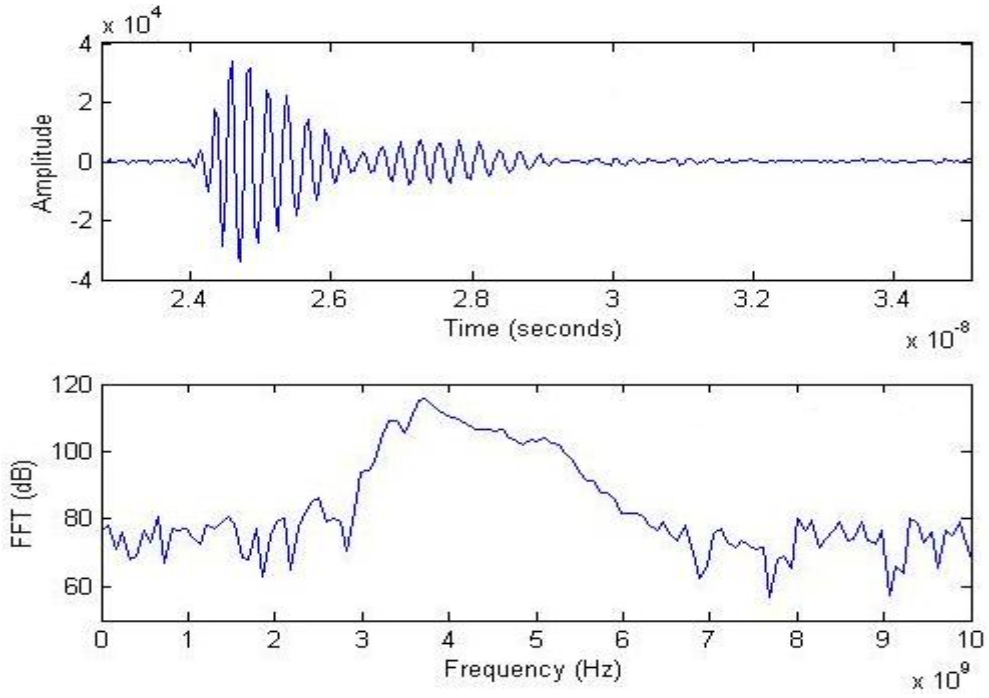


Figure 4.1 Transmitted monocycle pulse waveform of Time Domain PulsON 210 UWB radar and its Fourier spectrum.

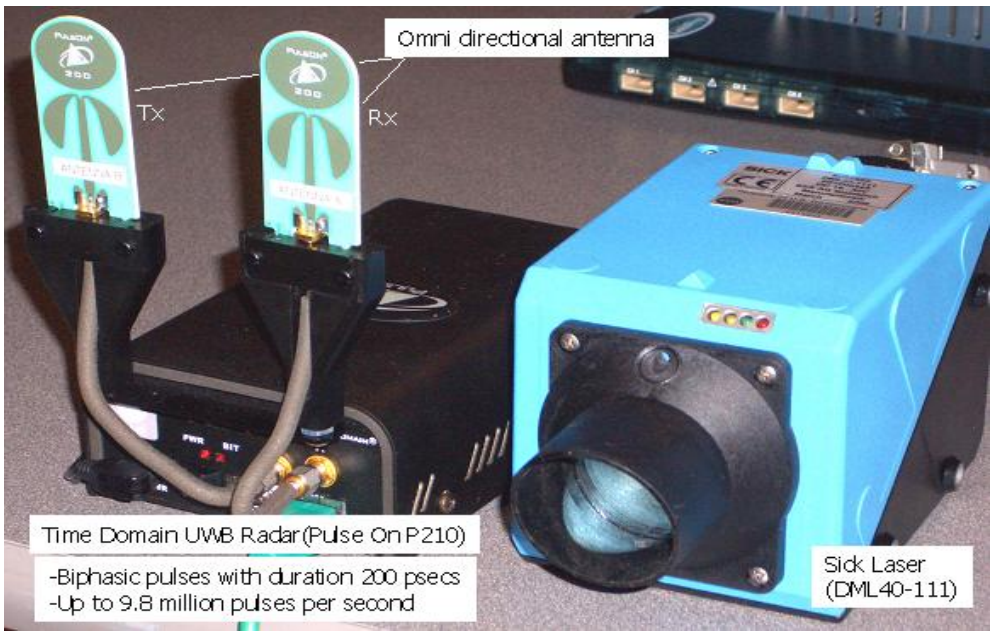


Figure 4.2 Time Domain PulsON 210 UWB radar.

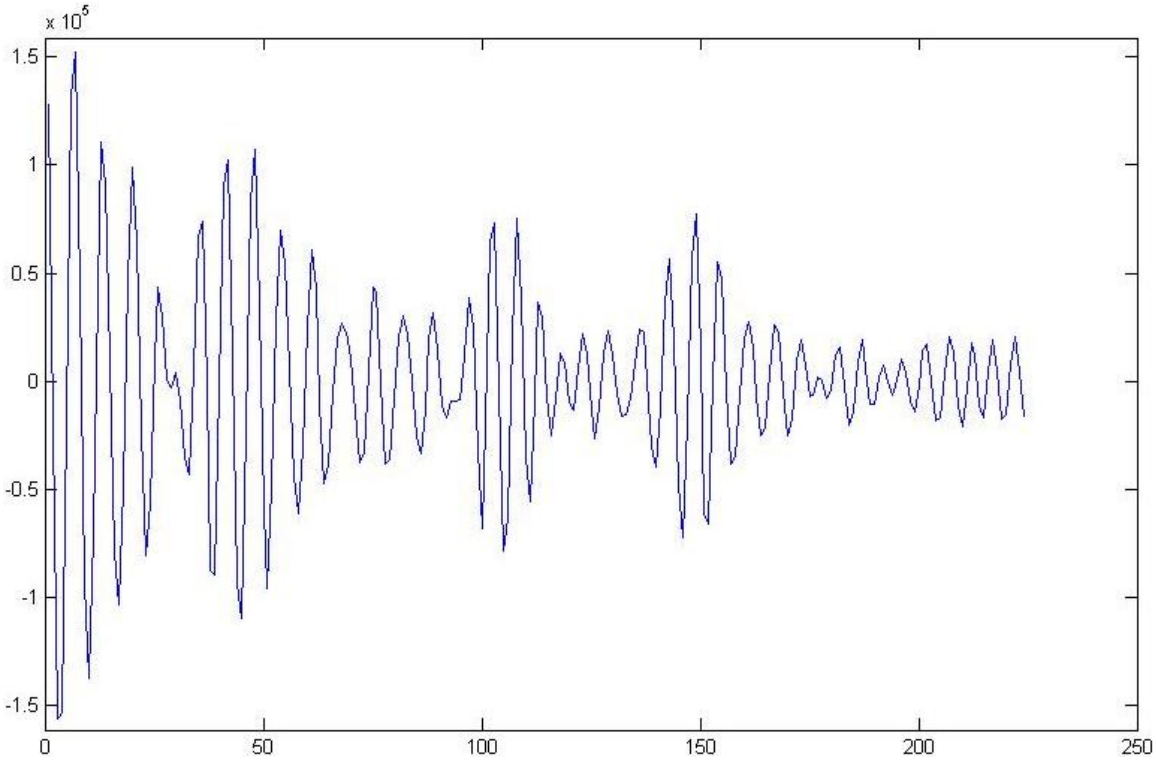


Figure 4.3 One radar scan waveform from Time Domain PulsON 210.

One of the most complicated matters that make UWB signal analysis challenging is the highly varied reflected waveform shapes obtained when the target is not stationary. Consider a simplified radar signal reflection from a local scattering element, the reflected pulse waveform can be determined as the convolution of the waveform of the impulse response characteristic of this local element, $h(x)$, with the function, $f(x)$, describing the incident signal. One geometrically complex target, say the human body, consists of multiple simple local scattering elements. When the target is not stationary, for example when the target person is walking, on the one hand, $h(x)$'s change due to changes of aspects of scattering elements; on the other hand, because of changes of aspects, the

reflected waveform will represent the superposition of multiple reflected pulses with different time delay orders. These two factors, together with others including but not limited to multiple reflections between scattering elements, make the radar waveform shape highly sensitive to the target configuration. In reality, it is observed that the UWB radar yields highly different range profiles even when the target person takes small motions like twisting or tilting the body a little.

4.3 Problem Statements and Characteristics

To prominently distinguish people from other targets and classify people's behaviors, an automatic pedestrian detection system should incorporate as many informative clues as possible. Several prominent features serve this goal, and can be categorized as static features and dynamic features. The static features usually reflect the information of target geometry and its variation structure, while the dynamic features extract the temporal structure among a sequence of radar scans, such as how the target range profiles evolve over time. Fusion of static and dynamic body biometrics will augment the performance of automatic human identification. This chapter researches how to extract a compact set of static features of the target to unravel the dominant statistical properties of the target images. Moreover, the projection of sequential target images onto the subspace spanned by the selected features accentuates the prominent target motion patterns, such as the gait, and therefore provides a sound platform to explore the target dynamics. Although a feature selection problem is explored here, it accounts for a different situation from that in

Chapters 2 and 3. Those chapters designed an adaptive feature selection algorithm based on a class separability criterion, while feature selection in the current problem mainly comes along with cluster representation, which is to approximate the prominent information inside a set of data by as few as possible feature components. More concretely, each target image can be geometrically transformed into one point in the high dimensional Euclidean space, then for a collection of target images, they correspond to a cluster of points that assume a complex high dimensional structure. But due to the redundancy in the random vector, the main variation structures of the data cluster will reside in a much lower dimensional subspace. Then the main task of this feature selection problem is to generate the representative template for a set of target images, locate the high information packing subspace, and explore their statistical and algebraic properties. The selected template or subspace should be adaptive to the gathered data, because the radar range profiles and radar signal dynamics are highly distinct for different target geometry, orientation, or motion patterns. There are several critical issues concerning this feature selection problem, which are addressed in the following structure. Section 4.4 provides preprocessing steps for the raw radar data, which augment the resulting algorithmic performance greatly; Section 4.5 generates a representative template for the target through Procrustes shape analysis, and discusses the statistical classification issue based on the derived template; finally, Section 4.6 implements a classic projection pursuit method to derive the principal components of the data variation structure in the tangent space, and shows that the principal components are also promising clues for target identification.

4.4 UWB Radar Signal Preprocessing

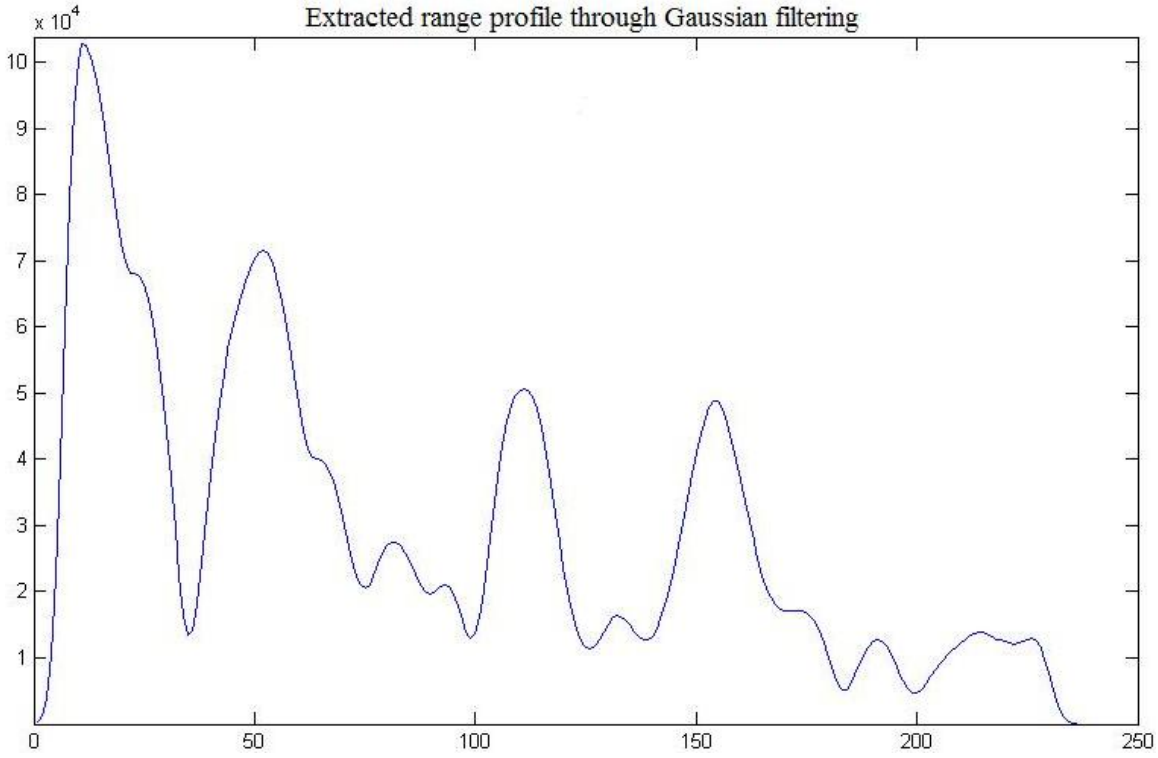
Without preprocessing, one scan of radar data from Time Domain PulsON 210 UWB radar is shown in Figure 4.3. The typical radar scan waveform has an amplitude modulated fast oscillation pattern, which motivates the approximation

$$r(x) \approx a(x)\sin(\omega x + \theta). \tag{4.2}$$

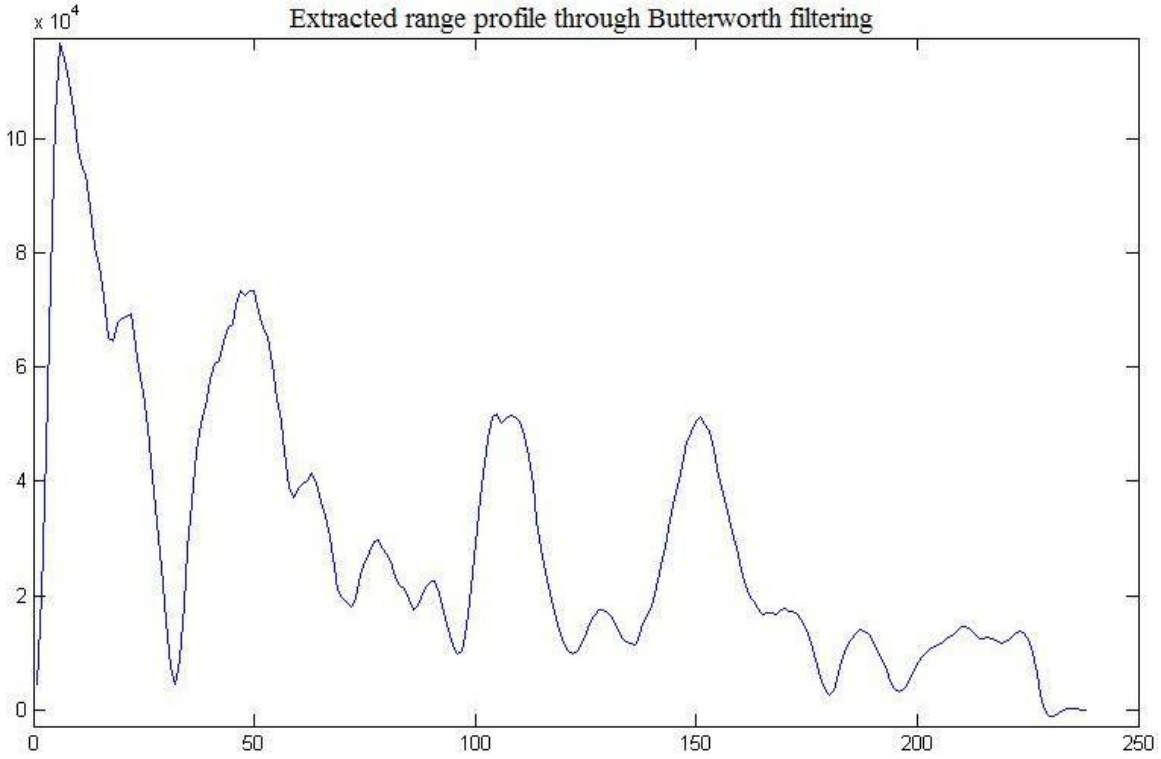
In (4.2), $r(x)$ is the radar scan waveform function with respect to the range, $\sin(\omega x + \theta)$ provides the high oscillation kernel with the phase of θ, and $a(x)$ modulates the amplitude of the oscillation kernel. When the target is not stationary, both the changes in $a(x)$ and θ affect the waveform shape of $r(x)$. In practice, phase change is not only hard to detect, but also has limited identifiability, because $\theta, \theta + 2\pi, \cdots$ are not differentiable based on the observed function value. So the detectable and differentiable information of $r(x)$ is the amplitude part, $a(x)$. Moreover, the fact that the shape of $a(x)$ directly relates to the target reflection geometry makes it an ideal source to generate prominent features for classification. The goal of this first preprocessing step is to separate the range profile, $a(x)$, from $r(x)$. The experimental results show that this preprocessing step augments the resulting algorithmic performance noticeably.

The range profile of $r(x)$ is a smooth function and concentrates the energy on lower frequency components than the high frequency oscillation kernel, so the method of low-pass filtering can be implemented to extract the range profile from $r(x)$. Two specific low-

pass filters, the Gaussian filter and the Butterworth filter, are implemented. Both of them yield similar resulting performance. Figure 4.4 plots the convolution result between the absolute value of the raw radar data and a Gaussian filter (a), and the convolution result between the absolute value of the raw radar data and a Butterworth filter (b). For more complete treatments of digital filter design techniques, please refer to [Oppenheim et al., 1999].



(a)



(b)

Figure 4.4 The convolution result between the absolute value of one raw scan data and a Gaussian filter (a), a Butterworth filter (b).

