5-24-2012

Information Theoretic Methods For Biometrics, Clustering, And Stemmatology

Po-Hsiang Lai
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

School of Engineering and Applied Science

Department of Electrical and Systems Engineering

Thesis Examination Committee:
Joseph A. O'Sullivan, Chair
R. Martin Arthur
Jr-Shin Li
Robert Pless
John W. Rohrbaugh
Heinz Schaettler

INFORMATION THEORETIC METHODS FOR BIOMETRICS, CLUSTERING, AND STEMMATOLOGY

by

Po-Hsiang Lai

A dissertation presented