A second preprocessing step is variable transformation [p76, Fukunaga, 1990], which is applicable for the positive random variable, X, whose distribution can be approximated by a gamma density. In this case, it is advantageous to convert the distribution to a normal-like one by applying the power transformation, i.e.,

$$Y = X^{v}, \quad (0 < v < 1). \tag{4.3}$$

Define

$$\gamma = \frac{E\{[Z - E(Z)]^4\}}{E^2\{[Z - E(Z)]^2\}} \tag{4.4}$$

where Z denotes a random variable, and $E\{\cdot\}$ the expectation operator. It can be obtained that if Z is a Gaussian random variable, then $\gamma = 3$. So the normal-like of Y is approximately achieved by selecting a value of v, such that

$$\gamma' = \frac{E\{[Y - E(Y)]^4\}}{E^2\{[Y - E(Y)]^2\}} \approx 3. \tag{4.5}$$

X is a gamma random variable, so

$$E\{Y^m\} = \frac{\alpha^{\beta+1}}{\Gamma(\beta+1)} \int_0^\infty x^{vm} x^\beta e^{-\alpha x} dx = \frac{\Gamma(\beta+1+mv)}{\alpha^{mv}\Gamma(\beta+1)} \tag{4.6}$$

where $\Gamma(x)$ is the gamma function, defined as

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du, \quad x > 0. \tag{4.7}$$

From (4.6), it can be derived that the value of γ' in (4.5) is independent of the choice of α, therefore the value of γ' is a function of β and v only. In practice, the selection, $v = 0.4$, is often suggested because γ' is close to 3 for a wide range of β when $v = 0.4$. This power transformation, $Y = X^{0.4}$, is applied to the extracted radar range profiles, and makes the data distribution much closer to that of a multivariate normal density.

The last preprocessing implemented in this study is segmentation, which is to extract the part of the radar waveform that corresponds to the location of human presence. In the following sections in this chapter, it is assumed that the preprocessing of low-pass filtering, power transformation, and segmentation has been carried out for the raw radar data.

4.5 Procrustes Shape Analysis and Tangent Space Inference

4.5.1 Procrustes Shape Analysis

In a 2D scenario, shape is very commonly used to refer to the appearance or silhouette of an object. Following the definition in [Dryen and Mardia, 1998], shape is all the geometrical information that remains when location, scale and rotational effects are filtered out from an object. Important aspects of Procrustes shape analysis are to obtain a measure of distance between shapes, and to estimate the average shape and shape variability from a random sample, which should be independent with respect to translation, scaling and rotation of the geometric objects. Translation and rotation invariance don't have a natural correspondence in 1D radar range profiles, because the range of the radar signal corresponding to the target presence can be segmented, and no general movements of targets are assumed to cause a circular shift of the radar waveform. But scaling invariance is left as an important factor because the same target reflection geometry on mildly different distances can yield the waveforms that are close in their shapes but different in their amplitudes. So the classic 2D Procrustes shape analysis has a reduced 1D version for radar signal analysis. For more complete treatments on the Procrustes shape analysis, please refer to [Chapters 3 and 4, Dryen and Mardia, 1998]. In the application aspect, [Boyd, 2001] and [Wang et al., 2004] successfully applied Procrustes shape analysis into computer vision based gait recognition. The following paragraphs briefly review the

definitions of full Procrustes fit, full Procrustes distance, and full Procrustes mean shape, and state their special counterparts in the context of radar range profiles.

Assume two shapes or silhouettes in the 2D space are represented by two vectors of k complex entries, say $y = [y_1, \ldots, y_k]^T$ and $w = [w_1, \ldots, w_k]^T$. Without loss of generality, assume these two configurations are centered, i.e., $y^H 1_k = w^H 1_k = 0$, where y^H means transpose of complex conjugate of y and 1_k is a length-k vector with all components being 1.

Definition 4.1 [p40, Dryen and Mardia, 1998]: The **full Procrustes fit** of w onto y is

$$w^P = (a + ib)1_k + \beta e^{i\theta} w \tag{4.8}$$

where (a, b, β, θ) are chosen to minimize

$$D^2(y, w) = \left\| y - (a + ib)1_k - \beta e^{i\theta} w \right\|^2 . \tag{4.9}$$

Proposition 4.1 [p40, Dryen and Mardia, 1998; Appendix B]: The full Procrustes fit has matching parameters

$$a + ib = 0\,, \quad \theta = \arg(w^H y)\,, \quad \beta = (w^H y y^H w)^{1/2} / (w^H w)\,. \tag{4.10}$$

So the full Procrustes fit of w onto y is

$$w^P = \frac{(w^H y y^H w)^{1/2}}{w^H w} e^{i \angle w^H y} w = \frac{\left| w^H y \right| e^{i \angle w^H y}}{w^H w} w = \frac{w^H y w}{w^H w}\,. \tag{4.11}$$

■

Note that the distance measure, $D(y,w)$, in full Procrustes fit is not symmetric in y and w unless $y^H y = w^H w$. Then a convenient standardization, $y^H y = w^H w = 1$, leads to the definition of full Procrustes distance.

Definition 4.2 [p41, Dryen and Mardia, 1998]: The **full Procrustes distance** between w and y is

$$d_F(w,y) = \inf_{\beta,\theta,a,b} \left\| \frac{y}{\|y\|} - \frac{w}{\|w\|}\beta e^{i\theta} - a - ib \right\| = \left\{ 1 - \frac{y^H w w^H y}{w^H w y^H y} \right\}^{1/2} \tag{4.12}$$

where the second equation comes from complex linear regression used in deriving Proposition 4.1, and it can be checked that $d_F(w,y)$ is invariant with respect to translation, rotation and scaling of configurations of w and y.

Definition 4.3 [p44, Dryen and Mardia, 1998]: The **full Procrustes mean shape** $[\hat{\mu}]$ is obtained by minimizing the sum of squared full Procrustes distances from each configuration w_i to an unkown unit size configuration μ, i.e.,

$$[\hat{\mu}] = \arg\inf_{\mu} \sum_{i=1}^{n} d_F^2(w_i, \mu). \tag{4.13}$$

Note that $[\hat{\mu}]$ is not a single configuration, instead, it is a set, whose elements have zero full Procrustes distance to the optimal unit size configuration, μ.

Proposition 4.2 [p45, Dryen and Mardia, 1998; Appendix B]: The full Procrustes mean shape, $[\hat{\mu}]$, can be found as the eigenvector, $\hat{\mu}$, corresponding to the largest eigenvalue of the complex sum of squares and products matrix

$$S = \sum_{i=1}^{n} w_i w_i^H / (w_i^H w_i). \tag{4.14}$$

All translation, scaling, and rotation of $\hat{\mu}$ are also solutions, but they all correspond to the same shape $[\hat{\mu}]$, i.e., have zero full Procrustes distance to $\hat{\mu}$.

■

Then the full Procrustes fits of $w_1, \ldots, w_n$ onto $\hat{\mu}$ is calculated from Proposition 4.1 as,

$$w_i^P = \frac{w_i^H \hat{\mu} w_i}{w_i^H w_i}, \quad i = 1, \ldots, n. \tag{4.15}$$

A convenient fact is that calculation of the full Procrustes mean shape can also be obtained by taking the arithmetic mean of the full Procrustes fits, i.e., $d_F(\frac{1}{n}\sum_{i=1}^{n} w_i^P, \hat{\mu}) = 0$ [p89, Dryen and Mardia, 1998; Appendix B].

Procrustes shape analysis provides a measure, full Procrustes distance, that quantifies the similarity of two planar configurations, and which is invariant with respect to translation, scaling, and rotation. Procrustes shape analysis also provides an elegant way to define the average shape, the full Procrustes mean shape, which can be viewed as a representative template of the target or class. All the forementioned full Procrustes concepts have their special counterparts in radar range profile context. More concretely, assume two

preprocessed target images are termed $y \in \Re_+^m$ and $w \in \Re_+^m$, then from Proposition 4.1, the full Procrustes fit of w onto y is

$$w^P = \frac{(w^T y y^T w)^{1/2}}{w^T w} w = \frac{w^T y w}{w^T w}. \tag{4.16}$$

The full Procrustes distance between w and y is

$$d_F(w, y) = \inf_{\beta} \left\| \frac{y}{\|y\|} - \frac{w}{\|w\|} \beta \right\| = \left(1 - \frac{(y^T w)^2}{\|w\|^2 \|y\|^2} \right)^{1/2}. \tag{4.17}$$

This distance is quite useful in quantifying the similarity between target image shapes, especially when they come from measurements at different target distances. Consider a sequence of preprocessed target images $\{a_1, \ldots, a_n\}$ with each $a_i \in \Re_+^m$, and a matrix $A = [a_1, \ldots, a_n]$, then one definition of the static template for A can be the full Procrustes mean shape, i.e., to find $u \in \Re^m$, such that it minimizes

$$\sum_{i=1}^{n} d_F^2(u, a_i) = \sum_{i=1}^{n} \left(1 - \frac{u^T a_i a_i^T u}{u^T u a_i^T a_i} \right). \tag{4.18}$$

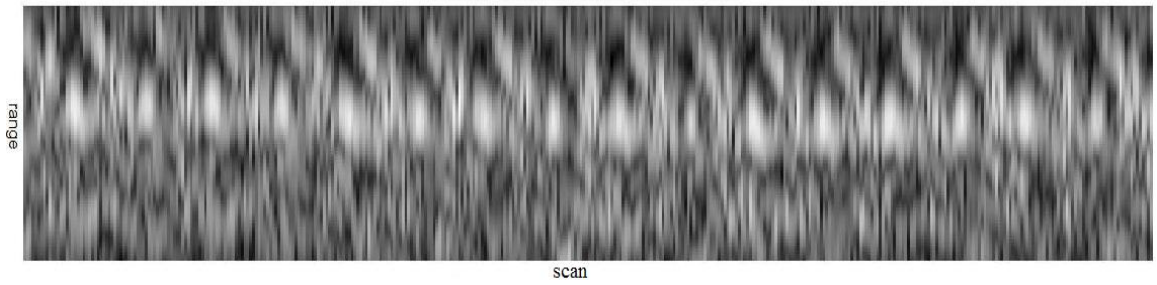
Proposition 4.2 shows that the optimal solution, u^*, turns out to be the eigenvector of $\sum_{i=1}^{n} \frac{a_i a_i^T}{a_i^T a_i}$ corresponding to the maximum eigenvalue. One notable fact is that if columns of A are normalized, i.e., $a_i^T a_i = 1$, then u^* is the first left singular vector of the matrix A. For the special treatement on singular value decomposition, please refer to [p291~p294,

Shores, 2006; p273~p276, Theodoridis and Koutroumbas, 2006; p331~p337, Strang, 2006; Chapter 5, Berrar, Dubitzky and Granzow, 2003].

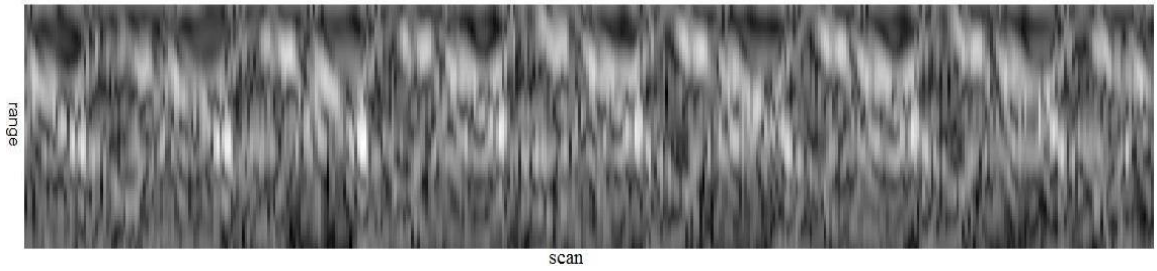
Apply the Procrustes shape analysis to the radar samples, the discrimination ability of the full Procrustes mean shape can be explored. Firstly, Table 4.1 lists the experimental configurations for gathering four data samples that were used to test the proposed algorithms. Figure 4.5 (a) ~ (d) provide the image show of these four data samples, which will also be referred to as sample I, II, III, and IV, respectively. Each sample consists of a sequence of 400 preprocessed radar scans, with each scan stored as a real vector of length 81 and shown as one column in the corresponding figure. To match with the numerical values of the data, the brighter one image pixel is, the higher numeric value it represents.

Scanning rate	42 scans/sec
Target identity	Hao Jiang
Target behavior	Walking in a treadmill
Target distance	1.5 ~ 1.8 m
Target orientation	I. Facing to radar; II. 45 degree to radar; III. Side to radar; IV. Back to radar;

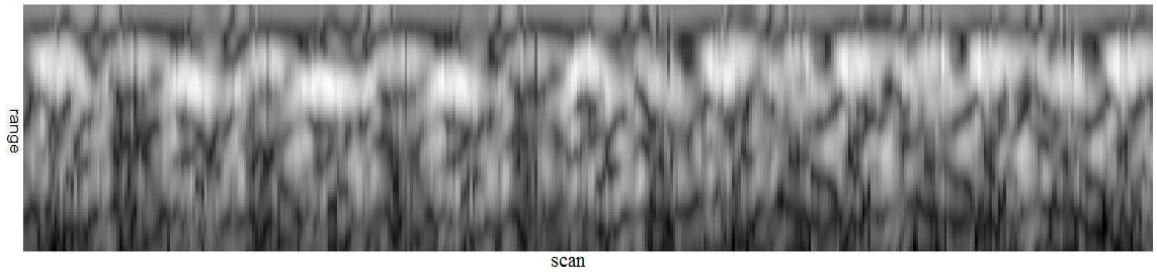
Table 4.1 Experimental configurations for gathering four sample data sets.



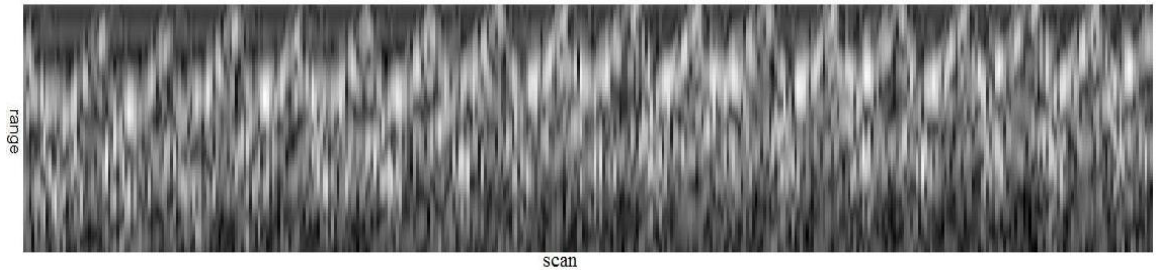
(a)



(b)



(c)



(d)

Figure 4.5 The image show of target images gathered from walking orientation I (a), walking orientation II (b), walking orientation III (c), walking orientation IV (d).

Based on the four sample data sets illustrated above, Figure 4.6 shows the calculated full Procrustes mean shape for each of them.

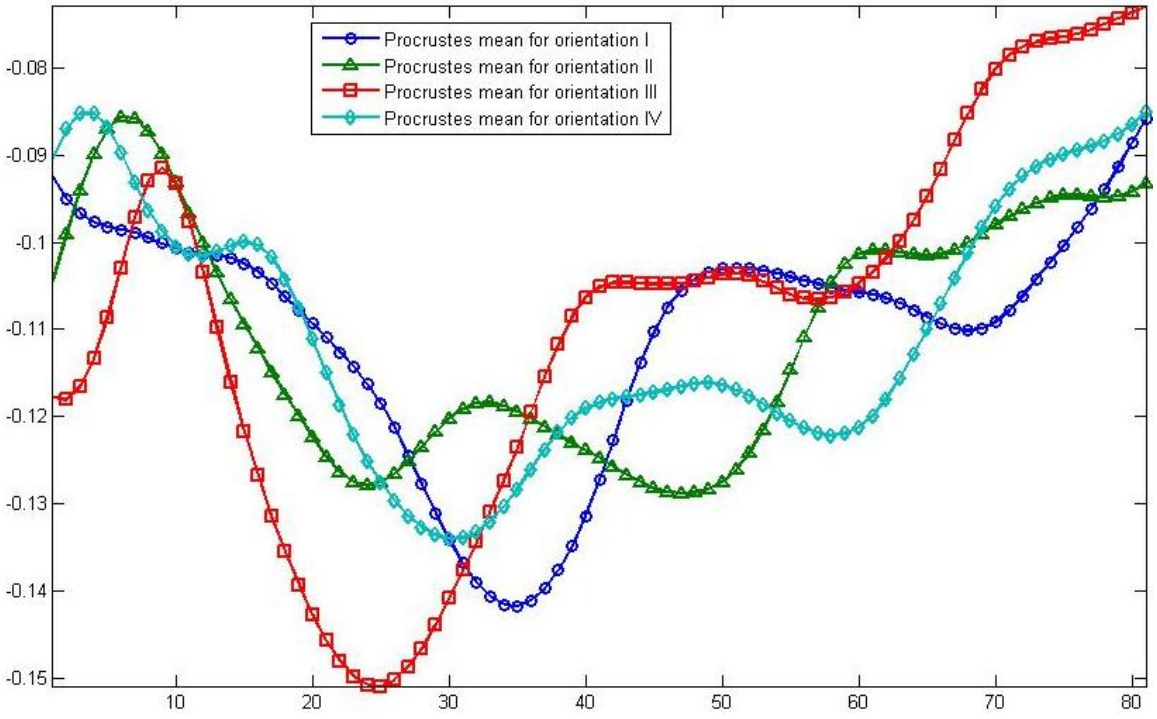
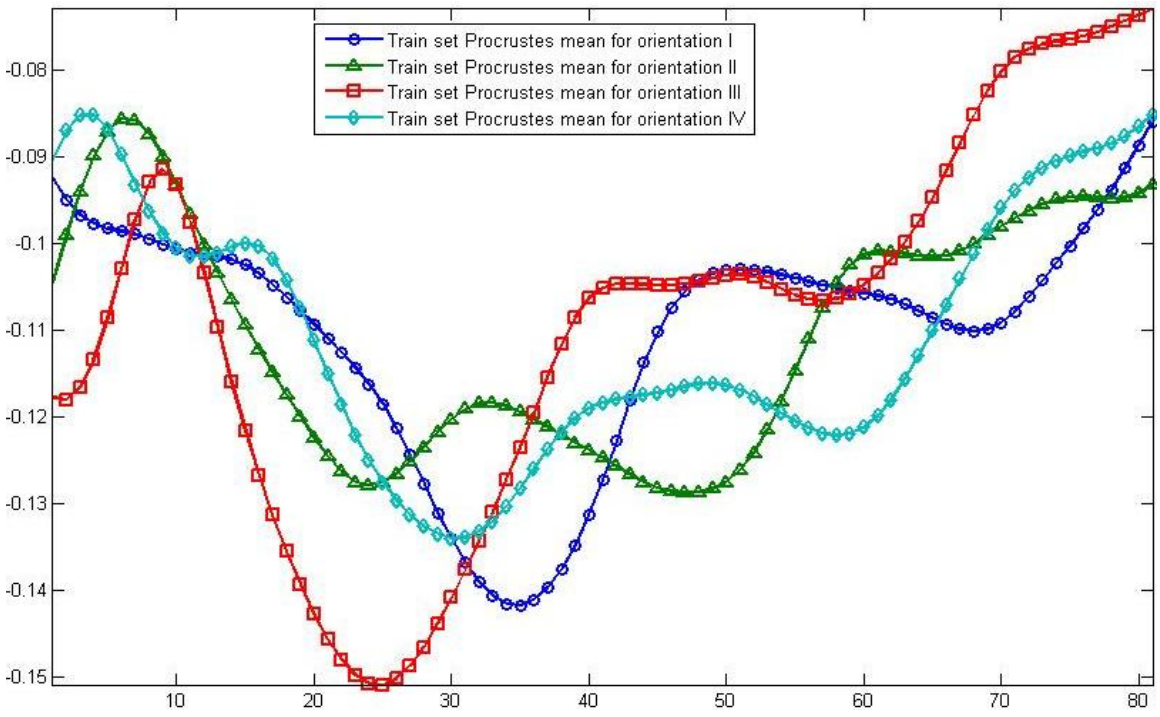


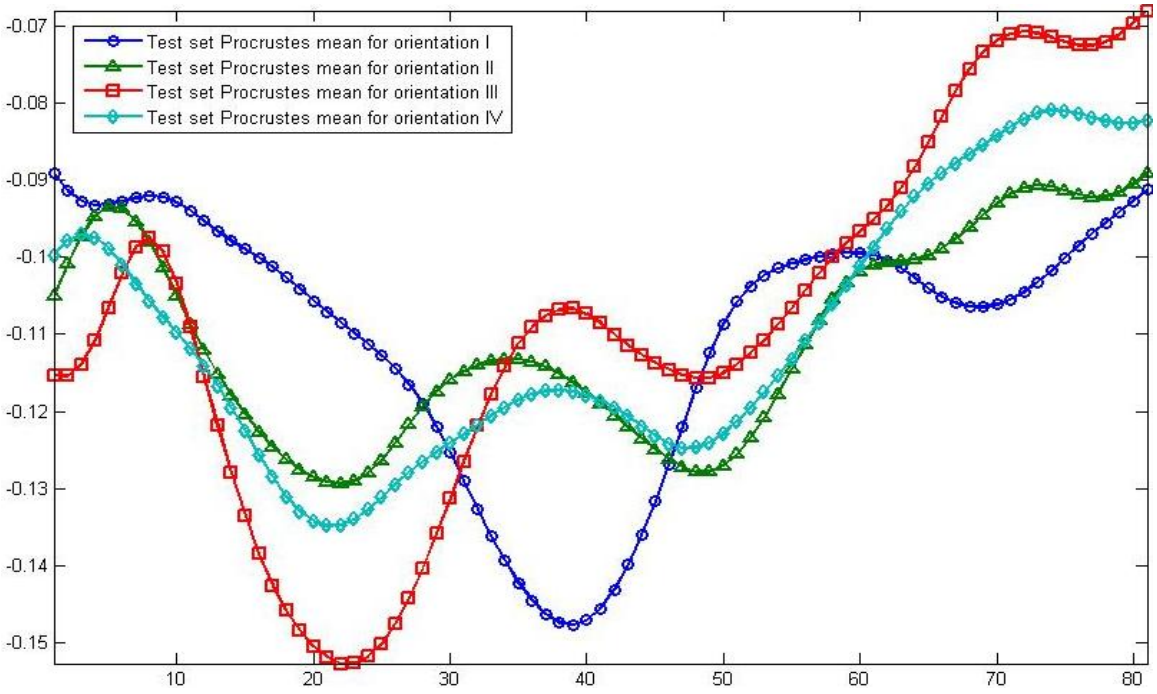
Figure 4.6 Four full Procrustes mean shapes corresponding to the four sample data I ~ IV.

From the experimental data samples, it can be observed that the full Procrustes mean shapes are distinct. Can the Procrustes mean shape be utilized as an informative clue for classification? This question is addressed first by an empirical test, then is generalized to a statistical inference problem which is addressed by the Hotelling one sample T^2 test under the assumption of a multivariate normal model.

To explore the discrimination power of the full Procrustes mean shape, each data sample is divided into two non-overlapping parts, the first part containing scans No.1 through No.280, and the second part containing scans No.281 through No.400. Using the terminology from the pattern recognition community, the first part is called the training set and the second part the test set. The intention of this division is to calculate the full Procrustes mean shapes for each part separately and compare them. Figure 4.7 (a) & (b) plot the four Procrustes mean shapes of samples I ~ IV for the training set and the test set, respectively.



(a)



(b)

Figure 4.7 Procrustes mean shapes of samples I ~ IV for the training set (a), the test set (b).

The dynamic time warping (DTW) distance measure is utilized to measure the similarity between different Procrustes mean shapes. The DTW measure calculates the similarity between two vectors based on their best nonlinear alignment version, hence being insensitive to the relative shift and relative stretch of one vector with respect to the other. In the following, the definition and calculation of the DTW distance measure is briefly reviewed. For a more complete cover of the DTW algorithm, please refer to [Salvador and Chan, 2007].

Definition 4.4 [Salvador and Chan, 2007]: Given two vectors, X and Y, of length $|X|$ and $|Y|$, respectively, $X(i)$, $Y(i)$ denote the i^{th} component of X, Y, respectively. Construct a warp path, $W = w_1, w_2, \ldots, w_K$, where K is the length of the warp path satisfying

$$\max(|X|,|Y|) \le K \le |X| + |Y|. \tag{4.19}$$

The k^{th} element of the warp path, W, is

$$w_k = (i, j) \tag{4.20}$$

where i is an index from the vector X and j is an index from the vector Y. Also let $w_{k,1}$ and $w_{k,2}$ denote the first and the second component of w_k, respectively. The warp path, W, must satisfy

$$w_1 = (1,1),\ w_K = (|X|,|Y|) \tag{4.21}$$

$$w_k = (i, j),\ w_{k+1} = (i', j') \ \Rightarrow\ i \le i' \le i+1,\ j \le j' \le j+1. \tag{4.22}$$

The distance of a warp path W, $Dist(W)$, is defined as

$$Dist(W) = \sum_{k=1}^{K} |X(w_{k,1}) - Y(w_{k,2})|. \tag{4.23}$$

The optimal warp path is the warp path with the minimum distance.

Proposition 4.3 [Salvador and Chan, 2007]: A dynamic programming approach is used to calculate the distance of the optimal warp path. A two-dimensional $|X| \times |Y|$ distance measure matrix, D, is constructed where the $(i, j)^{th}$ value, $D(i, j)$, is the distance of the optimal warp path that can be constructed from the two vectors, $X' = [X(1), \ldots, X(i)]$ and

$Y' = [Y(1), \ldots, Y(j)]$. The value, $D(|X|, |Y|)$, is termed the DTW distance measure between the vectors, X and Y. The dynamic programming algorithm to find $D(|X|, |Y|)$ is

$$D(i, j) = |X(i) - Y(j)| + \min\{D(i-1, j), D(i, j-1), D(i-1, j-1)\}. \qquad (4.24)$$

■

The DTW distance measures between different pairs of Procrustes mean shapes are tabulated in Table 4.2 as a matrix form, whose $(i, j)^{th}$ element represents the DTW distance measure between the training set Procrustes mean shape of sample i and the test set Procruestes mean shape of sample j.

DTW	I	II	III	IV
I	**0.13**	0.34	0.51	0.24
II	0.36	**0.14**	0.86	0.34
III	0.44	0.63	**0.16**	0.42
IV	0.28	0.19	0.66	**0.15**

Table 4.2 The DTW distance measures between the training set Procrustes mean shapes of samples I ~ IV, and the test set Procrustes mean shapes of samples I ~ IV.

In Table 4.2, all four diagonal elements are the minimum ones among their corresponding columns. This shows that all four test data will be classified correctly if the DTW distance measure between the Procrustes mean shapes is used as the decision criterion. In another word, this result verifies the intuition that the full Procrustes mean shape can be an informative feature for target identification. Furthermore, in the following it can be shown

that through the use of the Hotelling one sample T^2 test, one can carry out this comparison process under a statistical inference framework.

4.5.2 Tangent Space Inference

The tangent space is tangent to the unit size shape sphere on a particular point called the pole of the tangent projection. The projection pole is usually adaptively chosen to be an average shape obtained from the data set. Then the tangent space is a linear approximation of the unit size shape sphere in the vicinity of the projection pole. One prominent advantage of tangent space is that on this linear space, the classical multivariate statistical analysis techniques can be directly applied. For a detailed explanation of the geometric and algebraic properties in the tangent space, please refer to [Chapter 4, Dryen and Mardia, 1998]. Figure 4.8 gives a geometric view of the unit size shape sphere, a unit size Procrustes mean shape, γ, the tangent space on γ, $T(\gamma)$, the full Procrustes fit of w onto γ, w^P, the full Procrustes distance, $d_F(\gamma, w)$, and the tangent plane coordinate of w^P onto $T(\gamma)$, v. Taking the notation of Section 4.5.1 and Proposition 4.1, it is known that

$$w^P = \frac{w^H \gamma w}{w^H w}. \tag{4.25}$$

So the tangent plane coordinate, v, turns out to be

$$v = w^P - (\gamma^H w^P)\gamma = \frac{w^H \gamma w}{w^H w} - \frac{\gamma^H w^H \gamma w \gamma}{w^H w} = \frac{(w^H \gamma) w - \left| w^H \gamma \right|^2 \gamma}{\|w\|^2}. \tag{4.26}$$

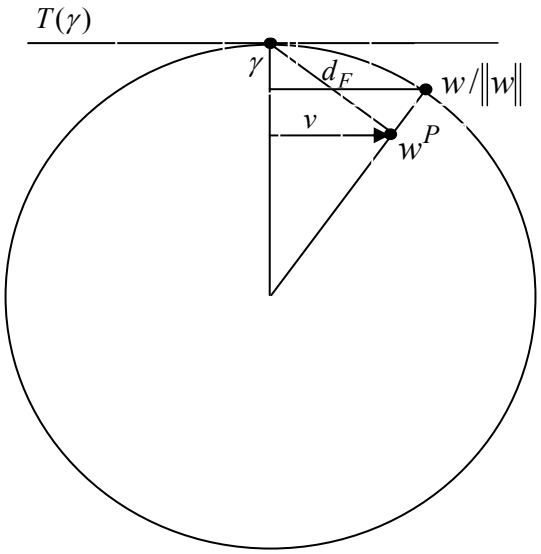


Figure 4.8 A geometric view of the unit size shape sphere, the tangent space, $T(\gamma)$, the Procrustes fit, w^P, the Procrustes distance, d_F, and the tangent plane coordinates, v.

In the UWB radar signal context, consider a hypothesis test on whether or not the Procrustes mean shape of a population has a specific shape $[\mu_0]$, i.e.,

$$H_0 : [\mu] = [\mu_0] \quad \text{versus} \quad H_1 : [\mu] \neq [\mu_0]. \tag{4.27}$$

Let $\{w_1, \ldots, w_n\}$ be a random sample of preprocessed radar range profiles with each $w_i \in \Re_+^m$, and $\hat{\mu}$ be their full Procrustes mean shape with unit Euclidean norm. The full Procrustes fits of $w_1, \ldots, w_n$ onto $\hat{\mu}$ are,

$$w_i^P = \frac{w_i^T \hat{\mu} w_i}{w_i^T w_i}, \quad i = 1, \ldots, n. \tag{4.28}$$

And the tangent plane coordinates of w_i^P onto $T(\hat{\mu})$, v_i, have been derived as

$$v_i = \frac{(w_i^T \hat{\mu}) w_i - \left| w_i^T \hat{\mu} \right|^2 \hat{\mu}}{\left\| w_i \right\|^2}, \quad i = 1, \ldots, n. \tag{4.29}$$

Similary, the full Procrustes fit of μ_0 onto $\hat{\mu}$, μ_0^P, and the tangent plane coordinate of μ_0^P onto $T(\hat{\mu})$, v_0, are,

$$\mu_0^P = \frac{\mu_0^T \hat{\mu} \mu_0}{\mu_0^T \mu_0} \quad \text{and} \quad v_0 = \frac{(\mu_0^T \hat{\mu}) \mu_0 - \left| \mu_0^T \hat{\mu} \right|^2 \hat{\mu}}{\left\| \mu_0 \right\|^2}. \tag{4.30}$$

To obtain the Hotelling one sample T^2 test, assume v_i obeys the standard multivariate normal model, i.e., $v_i \sim N(\eta, \sum)$ and v_i are independent, $i = 1, \ldots, n$. Denote $\bar{v} = \frac{1}{n} \sum_{i=1}^{n} v_i$ for the sample mean and $S = \frac{1}{n} \sum_{i=1}^{n} (v_i - \bar{v})(v_i - \bar{v})^T$ for the maximum likelihood estimation (MLE) of the covariance matrix, $\sum$, then the one sample Hotelling T^2 test statistic is given as

$$F = \{(n - M)/M\} (\bar{v} - v_0)^T S^{-1} (\bar{v} - v_0) \tag{4.31}$$

where $M = m - 1$ is the dimension of the tangent space and S^{-1} is the Moore-Penrose generalized inverse of S. From [p151, Dryen and Mardia, 1998; Proposition C.7, Appendix C], $F \sim F_{M, n-M}$ under H_0, so H_0 is rejected for large values of F.

The calculated one sample Hotelling T^2 statistic for the samples I ~ IV are summarized as follows. For each sample, $\{w_1, \ldots, w_{280}\}$ and $\{w_{281}, \ldots, w_{400}\}$ are the training set and the

test set, respectively, with each $w_i \in \Re_+^{81}$, and $\hat{\mu}$ is the test set Procrustes mean shape with unit norm. Consider a hypothesis test on whether the test population mean shape can be μ_0, which is taken as the Procrustes mean shape calculated from the training set. In other words, this hypothesis test is to evaluate whether the test population mean shape can match the template generated from the training set. The calculated Hotelling T^2 statistics are organized into a matrix shown in Table 4.3, whose $(i, j)^{th}$ element represents the F value when using the test set of sample i to match the mean shape generated from the training set of sample j.

F	I	II	III	IV
I	**2.45**	51.4	55.4	54.4
II	43.6	**2.57**	32.6	13.2
III	210.9	514.1	**12.8**	900.9
IV	28.4	14.0	42.1	**5.07**

Table 4.3 The F values when using the test sets of samples I ~ IV to match the training set Procrustes mean shapes from samples I ~ IV.

In Table 4.3, all four diagonal elements are the minimum ones among their corresponding rows. This shows that all four test data will be classified correctly if the F statistic is used as the decision criterion. But there is one unanticipated phenomenon: the random variable $F_{M,n-M}$ with parameters $M = 80$ and $n = 120$ has a very spiky p.d.f around 1, hence making even the best matches an extremely low p-level, e.g. $P(F_{80,40} > 2.45) = 0.0012$. This unusual occurrence comes from the fact that the rank of the sample covariance matrix,

$\rho(S)$, is assumed to be M when deriving the F statistic. Actually, $\rho(S) << M$, because there are strong temporal structures in the sequential radar data. What this rank insufficiency implies about the appropriate setup of parameters, (M, n), is still a mystery.

4.6 Shape Variability by Principal Component Analysis

The full Procrustes mean shape in Section 4.5 provides a way to define a representative template for a sequence of target images. But for a sequence of target images, they also exhibit significant variations around their mean shape. This section investigates the shape variation structure of radar range profiles, and verifies its use as one promising feature for the classification. If one views a target image as a specific realization of a random vector, and plots it in the high dimensional space, then because of the redundancy in the random vector components, the main structure of the data cluster from a sequence of target images will reside in a much lower dimensional subspace. This section is to reveal this main structure by the projection pursuit technique, which offers the selected low-dimensional orthogonal projection of data by optimizing an index of goodness - projection index. Classical projection indexes include, but are not limited to, variance [Jolliffe, 2002] and entropy [Hyvarinen, 1998]. This thesis, by no means, intends to cover this field even superficially. Academic treatements on projection pursuit methods can be found in [Jones and Sibson, 1987; Huber, 1985; Nason, 1992; Pajunen, 1998; Friedman and Tukey, 1974; Friedman, 1987]. In this section, one dominant projection pursuit method, principal component analysis, is adopted due to its mathematical neatness and experimental success.

Principal component analysis (PCA) was developed by Hotelling [Hotelling, 1933] after its origin by Karl Pearson [Pearson, 1901]. Generally speaking, PCA looks for the linear transformations, which provide "parsimonious summarization" of the data, losing in the process as little information as possible, and generate features that are mutually uncorrelated in order to avoid information redundancies. PCA is of fundamental significance in pattern recognition, e.g., face recognition [Turk and Pentland, 1991], and in a wide range of applications [p9, Jolliffe, 2002]. For self-containment and terminology introduction of this section, [Appendix D] briefly reviews one classic construction process for PCA and its two key properties. More complete coverage on PCA can be found in many nice references, such as [Jolliffe, 2002; p266~p272, Theodoridis and Koutroumbas, 2006; Chapter 8, Mardia, et al., 1979; Chapter 11, Anderson, 2003].

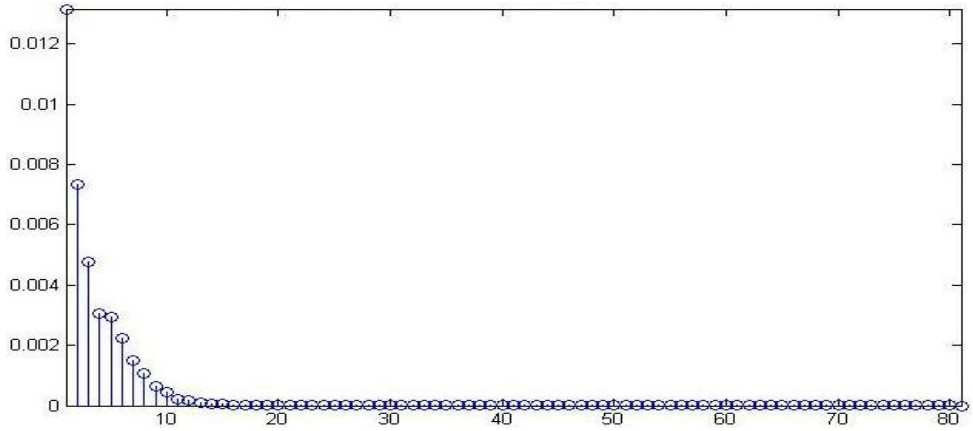
Principal component analysis of the sample covariance matrix in the tangent space provides an efficient way to analyze the main modes of shape variation. As in Section 4.5, consider a real $m \times n$ matrix $A = [a_1, \ldots, a_n]$, with the columns, $a_i \in \Re_+^m$, representing the sequential preprocessed target images. Assume $\hat{\mu}$ is the full Procrustes mean shape of $\{a_1, \ldots, a_n\}$ with unit Euclidean norm, then the full Procrustes fits of $a_1, \ldots, a_n$ onto $\hat{\mu}$ are,

$$a_i^P = \frac{a_i^T \hat{\mu} a_i}{a_i^T a_i}, \quad i = 1, \ldots, n \tag{4.32}$$

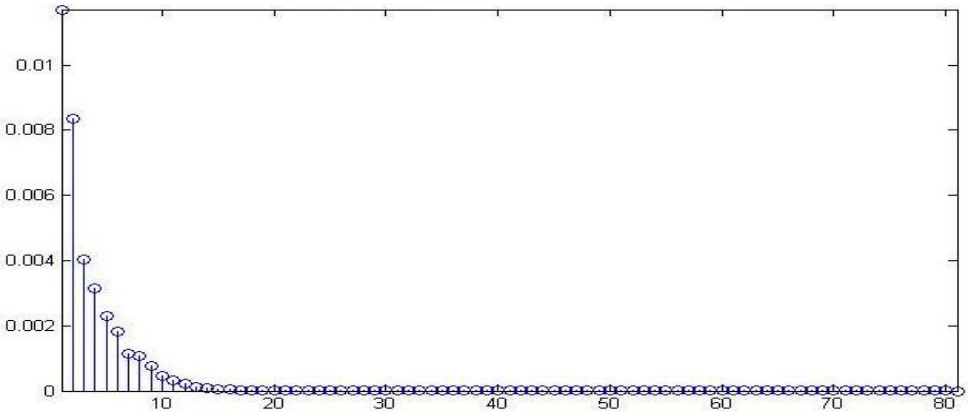
and the tangent plane coordinates of a_i^P onto $T(\hat{\mu})$, v_i, are given as

$$v_i = \frac{(a_i^T \hat{\mu})a_i - \left|a_i^T \hat{\mu}\right|^2 \hat{\mu}}{\|a_i\|^2}, \quad i = 1,\ldots,n. \tag{4.33}$$

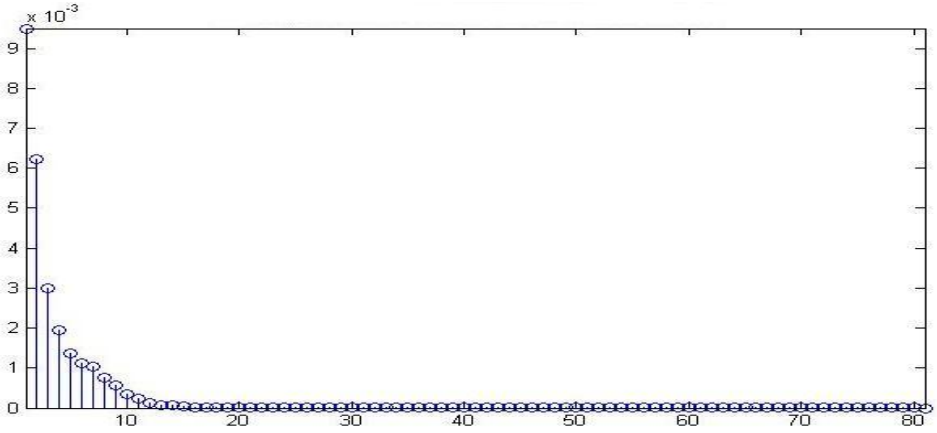
Stack v_i columnwise into a matrix, $G = [v_1,\ldots,v_n]$, and call G the matrix of tangent plane coordinates. $\{v_1,\ldots,v_n\}$ can be viewed as n realizations of a random vector X, then an unbiased estimator of $\sum_X$ is $S_u = \frac{1}{n-1}\sum_{i=1}^{n}(v_i - \bar{v})(v_i - \bar{v})^T$, where $\bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_i$. Figure 4.9 illustrates the eigenvalues of S_u for samples I ~ IV. Figure 4.10 plots the first four pairs of eigenvectors and principal components of S_u for each sample I ~ IV.



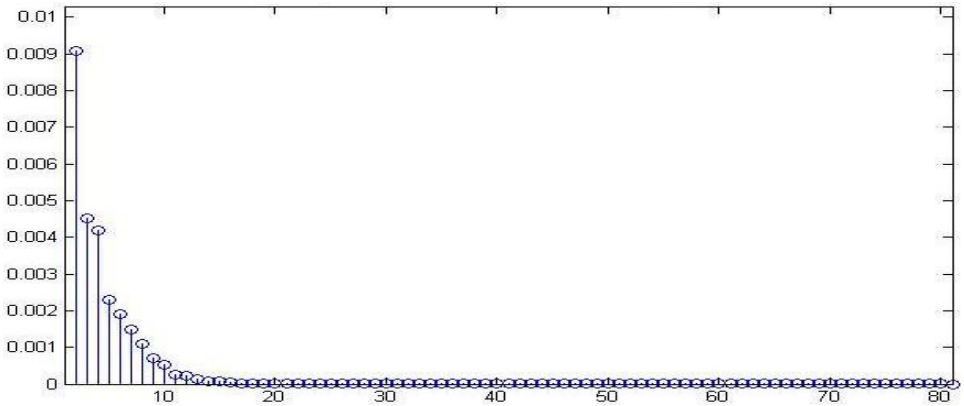
(a)



(b)

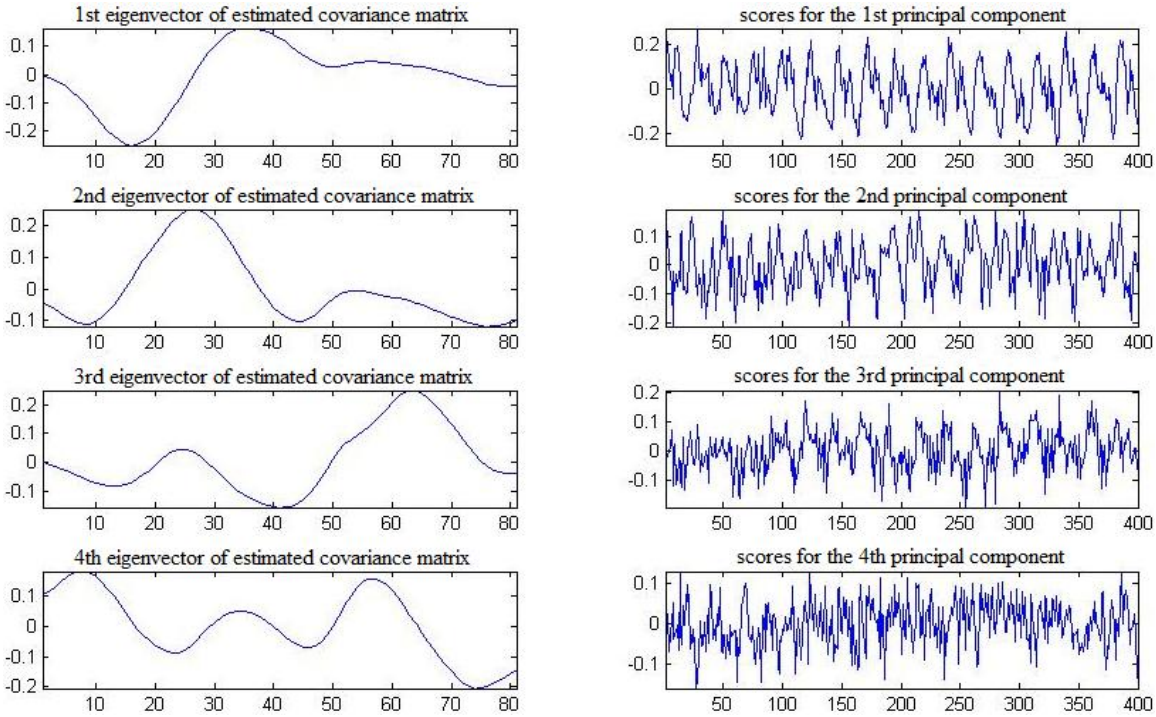


(c)

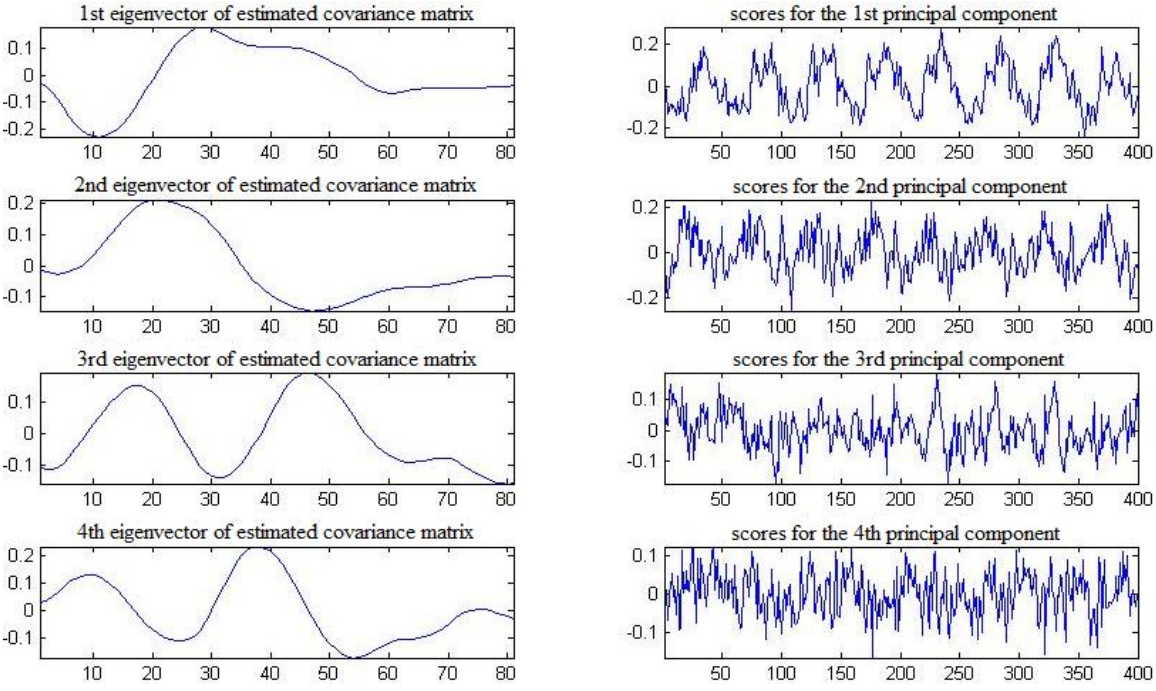


(d)

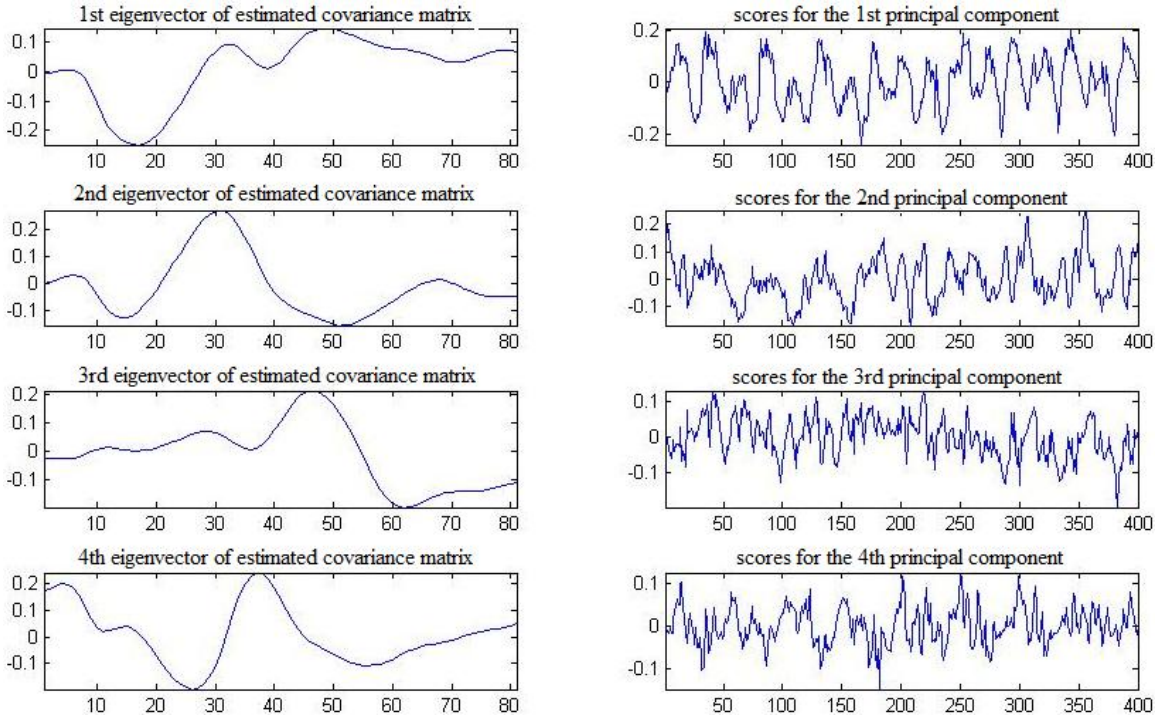
Figure 4.9 Eigenvalues of S_u, sample I (a), sample II (b), sample III (c), sample IV (d).


1st eigenvector of estimated covariance matrix
2nd eigenvector of estimated covariance matrix
3rd eigenvector of estimated covariance matrix
4th eigenvector of estimated covariance matrix
scores for the 1st principal component
scores for the 2nd principal component
scores for the 3rd principal component
scores for the 4th principal component

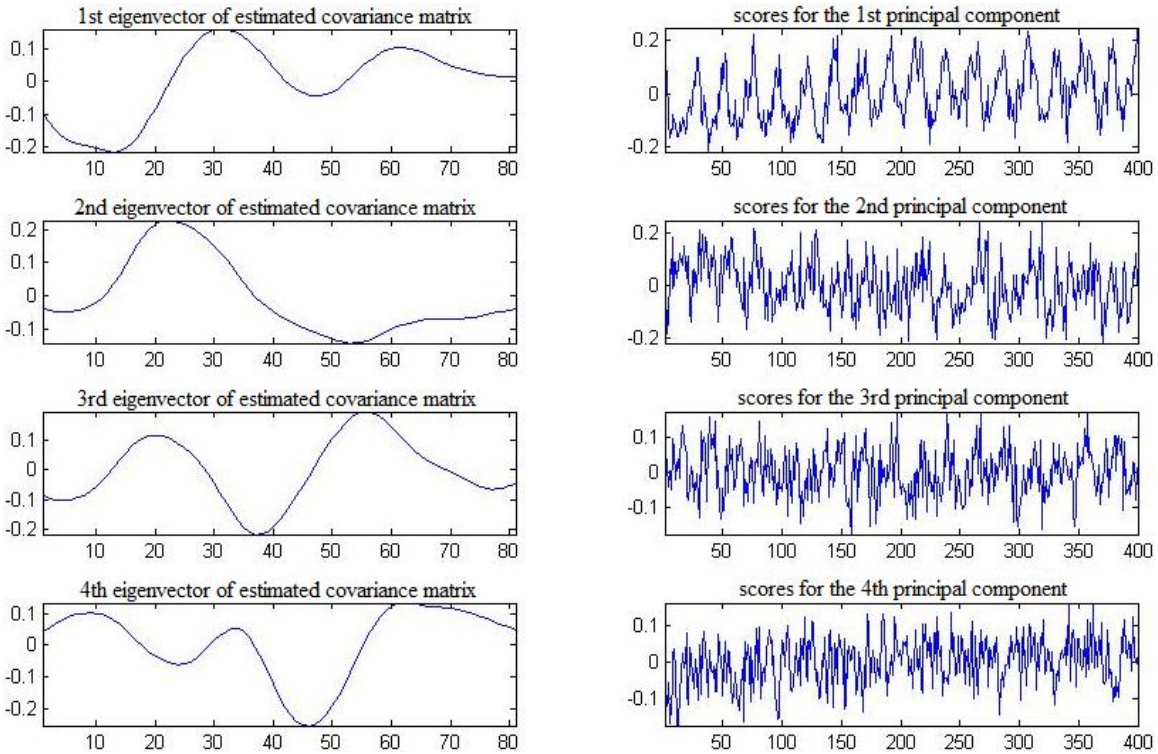
(a)


1st eigenvector of estimated covariance matrix
2nd eigenvector of estimated covariance matrix
3rd eigenvector of estimated covariance matrix
4th eigenvector of estimated covariance matrix
scores for the 1st principal component
scores for the 2nd principal component
scores for the 3rd principal component
scores for the 4th principal component

(b)


1st eigenvector of estimated covariance matrix
2nd eigenvector of estimated covariance matrix
3rd eigenvector of estimated covariance matrix
4th eigenvector of estimated covariance matrix
scores for the 1st principal component
scores for the 2nd principal component
scores for the 3rd principal component
scores for the 4th principal component

(c)


1st eigenvector of estimated covariance matrix
2nd eigenvector of estimated covariance matrix
3rd eigenvector of estimated covariance matrix
4th eigenvector of estimated covariance matrix
scores for the 1st principal component
scores for the 2nd principal component
scores for the 3rd principal component
scores for the 4th principal component

(d)

Figure 4.10 Eigenvectors of S_u (left panel) and scores of principal components of G (right panel) for sample I (a), sample II (b), sample III (c), sample IV (d).

The sparsity pattern of eigenvalues of S_u in Figure 4.9 and optimal representation properties of PCA [Propositions D.1 and D.2] lead the analysis to focus on the first several eigenvectors of S_u and the scores of corresponding PCs of the matrix of tangent plane coordinates, G. This is a reasonable simplification, because the top PCs have explained most of the variance inside the original data. For example, in the illustrated data sets, the top 4 out of 81 PCs account for 74.8%, 76.2%, 78%, and 75.9% of the total variance of G for samples I ~ IV, respectively. So the top principal components can capture the prominent target dynamics compactly. For example, in Figure 4.10 (a), the stepping pace of the experimental person in orientation I is clearly identified through the first PC. In Figure 4.10 (b), the left leg stepping pace and right leg stepping pace in orientation II are identified through the first and the second PCs, respectively. This can be understood better if it is noticed that the waveforms of the corresponding two eigenvectors have an approximate relative shift, hence accentuating the motion pattern of different body parts. In Figure 4.10 (c), the strong pace of one visible leg and relatively weak pace of another partially occluded leg in orientation III are both captured by the first PC. A closer look at the eigenvector reveals that this is achieved by putting contrast weights on the range of presence of two legs, respectively. In Figure 4.10 (d), the walking pace of the experimental person in orientation IV is clearly identified through the first PC.

It is promising to analyze the top principal components to extract target dynamic information, meanwhile, the corresponding eigenvectors of the sample covariance matrix, S_u, are quite worthy of further exploration. The Procrustes mean shape defines a static template, and the top eigenvectors of S_u build an orthonormal basis for a subspace which captures the dominant data variation structure in the tangent space. The combination of the Procrustes mean shape and the dominant data variation structure provides a more complete version of static features, which will augment the robustness of the radar target recognition provided one can compare not only the mean templates, but also the data variation structures. The similarity between two data variation structures can be measured through comparing their residing subspaces. This method was proposed in [Krzanowski, 1979], and is briefly reviewed in the following and [Appendix E].

Assume two random vectors, $X \in \Re^m$ and $Y \in \Re^m$, have the covariance matrices of $\sum_X$ and $\sum_Y$, respectively. Suppose the coefficients for the first p principal components (PCs) of X and Y are $\{\alpha_1,\ldots,\alpha_p\}$ and $\{\beta_1,\ldots,\beta_p\}$ respectively, that is, $\{\alpha_1,\ldots,\alpha_p\}$ are the first p orthonormal eigenvectors of $\sum_X$, and $\{\beta_1,\ldots,\beta_p\}$ are the first p orthonormal eigenvectors of $\sum_Y$. In order to compare these two sets of PCs, it is necessary to compare the two subspaces spanned by $\{\alpha_1,\ldots,\alpha_p\}$ and $\{\beta_1,\ldots,\beta_p\}$, respectively. The following two theorems from [Krzanowski, 1979] propose one elegant way to do it.

Proposition 4.3 [Krzanowski, 1979; Appendix E]: $\{\alpha_1,\ldots,\alpha_p\}$ and $\{\beta_1,\ldots,\beta_p\}$ denoting the same as in the above paragraph, construct two matrices, $L=[\alpha_1,\ldots,\alpha_p]$ and $M=[\beta_1,\ldots,\beta_p]$, the minimum angle between an arbitrary vector in the subspace $span\{\alpha_1,\ldots,\alpha_p\}$ and another arbitrary vector in the subspace $span\{\beta_1,\ldots,\beta_p\}$ is given by $\cos^{-1}(\sqrt{\lambda_1})$, where λ_1 is the largest eigenvalue of $K=L^TMM^TL$.

■

Proposition 4.4 [Krzanowski, 1979; Appendix E]: L, M and K are defined as in Proposition 4.3. Let λ_i, ν_i be the i-th largest eigenvalue and corresponding eigenvector of K, respectively. Take $w_i=L\nu_i$, then $\{w_1,\ldots,w_p\}$ and $\{MM^Tw_1,\ldots,MM^Tw_p\}$ form orthogonal vectors in $span\{\alpha_1,\ldots,\alpha_p\}$ and $span\{\beta_1,\ldots,\beta_p\}$, respectively. The angle between the i-th pair, (w_i,MM^Tw_i), is given by $\cos^{-1}(\sqrt{\lambda_i})$. Proposition 4.3 shows that w_1 and MM^Tw_1 give the two closest vectors when one is constrained to be in the subspace $span\{\alpha_1,\ldots,\alpha_p\}$ and the other in $span\{\beta_1,\ldots,\beta_p\}$. It follows that w_2 and MM^Tw_2 give directions, orthogonal to the previous ones, between which lies the next smallest possible angle between the subspaces.

■

Let θ_{ij} be the angle between α_i and β_j, i.e., $\cos(\theta_{ij})=\alpha_i^T\beta_j$, then $L^TM=\{\cos(\theta_{ij})\}$,

$$\sum_{i=1}^{p}\lambda_i=trace(L^TMM^TL)=\sum_{j=1}^{p}\sum_{i=1}^{p}\cos^2(\theta_{ij}). \tag{4.34}$$

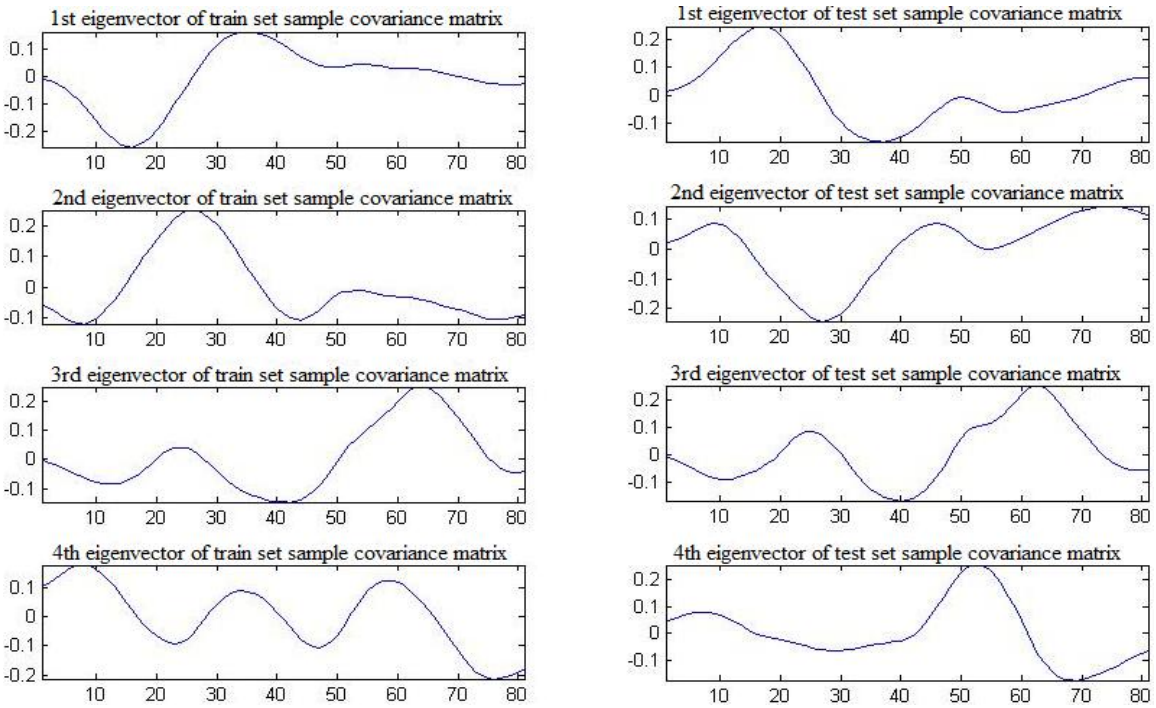
Thus the summation of the eigenvalues of $K=L^TMM^TL$ equals the sum of squares of

the cosines of the angles between each basis element of $span\{\alpha_1,\ldots,\alpha_p\}$ and each basis element of $span\{\beta_1,\ldots,\beta_p\}$. This sum is invariant with respect to whichever basis you select for $span\{\alpha_1,\ldots,\alpha_p\}$ and $span\{\beta_1,\ldots,\beta_p\}$ [Appendix E], so it can be used as a measure of total similarity between the two subspaces. Also it can be checked that

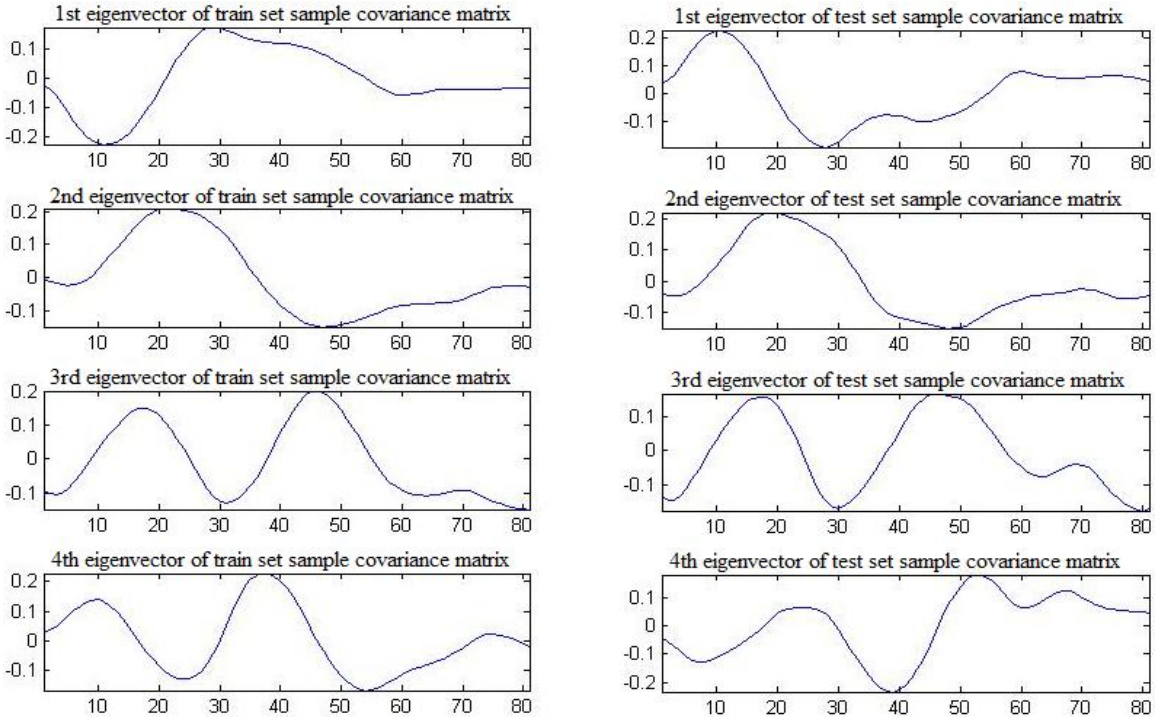
$$\sum_{i=1}^{p}\lambda_i = p\,, \quad \text{if } span\{\alpha_1,\ldots,\alpha_p\} = span\{\beta_1,\ldots,\beta_p\} \tag{4.35}$$

$$\sum_{i=1}^{p}\lambda_i = 0\,, \quad \text{if } span\{\alpha_1,\ldots,\alpha_p\} \perp span\{\beta_1,\ldots,\beta_p\}\,. \tag{4.36}$$

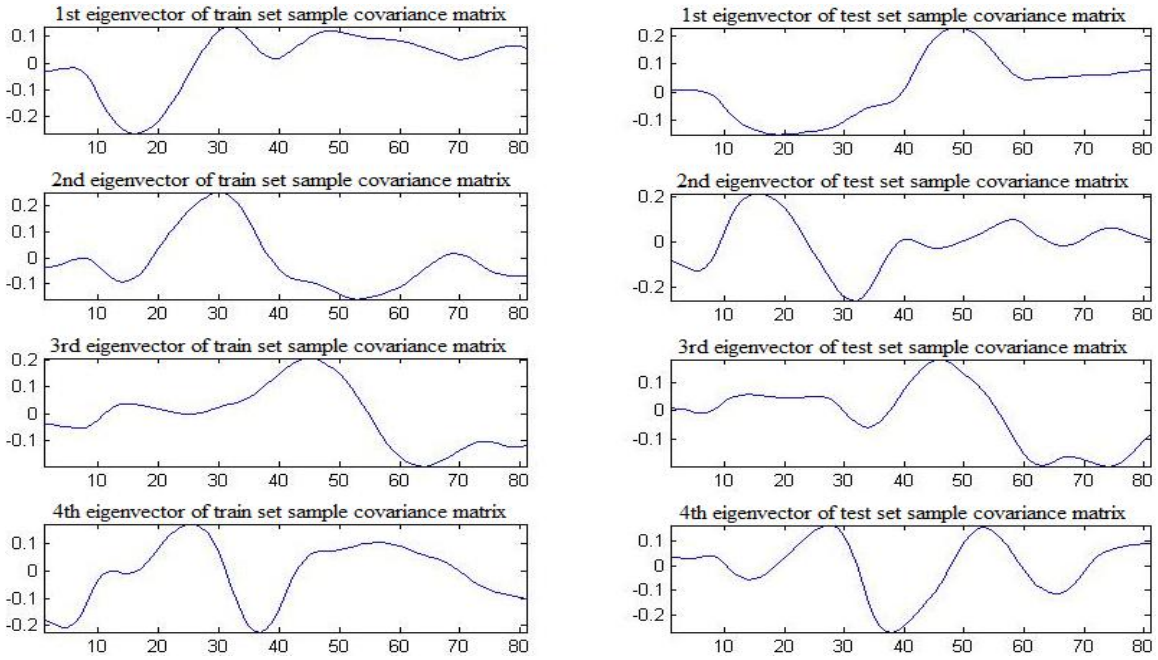
To explore the discrimination power of the principal components, each of samples I ~ IV is divided into two non-overlapping parts. The first part containing scans No.1 through No.280 is called the training set, and the second part containing scans No.281 through No.400 is called the test set. This test concerns comparing the PCs of the training set with the PCs of the test set from each sample. Figure 4.11 (a) ~ (d) plot the top four eigenvectors of S_u of the training set (left panel) and the test set (right panel) for samples I ~ IV, respectively.


1st eigenvector of train set sample covariance matrix
2nd eigenvector of train set sample covariance matrix
3rd eigenvector of train set sample covariance matrix
4th eigenvector of train set sample covariance matrix
1st eigenvector of test set sample covariance matrix
2nd eigenvector of test set sample covariance matrix
3rd eigenvector of test set sample covariance matrix
4th eigenvector of test set sample covariance matrix

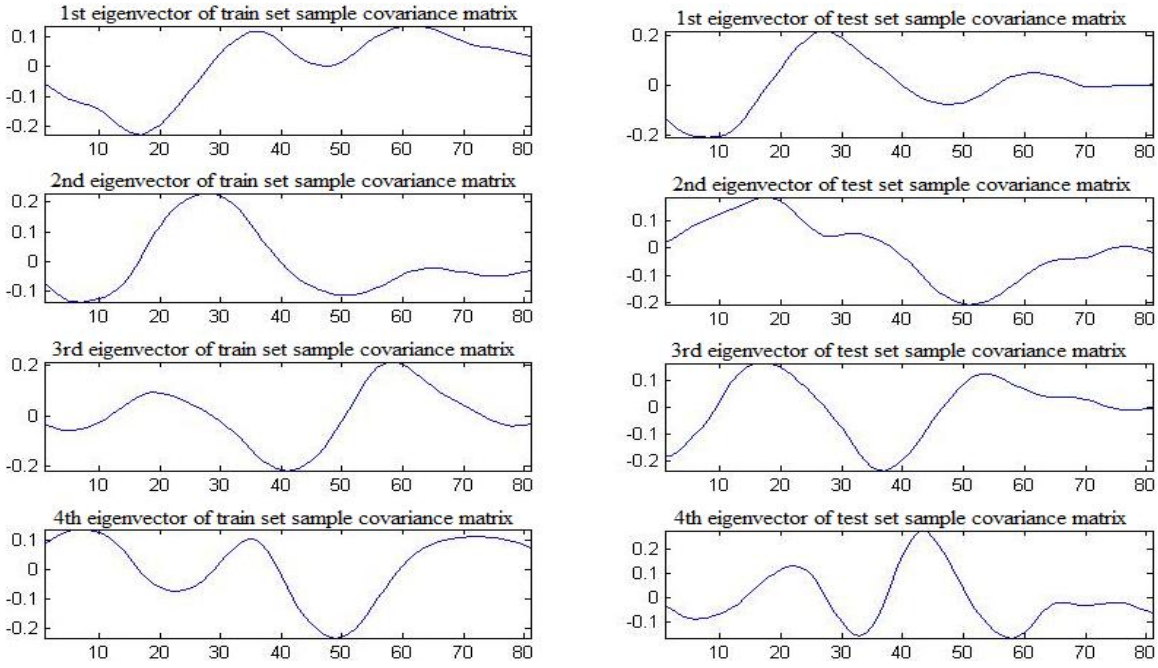
(a)


1st eigenvector of train set sample covariance matrix
2nd eigenvector of train set sample covariance matrix
3rd eigenvector of train set sample covariance matrix
4th eigenvector of train set sample covariance matrix
1st eigenvector of test set sample covariance matrix
2nd eigenvector of test set sample covariance matrix
3rd eigenvector of test set sample covariance matrix
4th eigenvector of test set sample covariance matrix

(b)



(c)



(d)

Figure 4.11 The top four eigenvectors of S_u of the training set (left panel) and the test set (right panel) for sample I (a), sample II (b), sample III (c), sample IV (d).

Denote $L=[\alpha_1,\ldots,\alpha_p]$, $M=[\beta_1,\ldots,\beta_p]$, and $K=L^T MM^T L$, where $\{\alpha_1,\ldots,\alpha_p\}$ and $\{\beta_1,\ldots,\beta_p\}$ are the first p eigenvectors of S_u for the training set and the test set, respectively, then the total similarity between two subspaces, $span\{\alpha_1,\ldots,\alpha_p\}$ and $span\{\beta_1,\ldots,\beta_p\}$, can be measured as $trace(K)$. The calculated values of $trace(K)$ are organized as two matrices shown in Table 4.4 (a) for $p=3$ and (b) for $p=4$. The $(i,j)^{th}$ element in the matrix represents the similarity measure, $trace(K)$, when comparing the first p PCs of the training set of sample i with the first p PCs of the test set of sample j.

$trace(K)$	I	II	III	IV
I	**2.88**	1.89	1.64	2.08
II	1.82	**2.89**	2.19	1.92
III	1.88	2.35	**2.55**	1.73
IV	2.52	1.83	1.60	**2.32**

(a)

$trace(K)$	I	II	III	IV
I	**3.14**	2.72	2.01	2.61
II	2.99	**3.84**	2.88	3.33
III	2.93	3.42	**2.94**	3.02
IV	**3.21**	3.34	2.61	**3.35**

(b)

Table 4.4 Comparing the first p PCs of the training sets of samples I ~ IV with the first p PCs of the test sets of samples I ~ IV, $p=3$ (a), $p=4$ (b).

In Table 4.4, all the matrix diagonal elements are the maximum ones among their corresponding columns, except the first data column in (b). This shows that most test data

will be classified correctly if the similarity measure, $trace(K)$, is used as the decision criterion. The exceptional case is with orientation I and $p = 4$. To understand this special case better, the cosines of the angles between columns of L and M are calculated. Specifically, two matrices, $L_I^T M_I$ and $L_{IV}^T M_I$, are computed. L_I, L_{IV} consist of the first 4 eigenvectors of S_u for the training set of sample I and IV, respectively, and M_I consists of the first 4 eigenvectors of S_u for the test set of sample I. Table 4.5 shows the elements of $L_I^T M_I$ in (a), $L_{IV}^T M_I$ in (b).

-0.98	-0.07	0.02	-0.02
0.08	**-0.96**	0.15	-0.13
0	0.16	**0.96**	-0.17
-0.1	-0.18	0.03	**0.42**

(a)

-0.85	0.25	0.34	-0.16
-0.12	**-0.86**	0.01	-0.46
0.36	-0.08	**0.85**	0.09
0.12	0.22	-0.09	**-0.62**

(b)

Table 4.5 Cosines of the angles between columns of L and M; the result of $L_I^T M_I$ is shown in (a), and $L_{IV}^T M_I$ in (b).

Table 4.5 (a) shows that the first 4 eigenvectors of S_u for the training set and the test set of sample I match very well except for the 4th eigenvector. Because the subspace similarity measure, $trace(K)$, doesn't consider the relative dominancy of each principal

component, two sets of principal components matching very well in all pairs except in the minor one may not have the best match when evaluated from the subspace similarity measure. This exceptional case doesn't impair the promising usefulness of the PC comparison as the classification feature, instead, it triggers a strong need to develop a more robust way to measure the similarity of data variation structures, which should take more factors into account, e.g., the relative dominancy of each principal component.

CHAPTER 5

SUMMARY AND FURTHER RESEARCH

5.1 Summary

Feature selection from measured data presents one significant challenge for diverse fields of science and engineering. Nowadays, as large and complex data sets become typical and usual, it is essential to extract the informative features revealing the statistic or stochastic mechanism which generates the data. This thesis explores the feature selection problem in two scenarios, pattern recognition and ultra-wideband radar signal analysis.

Classical pattern recognition methods view feature selection as an independent stage, which aims to find an optimal or suboptimal subset of candidate features with high separability scores based on a given gauge measure. There are deficiencies to this general approach in specific situations. Firstly, a feature may not necessarily be a good feature even when its separation score is high. Especially, this leads to a problem in our concerned application, neural signal decoding, in which a single neuron and its associated features have only weak classification ability. In such cases, one cannot guarantee a statistically significant difference between selected features and discarded ones by the gauge measure. Secondly, classical feature selection depends on the set of training data. As a result, feature selection and the following decision rule learning must start from

scratch if the pool of candidate patterns or the set of training samples changes. This adds a significant computational burden for some applications, such as neural signal decoding, whose data generation mechanism can be quite variable. To overcome these deficiencies, this thesis presents a new adaptive sequential feature selection algorithm which utilizes an information-theoretic measure and the nonparametric kernel density estimation method to sequentially reduce the complexity of the classification task, and finally outputs the probabilistic classification result and its variance estimate. The point is that the selected feature can be a function of not only the training samples, but also of the unlabeled test data.

Brain computer interfaces (BCIs) aim to utilize the biological understanding of brain functionality and operating mechanism to enable people to control external devices merely by thought. One important application of BCI technique is to construct neural prosthetic systems which tap into the thoughts of millions of paralytic patients who are deprived of motor abilities, but not cognitive functions. The design and construction of such an automatic mechanism involves challenges in diverse disciplines. What is concerned in this thesis is decoding a finite number of classes from neural recordings through a combined application of the adaptive sequential feature selection algorithm and information fusion methods. This combination presents a new classification scheme that is particularly well suited to neural signal decoding, as it fulfills many of the challenges that are specific to neural prosthetic systems, which include real time decoding, adaptation to non-stationary data generation mechanism, classification under significant overlaps of classes in the feature space, and robust information pooling from multiple neurons. Experimental results

show that the proposed neural decoding method outperforms the classical pattern recognition methods, the k -NN rules and the C-SVC classifier, in decoding performance.

The motivation behind the second research topic in this thesis is to augment pedestrian safety by developing a human presence detector and a human behavior classifier through the use of ultra-wideband (UWB) impulse radar. The UWB system is ideal for its abundant information content and precise positioning, however no systematic theory of UWB is available due to the highly complex signal transformations involved. Hence the well-known radar target recognition technique through the Doppler shift effect doesn't apply in UWB systems. Novel methods must be developed for this application. To prominently distinguish people from other targets and classify people's behaviors, information from both static and dynamic biometric features should be pooled. The static features can reflect information about the target geometry and its variation structure, while the dynamic features extract the temporal structure among a sequence of radar scans. This thesis is devoted to the problem of generating the static template for a sequence of target images, locating the high information packing subspace, and exploring their statistical and algebraic properties. Firstly, the introduction of the preprocessing of range profile extraction makes radar data more amiable for further analysis. Next, the Procrustes shape analysis is utilized to generate a representative template, the Procrustes mean shape, for the radar data set, and the statistical inference about the Procrustes mean shape is carried out in the tangent space through the Hotelling one sample T^2 test. After that, the waveform shape variation structure in the tangent space is analyzed through principal component analysis (PCA).

PCA analysis not only provides a new kind of static features for classification, but also accentuates the prominent dynamics of the target motion. Undoubtedly, the combination of the Procrustes shape analysis and shape variability analysis forms a more complete platform to extract static features for the UWB radar signal.

5.2 Further Research

In this thesis, the combination of the sequential feature selection algorithm and the information fusion methods shows the potential advantage of robustly assembling the information of multiple neurons. Despite this initial success, deeper work should be done to improve the match between the algorithmic design and the neural operation mechanism. This work includes,

(1) In the adaptive sequential feature selection algorithm, T is the threshold set to make the decision of retaining or removing sequentially, and t is the threshold set to determine whether an early termination of the sequential process is necessary. Current settings of T and t are empirical, but should be developed in a more rigorous fashion.

(2) Minimizing the average Kullback-Leibler distance is one way to justify the information fusion method – the product rule. This information fusion strategy assumes no data variation for the estimate of the posterior probability distribution from each classifier. The next question is how to generalize this information fusion method to the case where each classifier outputs not only an estimate of the posteriori probabilities, but

also an estimate of their variances.

In radar signal analysis, this thesis exhibits the discrimination ability of static features generated from Procrustes shape analysis and shape variability analysis. However the target dynamic information can also be explored as informative identification cues due to its correlation with behavioral characteristics. Fusion of static and dynamic features will augment the performance of automatic target identification further. Figure 5.1 gives a diagrammatic description of the design of an automatic target recognition system along this line of reasoning.

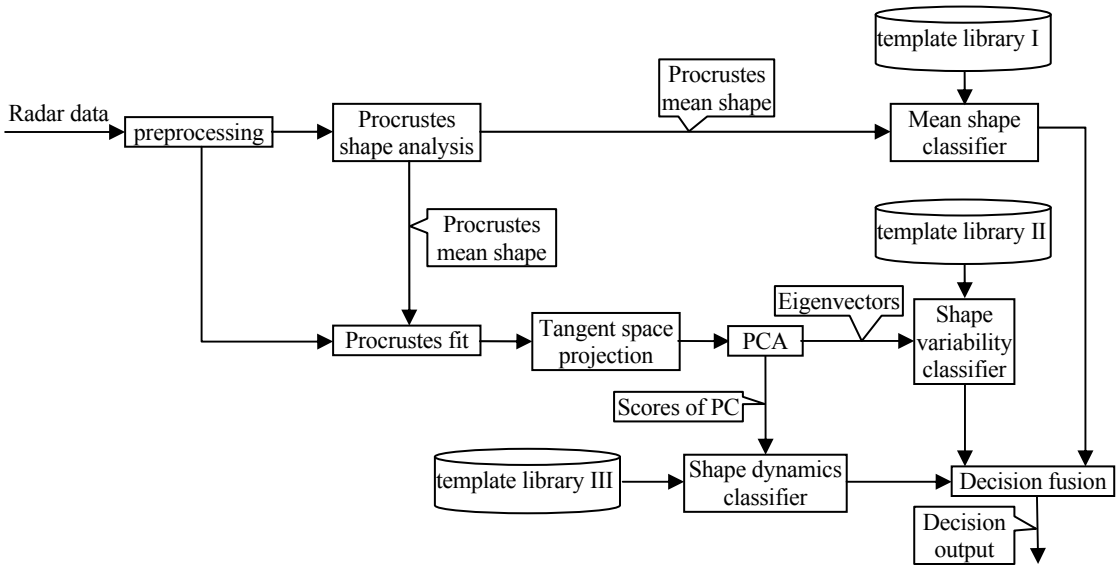


Figure 5.1 Diagrammatic description of the design of an automatic UWB radar target identification system using both the static and dynamic features.

To make this design more robust and efficient, some important open problems should be explored:

(3) As an extension of current fixed-range segmentation and range profile extraction through Gaussian kernel, more solid choices of preprocessing methods will augment the robustness of shape analysis, principal component analysis and dynamic analysis.

(4) The one sample Hotelling T^2 test is an elegant way to make the shape inference in the tangent space. But when there are temporal relations among the sequential radar scans, it remains a mystery how to appropriately set up the parameters of (M, n) for the F statistic (4.31).

(5) Although the similarity between two data variation structures can be measured through comparing their residing subspaces, this comparison is deficient in robustness, because it fails to take into account the relative dominancy of each principal component. This deficiency motivates a need to develop a robust way to measure the similarity of two data variation structures.

(6) This thesis doesn't touch upon dynamic feature selection, but it is a natural extension to include target dynamic signatures as classification clues, because the same people may have different physiological coordination of body parts when in different motion patterns. Time series analysis [Box et al., 1994] will be an appropriate technique to quantify this coordination. Through postulating relationships between variables, time series analysis aims to estimate the coefficients in this relationship and test hypotheses about it. One specific branch of time series modeling techniques, structural time series analysis [Harvey, 1989; Durbin and Koopman, 2001], may help for the human behavior classification. One prominent characteristic of the human gait is its quasi-periodic pattern, which can be captured by the principal component analysis. Structural time

series models are models which are formulated directly in terms of components of interest, such as a cyclic component. It has an intuitive appeal for this application. There are many ways in which such a model may be formulated. For example, one may assume that,

observed time series = trend + cycle + seasonal adjustment + noise.

Structural time series analysis provides a systematic way to model and estimate the components, and test the goodness of fit, which awaits further detailed exploration.

BIBLIOGRAPHY

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, third edition. John Wiley & Sons, Inc.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C. (1994). *Time Series Analysis, Forcasting and Control*, third edition. Prentice-Hall, Inc.

Boyd, J. E. (2001). Video phase-locked loops in gait recognition. *Proceedings of Eighth IEEE International Conference on Computer Vision*, 1:696-703.

Breiman, L., Friedman, J., Stone, C. J., Olshen, R. A. (1984). *Classification and Regression Trees*. Chapman & Hall/CRC.

Cao, S. (2003). *Spike Train Characterization and Decoding for Neural Prosthetic Devices*. Ph.D. thesis, California Institute of Technology.

Chang, C., Lin, C. (2001). LIBSVM : a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cover, T. M., Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley & Sons, Inc.

Devroye, L., Gyorfi, L., Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York.

Donoghue, J. P. (2002). Connecting cortex to machines: recent advances in brain interfaces. *Nature Neuroscience*, 5 Supplement:1085-1088.

Dryen, I. L., Mardia, K. V. (1998). *Statistical Shape Analysis*. John Wiley & Sons Ltd.

Duda, R. O., Hart, P. E., Stork, D. G. (2001). *Pattern Classification*, second edition. John Wiley & Sons, Inc.

Durbin, J., Koopman, S. J. (2001). *Time Series Analysis by State Space Methods*. Oxford University Press.

Federal Communications Commission. (2002). Revision of part 15 of the commission's rules regarding ultra-wideband transmission systems: first report and order. *ET-Docket*: 98-153.

Friedman, J. H., Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, c-23(9):881-890.

Friedman, J. H. (1987). Exploratory projection pursuit. *Journal of the American Statistical Association*, 82(397):249-266.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*, second edition. Academic Press.

Ghavami, M., Michael, L. B., Kohno, R. (2004). *Ultra Wideband Signals and Systems in Communication Engineering*. John Wiley & Sons, Ltd.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417-441,498-520.

Huber, P. J. (1985). Projection pursuit. *The Annals of Statistics*, 13(2):435-475.

Hsu, C., Chang, C., Lin, C. (2007). A practical guide to support vector classification.

Hyvarinen, A. (1998). New approximations of differential entropy for independent component analysis and projection pursuit. In *Advances in Neural Information Processing Systems*, 10:273-279. MIT Press.

Isaacs, R., Weber, D., Schwartz, A. (2000). Work toward real-time control of a cortical prosthesis. *IEEE Trans. Rehabil. Eng*. 8:196-198.

Jolliffe, I. T. (2002). *Principal Component Analysis*, second edition. Springer-Verlag New York, Inc.

Jones, M. C., Sibson, R. (1987). What is projection pursuit. *Journal of the Royal Statistical Society*, series A, 150(1):1-37.

Koch, C. (1999). *Biophysics of Computation, information processing in single neurons*. Oxford University Press.

Krzanowski, W. J. (1979). Between-groups comparison of principal components. *Journal of the American Statistical Association*, 74(367):703-707.

Mallat, S. (1999). *A Wavelet Tour of Signal Processing*, second edition. Academic Press Inc.

Mardia, K. V., Kent, J. T., Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press Inc. (London) Ltd.

Meeker, D., Shenoy, K. V., Cao, S., Pesaran, B., Scherberger, H., Jarvis, M., Buneo, C. A., Batista, A. P., Kureshi, S. A., Mitra, P. P., Burdick, J. W., Andersen, R. A. (2001). Cognitive control signals for prosthetic systems. *Society for Neuroscience*, 27.

Miller, D. J., Yan, L. (1999). Critic-driven ensemble classification. *IEEE Transaction on Signal Processing*, 47(10):2833-2844.

Musallam, S., Corneil, B. D., Greger, B., Scherberger, H., Andersen, R. A. (2004). Cognitive control signals for neural prosthetics. *Science*, 305:258-262.

Nason, G. P. (1992). *Design and Choice of Projection Indices*. Ph.D. thesis, University of Bath.

Nicolelis, M. (2001). Actions from thoughts. *Nature*, 409:403-407.

Nicolelis, M. (2002). The amazing adventures of robotrat. *Trends in Cognitive Sciences*, 6:449-450.

Oppenheim, A. V., Schafer, R. W., Buck, J. R. (1999). *Discrete Time Signal Processing*, second edition. Prentice-Hall, Inc.

Pajunen, P. (1998). *Extensions of Linear Independent Component Analysis: Neural and Information-Theoretic Methods*. Ph.D. thesis, Helsinki University of Technology.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559-572.

Peden, M., Scurfield, R., Sleet, D., Mohan, D., Hyder, A. A., Jarawan, E., Mathers, C., editors. (2004). *World Report on Road Traffic Injury Prevention*. World Health Organization.

Rice, J. A. (1995). *Mathematical Statistics and Data Analysis*, second edition. Duxbury Press.

Richards, M. A. (2005). *Fundamentals of Radar Signal Processing*. McGraw-Hill.

Rieke, F., Warland, D., Steveninck, R., Bialek, W. (1997). *Spikes - Exploring the Neural Code*. Massachusetts Institute of Technology.

Salvador, S., Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11:561-580

Santhanam, G., Ryu, S. I., Yu, B. M., Afshar, A., Shenoy, K. V. (2006). A high-performance brain-computer interface. *Nature*. 442:195-198.

Schwartz, A. B., Moran, D. W. (2000). Arm trajectory and representation of movement processing in motor cortical activity. *European Journal of Neuroscience*, 12(6):1851-1856.

Scott, D. W. (1992). *Multivariate Density Estimation, Theory, Practice, and Visulization*. John Wiley & Sons, Inc.

Shen, X., Guizani, M., Qiu, R. C., Le-Ngoc, T., editors. (2006). *Ultra-wideband Wireless Communications and Networks*. John Wiley & Sons Ltd.

Shenoy, K. V., Meeker, D., Cao, S., Kureshi, S. A., Pesaran, B., Mitra, P., Buneo, C. A., Batista, A. P., Burdick, J. W., Andersen, R. A. (2003). Neural prosthetic control signals from plan activity. *NeuroReport*, 14:591-596.

Shores, T. S. (2006). *Applied Linear Algebra and Matrix Analysis*. Springer.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC.

Strang, G. (2006). *Linear Algebra and Its Applications*, 4th edition. Thomson, Brooks/Cole.

Taylor, J. D. (2001). *Ultra Wideband Radar Technology*. CRC Press LLC.

Theodoridis, S., Koutroumbas, K. (2006). *Pattern Recognition*, third edition. Elsevier (USA).

Turk, M., Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71-86

Wang, L., Ning, H., Tan, T., Hu, W. (2004). Fusion of static and dynamic body biometrics for gait detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(2):149-158.

Wessberg, J., Stambaugh, C., Kralik, J., Beck, P., Laubach, M., Chapin, J., Kim, J., Biggs, S., Srinivasan, M., Nicolelis, M. (2000). Real-time prediction of hand trajectory by emsembles of cortical neurons in primiates. *Nature*, 408(6810):361-365.

Zhu, X. (2001). *Fundamentals of Applied Information Theory*. Tsinghua University Press.

APPENDIX A

KERNEL DENSITY ESITMATION

This appendix states and proves the kernel density estimation theorem, for more broad treatements on the topic of non-parametric density estimation, please see [Silverman, 1986; Scott, 1992], which also generalize the univariate kernel methods to the multivariate case.

Definition A.1: Assume that the random variable $U \sim f(u)$, and $\{U^{(1)}, U^{(2)}, ..., U^{(n)}\}$ is a set of i.i.d. samples from the distribution of U, then the kernel density estimation for $f(u)$ is,

$$\hat{f}(u) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{u - U^{(i)}}{h}) = \frac{1}{n}\sum_{i=1}^{n} K_h(u - U^{(i)})$$

where $K_h(t) = K(t/h)/h$, and the function $K(\cdot)$ satisfies the following properties:

$$\int K(w)dw = 1, \quad \int wK(w)dw = 0, \quad \int w^2 K(w)dw = \sigma_K^2 > 0.$$

Proposition A.1 [Chapter 3, Silverman, 1986]: For a univariate kernel density estimator,

$$bias(u) = E\hat{f}(u) - f(u) = \frac{1}{2}\sigma_K^2 h^2 f''(u) + \mathrm{O}(h^4)$$

$$\mathrm{var}(u) = \mathrm{var}(\hat{f}(u)) = \frac{f(u)R(K)}{nh} - \frac{f^2(u)}{n} + O(\frac{h}{n})$$

$$IMSE = ISB + IV = \frac{R(K)}{nh} + \frac{1}{4}\sigma_K^4 h^4 R(f^{"})$$

where $IMSE$, ISB, and IV are, respectively, abbreviations for integral of the mean square error, integral of square of the bias, and integral of the variance, and $R(f) = \int f^2(u)du$. Moreover, when h takes the value,

$$h^* = \left(\frac{R(K)}{\sigma_K^4 R(f^{"})}\right)^{1/5} n^{-1/5}$$

the minimum value of the integral of mean square error is:

$$IMSE^* = \frac{5}{4}[\sigma_K R(K)]^{4/5} R(f^{"})^{1/5} n^{-4/5}.$$

Proof:

$$\hat{f}(u) = \frac{1}{n}\sum_{i=1}^{n} K_h(u - U^{(i)})$$

$$\Rightarrow \quad E\{\hat{f}(u)\} = EK_h(u-U), \text{ and } \operatorname{var}\{\hat{f}(u)\} = \frac{1}{n}\operatorname{var} K_h(u-U)$$

We have

$$EK_h(u-U) = \int \frac{1}{h}K(\frac{u-t}{h})f(t)dt = \int K(w)f(u-hw)dw$$

$$= f(u)\int K(w) - hf^{'}(u)\int wK(w) + \frac{1}{2}h^2 f^{"}(u)\int w^2 K(w) + ...$$

and

$$\operatorname{var} K_h(u-U) = E\left[\frac{1}{h}K(\frac{u-U}{h})\right]^2 - \left[E\frac{1}{h}K(\frac{u-U}{h})\right]^2$$

where

$$E\left[\frac{1}{h}K(\frac{u-U}{h})\right]^2 = \int \frac{1}{h^2}K\left(\frac{u-t}{h}\right)^2 f(t)dt = \int \frac{1}{h}K(w)^2 f(u-hw)dw \approx \frac{f(u)R(K)}{h}.$$

Therefore, if we denote $R(f) = \int f^2(u)du$,

$$bias(u) = EK_h(u-U) - f(u) = \frac{1}{2}\sigma_K^2 h^2 f''(u) + \mathrm{O}(h^4)$$

$$ISB = \int (EK_h(u-U) - f(u))^2 du = \frac{1}{4}\sigma_K^4 h^4 R(f'') + O(h^6)$$

and

$$\mathrm{var}(u) = \mathrm{var}(\hat{f}(u)) = \frac{f(u)R(K)}{nh} - \frac{f^2(u)}{n} + O(\frac{h}{n})$$

$$IV = \int \mathrm{var}(\hat{f}(u))du = \frac{R(K)}{nh} - \frac{R(f)}{n} + \ldots$$

$$IMSE = ISB + IV = \frac{R(K)}{nh} + \frac{1}{4}\sigma_K^4 h^4 R(f'').$$

Finally, h^* is obtained by setting the derivative of $IMSE$ with respect to h to be 0.

■

APPENDIX B

PROCRUSTES SHAPE ANALYSIS

In a 2D scenario, shape is very commonly used to refer to the appearance or silhouette of an object. Procrustes shape analysis is to analyze the geometrical information that remains when location, scale and rotational effects are filtered out from an object. This appendix briefly reviews the definitions of full Procrustes fit, full Procrustes distance, and full Procrustes mean shape, whose much more complete treatments are in [Dryden and Mardia, 1998].

Assume two shapes or silhouettes in the 2D space are represented by two vectors of k complex entries, say $y=[y_1,\ldots,y_k]^T$ and $w=[w_1,\ldots,w_k]^T$. Without loss of generality, assume these two configurations are centered, i.e., $y^H 1_k = w^H 1_k = 0$, where y^H means transpose of complex conjugate of y and 1_k is a length-k vector with all components being 1.

Definition B.1 [p40, Dryen and Mardia, 1998]: The **full Procrustes fit** of w onto y is

$$w^P = (a+ib)1_k + \beta e^{i\theta} w$$

where (a,b,β,θ) are chosen to minimize

$$D^2(y,w) = \left\| y - (a+ib)1_k - \beta e^{i\theta} w \right\|^2 .$$

Proposition B.1 [p40, Dryen and Mardia, 1998]: The full Procrustes fit has matching parameters

$$a + ib = 0\,, \quad \theta = \arg(w^H y)\,, \quad \beta = (w^H y y^H w)^{1/2} / (w^H w)\,.$$

So the full Procrustes fit of w onto y is

$$w^P = \frac{(w^H y y^H w)^{1/2}}{w^H w} e^{i\angle w^H y} w = \frac{\left| w^H y \right| e^{i\angle w^H y}}{w^H w} w = \frac{w^H y w}{w^H w}\,.$$

Proof:

$$D^2(y,w) = \left\| y - (a+ib)1_k - \beta e^{i\theta} w \right\|^2$$

$$= y^H y + \beta^2 w^H w + k(a^2 + b^2) - \beta e^{i\theta} y^H w - \beta e^{-i\theta} w^H y$$

Obviously, the minimizing $\hat{a}$, $\hat{b}$ are 0. Let us denote $y^H w = re^{i\varphi}$, $r > 0$, then

$$-\beta e^{i\theta} y^H w - \beta e^{-i\theta} w^H y = -r\beta(e^{i(\theta+\varphi)} + e^{-i(\theta+\varphi)}) = -2r\beta\cos(\theta+\varphi)\,.$$

So, $\hat{\theta} = -\varphi = \arg(w^H y)$. Now we have

$$D^2(y,w) = y^H y + \beta^2 w^H w - 2r\beta\,.$$

Lastly, by

$$\frac{\partial D^2}{\partial \beta} = 2\beta w^H w - 2r = 0 \;\Rightarrow\; \hat{\beta} = \frac{r}{w^H w} = \frac{(w^H y y^H w)^{1/2}}{w^H w}\,.$$

■

Definition B.2 [p41, Dryen and Mardia, 1998]: The **full Procrustes distance** between w and y is

$$d_F(w,y) = \inf_{\beta,\theta,a,b} \left\| \frac{y}{\|y\|} - \frac{w}{\|w\|}\beta e^{i\theta} - a - ib \right\| = \left\{ 1 - \frac{y^H w w^H y}{w^H w y^H y} \right\}^{1/2}$$

where the second equation comes from the complex linear regression used in deriving Result 1, and it can be checked that $d_F(w,y)$ is invariant with respect to translation, rotation, and scaling of configurations of w and y.

Definition B.3 [p44, Dryen and Mardia, 1998]: The **full Procrustes mean shape** $[\hat{\mu}]$ is obtained by minimizing the sum of squared full Procrustes distances from each configuration w_i to an unkown unit size configuration μ, i.e.,

$$[\hat{\mu}] = \arg \inf_{\|\mu\|=1} \sum_{i=1}^{n} d_F^2(w_i, \mu).$$

Note that $[\hat{\mu}]$ is not a single configuration, instead, it is a set, whose elements have 0 full Procrustes distance to the optimal unit size configuration μ.

Proposition B.2 [p44, Dryen and Mardia, 1998]: The full Procrustes mean shape, $[\hat{\mu}]$, can be found as the eigenvector, $\hat{\mu}$, corresponding to the largest eigenvalue of the complex sum of squares and products matrix

$$S = \sum_{i=1}^{n} w_i w_i^H / (w_i^H w_i).$$

All translation, scaling, and rotation of $\hat{\mu}$ are also solutions, but they all correspond to the same shape $[\hat{\mu}]$, i.e., have 0 full Procrustes distance to $\hat{\mu}$.

Proof:

$$\sum_{i=1}^{n} d_F^2(w_i, \mu) = \sum_{i=1}^{n} \left\{ 1 - \frac{\mu^H w_i w_i^H \mu}{w_i^H w_i \mu^H \mu} \right\}$$

Under the constraint of $\mu^H \mu = 1$, we have

$$\sum_{i=1}^{n} d_F^2(w_i, \mu) = n - \mu^H \left(\sum_{i=1}^{n} \frac{w_i w_i^H}{w_i^H w_i} \right) \mu = n - \mu^H S \mu$$

$$\Rightarrow \quad \hat{\mu} = \arg \sup_{\|\mu\|=1} \mu^H S \mu .$$

The well known linear algebra fact states that $\hat{\mu}$ is the first eigenvector of matrix S.

■

Proposition B.3 [p89, Dryen and Mardia, 1998]: The arithmetic mean of the full Procrustes fits, $\frac{1}{n}\sum_{i=1}^{n} w_i^P$, has the same shape as the full Procrustes mean, i.e.,

$$d_F\left(\frac{1}{n}\sum_{i=1}^{n} w_i^P, \hat{\mu}\right) = 0 .$$

Proof:

Because $d_F(w_i, \mu)$ is invariant with respect to the scaling of its arguments, so

$$\inf_{\|\mu\|=1} \sum_{i=1}^{n} d_F^2(w_i, \mu) = \inf_{\mu} \sum_{i=1}^{n} d_F^2(w_i, \mu) .$$

Without restricting the size of μ, let us denote the Procrustes fit of w_i onto μ as w_i^P, then

$$\sum_{i=1}^{n} d_F^2(w_i, \mu) = \sum_{i=1}^{n} \left\| \mu - w_i^P \right\|^2$$

is a quadratic function with respect to μ, and is minimized by

$$\mu = \frac{1}{n} \sum_{i=1}^{n} w_i^P \ .$$

■

APPENDIX C

HOTELLING SAMPLE STATISTIC

This appendix derives the one sample Hotelling T^2 tests, whose much more complete treatments are in [Mardia et al., 1979].

Suppose that $X = \{x_1, \ldots, x_n\}$ is a random sample from a population with p.d.f $f(x;\theta)$, where θ is a parameter vector. The likelihood function of sample X is

$$L(X;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

and the corresponding log likelihood function is

$$l(X;\theta) = \sum_{i=1}^{n} \log f(x_i;\theta).$$

As we know, the likelihood (log likelihood) function is central to the theory of statistical inference.

Example C.1 [p97, Mardia et al., 1979]: Suppose $X = \{x_1, \ldots, x_n\}$ is a random sample from $N_p(\mu, \Sigma)$, where $N_p(\mu, \Sigma)$ means p-variate normal distribution with mean, μ, and covariance matrix Σ, then,

$$l(X;\mu,\Sigma) = -\frac{n}{2}\log|2\pi\,\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T\,\Sigma^{-1}(x_i-\mu).$$

Take $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $d = \bar{x} - \mu$, then,

$$(x_i - \mu)^T \Sigma^{-1}(x_i - \mu) = (x_i - \bar{x})^T \Sigma^{-1}(x_i - \bar{x}) + d^T \Sigma^{-1} d + 2d^T \Sigma^{-1}(x_i - \bar{x})$$

$$\sum_{i=1}^{n}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu) = \sum_{i=1}^{n}(x_i - \bar{x})^T \Sigma^{-1}(x_i - \bar{x}) + nd^T \Sigma^{-1} d\,.$$

Writing $nS = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$, we have,

$$\sum_{i=1}^{n}(x_i - \bar{x})^T \Sigma^{-1}(x_i - \bar{x}) = \sum_{i=1}^{n} trace(\Sigma^{-1}(x_i - \bar{x})(x_i - \bar{x})^T) = trace(n\Sigma^{-1} S)\,.$$

So,

$$l(X;\mu,\Sigma) = -\frac{n}{2}\left(\log|2\pi\Sigma| + trace(\Sigma^{-1} S) + d^T \Sigma^{-1} d\right).$$

Proposition C.1 [p103, Mardia et al., 1979]: Suppose $X = \{x_1, \ldots, x_n\}$ is a random sample from $N_p(\mu, \Sigma)$, $\Sigma > 0$, maximum likelihood estimation (MLE) of μ and Σ are

$$\hat{\mu} = \bar{x}\,, \quad \hat{\Sigma} = S\,.$$

Proof:

Using new parameters $d = \bar{x} - \mu$, $V = \Sigma^{-1}$, then the log likelihood function becomes,

$$l(X;\mu,V) = -\frac{n}{2}\left(p\log(2\pi) - \log|V| + trace(VS) + d^T V d\right)$$

$$\frac{\partial l}{\partial d} = -nVd = 0 \;\Rightarrow\; d = 0 \;\Rightarrow\; \hat{\mu} = \bar{x}\,.$$

V is symmetric positive definite, from matrix differentiation we have,

$$\frac{\partial \log|V|}{\partial V} = 2\sum - Diag\sum, \qquad \frac{\partial trace(VS)}{\partial V} = 2S - DiagS$$

$$\frac{\partial d^T V d}{\partial V} = \frac{\partial trace(V dd^T)}{\partial V} = 2dd^T - Diag(dd^T).$$

So, $\frac{\partial l}{\partial V} = \frac{n}{2}(2M - DiagM)$, if we denote $M = \sum - S - dd^T$. Then let $\frac{\partial l}{\partial V} = 0$, we get

$M = 0$, i.e., $\hat{\sum} = S + dd^T = S + (\bar{x} - \hat{\mu})(\bar{x} - \hat{\mu})^T = S$.

■

Definition C.1 [p123, Mardia et al., 1979]: Suppose that $X = \{x_1, \ldots, x_n\}$ is a random sample from a distribution with parameter θ, and $H_0 : \theta \in \Omega_0$ and $H_1 : \theta \in \Omega_1$ are any two hypotheses, then the likelihood ratio (LR) statistic for testing H_0 against H_1 is defined as

$$\lambda(X) = L_0^* / L_1^*$$

where L_i^* is the largest value which the likelihood function takes in region Ω_i, $i \in \{0,1\}$. Equivalently, we can use the statistic,

$$-2\log\lambda = 2(l_1^* - l_0^*), \quad \text{with } l_i^* = \log(L_i^*),\ i \in \{0,1\}.$$

Definition C.2 [p66, Mardia et al., 1979]: If $M(p \times p)$ can be written as $M = XX^T$, where $X(p \times n)$ is a data matrix whose each column is i.i.d. $N_p(0, \sum)$, then M is said to have a **Wishart distribution** with scale matrix $\sum$ and degrees of freedom parameter n; we write $M \sim W_p(\sum, n)$.

Proposition C.2 [p69, Mardia et al., 1979]: Suppose $X = \{x_1, \ldots, x_n\}$ is a random sample from $N_p(\mu, \Sigma)$, $\Sigma > 0$, then the sample mean, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and the sample covariance, $S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$, are independent, and

$$\bar{x} \sim N_p(\mu, n^{-1}\Sigma), \quad nS \sim W_p(\Sigma, n-1).$$

Proof:

Please refer to [p69, Mardia, Kent and Bibby, 1979].

■

Definition C.3 [p74, Mardia et al., 1979]: If α can be written as $nd^T M^{-1} d$ where d and M are independently distributed as $N_p(0, I)$ and $W_p(I, n)$, then we say that α has the **Hotelling T^2 distribution** with parameters p and n. We write $\alpha \sim T^2(p, n)$.

Proposition C.3 [p74, Mardia et al., 1979]: If x and M are independently distributed as $N_p(\mu, \Sigma)$ and $W_p(\Sigma, n)$, then

$$n(x-\mu)^T M^{-1}(x-\mu) \sim T^2(p, n).$$

Proof:

Let $d^* = \Sigma^{-1/2}(x-\mu)$ and $M^* = \Sigma^{-1/2} M \Sigma^{-1/2}$, then d^*, M^* satisfy the requirements in Definition C.3.

■

Proposition C.4 [p74, Mardia et al., 1979]: If $\bar{x}$ and S are the sample mean and sample covariance matrix of a sample of size n from $N_p(\mu, \Sigma)$, then

$$(n-1)(\bar{x}-\mu)^T S^{-1}(\bar{x}-\mu) \sim T^2(p, n-1).$$

Proof:

$x_i \sim N_p(\mu, \Sigma) \Rightarrow \bar{x} \sim N_p(\mu, n^{-1}\Sigma)$, $nS \sim W_p(\Sigma, n-1)$

Take $d = n^{1/2}(\bar{x}-\mu)$, $M = nS$, then $d \sim N(0, \Sigma)$. From Proposition C.3, we got

$$(n-1)(\bar{x}-\mu)^T S^{-1}(\bar{x}-\mu) = (n-1)d^T M^{-1} d \sim T^2(p, n-1).$$

■

Proposition C.5 [p74, Mardia et al., 1979]

$$T^2(p,n) = \frac{np}{n-p+1} F_{p,n-p+1}.$$

Proof:

Please refer to [p74, Mardia et al., 1979].

■

Proposition C.6 [p75, Mardia et al., 1979]: If $\bar{x}$ and S are the sample mean and sample covariance matrix of a sample of size n from $N_p(\mu, \Sigma)$, then

$$\{(n-p)/p\}(\bar{x}-\mu)^T S^{-1}(\bar{x}-\mu) \sim F_{p,n-p}.$$

Proof:

This result directly follows from Proposition C.5.

■

Proposition C.7 [p125, Mardia et al., 1979]: Suppose $X = \{x_1, \ldots, x_n\}$ is a random sample from $N_p(\mu, \Sigma)$, $\Sigma > 0$ is fixed but unknown, then the LR statistic for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ is known as one sample Hotelling T^2 statistic.

Proof:

Under H_0, $\hat{\mu} = \mu_0$ and $\hat{\Sigma} = S + dd^T$, with $d = \bar{x} - \mu_0$; under H_1, $\hat{\mu} = \bar{x}$ and $\hat{\Sigma} = S$. So,

$$l_0^* = -\frac{n}{2}\left(p\log(2\pi) + \log\left|S + dd^T\right| + trace[(S + dd^T)^{-1}S] + d^T(S + dd^T)^{-1}d\right).$$

By

$$\log\left|S + dd^T\right| = \log|S| + \log\left|I + S^{-1}dd^T\right| = \log|S| + \log(1 + d^T S^{-1}d)$$

$$trace[(S + dd^T)^{-1}S] + d^T(S + dd^T)^{-1}d = trace[(S + dd^T)^{-1}S + (S + dd^T)^{-1}dd^T] = p$$

$$\Rightarrow \quad l_0^* = -\frac{n}{2}\left(p\log(2\pi) + \log|S| + \log(1 + d^T S^{-1}d) + p\right).$$

Obviously, $l_1^* = -\frac{n}{2}\left(p\log(2\pi) + \log|S| + p\right)$, so

$$-2\log\lambda = 2(l_1^* - l_0^*) = n\log(1 + d^T S^{-1}d).$$

From Results 4 and 6, we got that,

$$(n-1)d^T S^{-1}d \sim T^2(p, n-1) \quad \text{or} \quad \{(n-p)/p\}d^T S^{-1}d \sim F_{p,n-p}.$$

■

APPENDIX D

PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) is a statistical technique that can be constructed by several ways, one commonly cited of which is stated in this appendix. By stating a few directly useful properties of PCA for radar signal analysis, we by no means, tend to give an even superficial survey of this ever-growing topic. For deeper and more complete coverage of PCA and its applications, please refer to [Jolliffe, 2002]. [p266~p272, Theodoridis and Koutroumbas, 2006; Chapter 8, Mardia et al., 1979; Chapter 11, Anderson, 2003] are also nice and shorter materials to explain some general properties of PCA. For simplifying the presentation, all the following properties of PCA are proved under the assumption that all eigenvalues of whichever covariance matrix concerned are positive and distinct.

One PCA construction: Assume a random vector X, taking values in $\Re^m$, has a mean and covariance matrix of μ_X and Σ_X, respectively. $\lambda_1 > \lambda_2 > \cdots > \lambda_m > 0$ are ordered eigenvalues of Σ_X, such that the i-th eigenvalue of Σ_X means the i-th largest of them. Similarly, a vector α_i is the i-th eigenvector of Σ_X when it corresponds to the i-th eigenvalue of Σ_X. To derive the form of principal components (PCs), consider the optimization problem of maximizing $\text{var}[\alpha_1^T X] = \alpha_1^T \Sigma_X \alpha_1$, subject to $\alpha_1^T \alpha_1 = 1$. The

Lagrange multiplier method is used to solve this question.

$$L(\alpha_1, \phi_1) = \alpha_1^T \Sigma_X \alpha_1 + \phi_1(\alpha_1^T \alpha_1 - 1)$$

$$\frac{\partial L}{\partial \alpha_1} = 2\Sigma_X \alpha_1 + 2\phi_1 \alpha_1 = 0 \Rightarrow \Sigma_X \alpha_1 = -\phi_1 \alpha_1 \Rightarrow \text{var}[\alpha_1^T X] = -\phi_1 \alpha_1^T \alpha_1 = -\phi_1 .$$

Because $-\phi_1$ is the eigenvalue of Σ_X, with α_1 being the corresponding normalized eigenvector, $\text{var}[\alpha_1^T X]$ is maximized by choosing α_1 to be the first eigenvector of Σ_X. In this case, $z_1 = \alpha_1^T X$ is named the first PC of X, α_1 is the vector of coefficients for z_1, and $\text{var}(z_1) = \lambda_1$.

To find the second PC, $z_2 = \alpha_2^T X$, we need to maximize $\text{var}[\alpha_2^T X] = \alpha_2^T \Sigma_X \alpha_2$ subject to z_2 being uncorrelated with z_1. Because $\text{cov}(\alpha_1^T X, \alpha_2^T X) = 0 \Rightarrow \alpha_1^T \Sigma_X \alpha_2 = 0 \Rightarrow \alpha_1^T \alpha_2 = 0$, this problem is equivalently set as maximizing $\alpha_2^T \Sigma_X \alpha_2$, subject to $\alpha_1^T \alpha_2 = 0$, and $\alpha_2^T \alpha_2 = 1$. We still make use of the Lagrange multiplier method.

$$L(\alpha_2, \phi_1, \phi_2) = \alpha_2^T \Sigma_X \alpha_2 + \phi_1 \alpha_1^T \alpha_2 + \phi_2(\alpha_2^T \alpha_2 - 1)$$

$$\frac{\partial L}{\partial \alpha_2} = 2\Sigma_X \alpha_2 + \phi_1 \alpha_1 + 2\phi_2 \alpha_2 = 0$$

$$\Rightarrow \alpha_1^T (2\Sigma_X \alpha_2 + \phi_1 \alpha_1 + 2\phi_2 \alpha_2) = 0 \Rightarrow \phi_1 = 0$$

$$\Rightarrow \Sigma_X \alpha_2 = -\phi_2 \alpha_2 \Rightarrow \alpha_2^T \Sigma_X \alpha_2 = -\phi_2 .$$

Because $-\phi_2$ is the eigenvalue of Σ_X, with α_2 being the corresponding normalized eigenvector, $\text{var}[\alpha_2^T X]$ is maximized by choosing α_2 to be the second eigenvector of

$\sum_X$. In this case, $z_2 = \alpha_2^T X$ is named the second PC of X , α_2 is the vector of coefficients for z_2, and $\text{var}(z_2) = \lambda_2$. Continuing in this way, it can be shown that the i-th PC $z_i = \alpha_i^T X$ is constructed by selecting α_i to be the i-th eigenvector of $\sum_X$, and has variance of λ_i. The key result in regards to PCA is that the principal components are the only set of linear functions of original data that are uncorrelated and have orthogonal vectors of coefficients.

Proposition D.1 [Jolliffe, 2002]: For any positive integer $p \le m$, let $B = [\beta_1, \beta_2, \ldots, \beta_p]$ be an real $m \times p$ matrix with orthonormal columns, i.e., $\beta_i^T \beta_j = \delta_{ij}$, and $Y = B^T X$. Then the trace of covariance matrix of Y is maximized by taking $B = [\alpha_1, \alpha_2, \ldots, \alpha_p]$, where α_i is the i-th eigenvector of $\sum_X$.

Proof:

Because $\sum_X$ is symmetric with all distinct eigenvalues, so $\{\alpha_1, \alpha_2, \ldots, \alpha_m\}$ is an orthonormal basis with α_i being the i-th eigenvector of $\sum_X$, and we can represent the columns of B as

$$\beta_i = \sum_{j=1}^{m} c_{ji} \alpha_j , \quad i = 1, \ldots, p .$$

So we have

$$B = PC$$

where $P=[\alpha_1,\ldots,\alpha_m]$, $C=\{c_{ij}\}$ is an $m\times p$ matrix. Then, $P^T\sum_X P=\Lambda$, with Λ being a diagonal matrix whose k-th diagonal element is λ_k, and the covariance matrix of Y is,

$$\sum_Y = B^T\sum_X B = C^T P^T\sum_X PC = C^T\Lambda C = \lambda_1 c_1 c_1^T + \cdots + \lambda_m c_m c_m^T$$

where c_i^T is the i-th row of C. So,

$$trace(\sum_Y) = \sum_{i=1}^{m}\lambda_i trace(c_i c_i^T) = \sum_{i=1}^{m}\lambda_i trace(c_i^T c_i) = \sum_{i=1}^{m}\lambda_i c_i^T c_i = \sum_{i=1}^{m}(\sum_{j=1}^{p} c_{ij}^2)\lambda_i .$$

Because $C^TC = B^TPP^TB = B^TB = I$, so $trace(C^TC)=\sum_{i=1}^{m}\sum_{j=1}^{p} c_{ij}^2 = p$, and the columns of C are orthonormal. By the Gram-Schmidt method, C can expand to D, such that D has its columns as an orthonormal basis of $\Re^m$ and contains C as its first p columns. D is square shape, thus being an orthogonal matrix and having its rows as another orthonormal basis of $\Re^m$. One row of C is a part of one row of D, so $\sum_{j=1}^{p} c_{ij}^2 \le 1$, $i=1,\ldots,m$.

Considering the constraints $\sum_{j=1}^{p} c_{ij}^2 \le 1$, $\sum_{i=1}^{m}\sum_{j=1}^{p} c_{ij}^2 = p$ and the objective $\sum_{i=1}^{m}(\sum_{j=1}^{p} c_{ij}^2)\lambda_i$. We derive that $trace(\sum_Y)$ is maximized if $\sum_{j=1}^{p} c_{ij}^2 = 1$ for $i=1,\ldots,p$, and $\sum_{j=1}^{p} c_{ij}^2 = 0$ for $i=p+1,\ldots,m$. When $B=[\alpha_1,\alpha_2,\ldots,\alpha_p]$, straightforward calculation yields that C is an all-zero matrix except $c_{ii}=1$, $i=1,\ldots,p$. This fulfills the maximization condition. Actually, by taking $B=[\gamma_1,\gamma_2,\ldots,\gamma_p]$, where $\{\gamma_1,\gamma_2,\ldots,\gamma_p\}$ is any orthonormal basis of

the subspace of $span\{\alpha_1, \alpha_2, \ldots, \alpha_p\}$, the maximization condition is also satisfied, thus yielding the same trace of covariance matrix of Y.

■

Proposition D.2 [Jolliffe, 2002]: Suppose that we wish to approximate the random vector X by its projection onto a subspace spanned by columns of B, where $B = [\beta_1, \beta_2, \ldots, \beta_p]$ is a real $m \times p$ matrix with orthonormal columns, i.e., $\beta_i^T \beta_j = \delta_{ij}$. If σ_i^2 is the residual variance for each component of X, then $\sum_{i=1}^{m} \sigma_i^2$ is minimized if $B = [\alpha_1, \alpha_2, \ldots, \alpha_p]$, where $\{\alpha_1, \alpha_2, \ldots, \alpha_p\}$ are the first p eigenvectors of Σ_X. In other words, the trace of convariance matrix of $X - BB^T X$ is minimized if $B = [\alpha_1, \alpha_2, \ldots, \alpha_p]$. When $E(X) = 0$, which is a commonly applied preprocessing step in data analysis methods, this property is saying that $E\left\|X - BB^T X\right\|^2$ is minimized if $B = [\alpha_1, \alpha_2, \ldots, \alpha_p]$.

Proof:

The projection of a random vector X onto a subspace spanned by columns of B is $\hat{X} = BB^T X$. Then the residual vector is $\varepsilon = X - BB^T X$, which has a convariance matrix

$$\Sigma_\varepsilon = (I - BB^T)\Sigma_X (I - BB^T).$$

Then,

$$\sum_{i=1}^{m} \sigma_i^2 = trace(\Sigma_\varepsilon) = trace(\Sigma_X - \Sigma_X BB^T - BB^T \Sigma_X + BB^T \Sigma_X BB^T).$$

Also, we know

$$trace(\sum_X BB^T) = trace(BB^T \sum_X) = trace(B^T \sum_X B)$$

$$trace(BB^T \sum_X BB^T) = trace(B^T \sum_X BB^T B) = trace(B^T \sum_X B).$$

The last equation comes from the fact that B has orthonormal columns.

So,

$$\sum_{i=1}^{m} \sigma_i^2 = trace(\sum_X) - trace(B^T \sum_X B).$$

To minimize $\sum_{i=1}^{m} \sigma_i^2$, it suffices to maximize $trace(B^T \sum_X B)$. This can be done by choosing $B = [\alpha_1, \alpha_2, \ldots, \alpha_p]$, where $\{\alpha_1, \alpha_2, \ldots, \alpha_p\}$ are the first p eigenvectors of $\sum_X$, according to Proposition D.1 stated above.

■

APPENDIX E

COMPARISON OF PRINCIPAL COMPONENTS

Assume two random vectors, $X \in \Re^m$ and $Y \in \Re^m$, have the covariance matrices of $\sum_X$ and $\sum_Y$, respectively. Suppose the coefficients for the first p principal components (PCs) of X and Y are $\{\alpha_1,\ldots,\alpha_p\}$ and $\{\beta_1,\ldots,\beta_p\}$, respectively, that is, $\{\alpha_1,\ldots,\alpha_p\}$ are the first p orthonormal eigenvectors of $\sum_X$, and $\{\beta_1,\ldots,\beta_p\}$ are the first p orthonormal eigenvectors of $\sum_Y$. In order to compare these two sets of PCs, it is necessary to compare the two subspaces spanned by $\{\alpha_1,\ldots,\alpha_p\}$ and $\{\beta_1,\ldots,\beta_p\}$, respectively. The following two theorems in [Krzanowski, 1979] propose one rigorous way to analyze it.

Proposition E.1 [Krzanowski, 1979]: Denote $L=[\alpha_1,\ldots,\alpha_p]$ and $M=[\beta_1,\ldots,\beta_p]$, the minimum angle between an arbitrary vector in the subspace $span\{\alpha_1,\ldots,\alpha_p\}$ and another arbitrary vector in the subspace $span\{\beta_1,\ldots,\beta_p\}$ is given by $\cos^{-1}(\sqrt{\lambda_1})$, where λ_1 is the largest eigenvalue of $K = L^T MM^T L$.

Proof:

Arbitrarily select one vector from $span\{\alpha_1,\ldots,\alpha_p\}$, and represent it as $w_1 = L\nu_1$, then the projection of w_1 onto $span\{\beta_1,\ldots,\beta_p\}$ is given by $MM^T w_1$. Due to the geometry

property of projection, to find the minimum angle between two arbitrary vectors in $span\{\alpha_1,\ldots,\alpha_p\}$ and $span\{\beta_1,\ldots,\beta_p\}$ is to find w_1, such that the angle between w_1 and $MM^T w_1$, say θ_1, is minimal.

From $\cos(\theta_1) = \dfrac{w_1^T MM^T w_1}{\|w_1\|\|MM^T w_1\|}$, and $L^T L = M^T M = I \Rightarrow \cos^2(\theta_1) = \dfrac{v_1^T L^T MM^T Lv_1}{v_1^T v_1}$.

So minimizing θ_1 is equivalent to maximizing $v_1^T L^T MM^T Lv_1$, subject to $v_1^T v_1 = 1$. The closed form solution from the well known linear algebra fact is given by:

When v_1 is the first eigenvector of $K = L^T MM^T L$, $\cos^2(\theta_1) = \lambda_1$, which is the largest eigenvalue of K.

One point to note is that all eigenvalues, λ_i, of K satisfy $0 \le \lambda_i \le 1$, which is verified by:

$L^T MM^T Lx = \lambda x \Rightarrow x^T L^T MM^T Lx = \lambda x^T x \Rightarrow y^T MM^T y = \lambda x^T x$, where $y = Lx$.

$M^T y$ is the projection coefficients of y onto $span\{\beta_1,\ldots,\beta_p\}$, with $\{\beta_1,\ldots,\beta_p\}$ being an orthonormal set, so

$$\|M^T y\| \le \|y\| \Rightarrow \lambda x^T x = y^T MM^T y = \|M^T y\|^2 \le \|y\|^2 = x^T L^T Lx = x^T x \Rightarrow \lambda \le 1$$

■

Proposition E.2 [Krzanowski, 1979]: L, M, and K are defined as in Proposition E.1, Let λ_i, v_i be the i-th largest eigenvalue and corresponding eigenvector of K, respectively. Take $w_i = Lv_i$, then $\{w_1,\ldots,w_p\}$ and $\{MM^T w_1,\ldots,MM^T w_p\}$ form orthogonal vectors in $span\{\alpha_1,\ldots,\alpha_p\}$ and $span\{\beta_1,\ldots,\beta_p\}$, respectively. The angle between the i-th pair w_i,

$MM^T w_i$ is given by $\cos^{-1}(\sqrt{\lambda_i})$. Proposition E.1 shows that w_1 and $MM^T w_1$ give the two closest vectors when one is constrained to be in the subspace $span\{\alpha_1,\ldots,\alpha_p\}$ and the other in $span\{\beta_1,\ldots,\beta_p\}$. It follows that w_2 and $MM^T w_2$ give directions, orthogonal to the previous ones, between which lies the next smallest angle between the subspaces.

Proof:

Arbitrarily select one vector from $span\{\alpha_1,\ldots,\alpha_p\}$, which is orthogonal to w_1, and represent it as $w_2 = L v_2$, then the projection of w_2 onto $span\{\beta_1,\ldots,\beta_p\}$ is given by $MM^T w_2$. The angle between w_2 and $MM^T w_2$, say θ_2, satisfies

$$\cos^2(\theta_2) = \frac{v_2^T L^T MM^T L v_2}{v_2^T v_2}.$$

So we need to maximize $v_2^T L^T MM^T L v_2$, subject to $v_2^T v_2 = 1$, and $v_2^T v_1 = 0$. By the same Lagrange multiplier method we use for PCA construction in [Appendix B], it turns out that the optimal solution is

$$\cos^2(\theta_2) = \lambda_2, \text{ when } v_2 \text{ is the second eigenvector of } K = L^T MM^T L.$$

It is also true that $w_1 \perp w_2$, because $w_1^T w_2 = v_1^T v_2 = 0$, and that $MM^T w_1 \perp MM^T w_2$, because $w_2^T MM^T MM^T w_1 = w_2^T MM^T w_1 = v_2^T L^T MM^T L v_1 = \lambda_1 v_2^T v_1 = 0$.

Continuing in this way, the conclusion of this theorem is reached.

■

Let θ_{ij} be the angle between α_i and β_j, i.e., $\cos(\theta_{ij}) = \alpha_i^T \beta_j$, then $L^T M = \{\cos(\theta_{ij})\}$, so we have

$$\sum_{i=1}^{p}\lambda_i = trace(L^T MM^T L) = \sum_{j=1}^{p}\sum_{i=1}^{p}\cos^2(\theta_{ij}).$$

Thus the summation of the eigenvalues of $K = L^T MM^T L$ equals the sum of squares of the cosines of the angles between each basis element of $span\{\alpha_1,\ldots,\alpha_p\}$ and each basis element of $span\{\beta_1,\ldots,\beta_p\}$. This sum is invariant with respect to whichever basis you select for $span\{\alpha_1,\ldots,\alpha_p\}$ and $span\{\beta_1,\ldots,\beta_p\}$. In more detail, let

$$\widetilde{L} = [\widetilde{\alpha}_1,\ldots,\widetilde{\alpha}_p] = [\alpha_1,\ldots,\alpha_p]P = LP, \text{ and } \widetilde{M} = [\widetilde{\beta}_1,\ldots,\widetilde{\beta}_p] = [\beta_1,\ldots,\beta_p]Q = MQ,$$

where P, Q are $p\times p$ orthogonal matrices, i.e., $P^T P = PP^T = I$, and $Q^T Q = QQ^T = I$. If $\widetilde{\theta}_{ij}$ is the angle between $\widetilde{\alpha}_i$ and $\widetilde{\beta}_j$, then

$$\sum_{j=1}^{p}\sum_{i=1}^{p}\cos^2(\widetilde{\theta}_{ij}) = trace(\widetilde{L}^T\widetilde{M}\widetilde{M}^T\widetilde{L}) = trace(P^T L^T MQQ^T M^T LP)$$

$$= trace(P^T L^T MM^T LP) = trace(M^T LPP^T L^T M)$$

$$= trace(M^T LL^T M) = trace(L^T MM^T L) = \sum_{j=1}^{p}\sum_{i=1}^{p}\cos^2(\theta_{ij}).$$

So, this sum can be used as a measure of total similarity between the two subspaces. It can be checked that if $span\{\alpha_1,\ldots,\alpha_p\} = span\{\beta_1,\ldots,\beta_p\}$, $\sum_{i=1}^{p}\lambda_i = p$, and if $span\{\alpha_1,\ldots,\alpha_p\} \perp span\{\beta_1,\ldots,\beta_p\}$, $\sum_{i=1}^{p}\lambda_i = 0$.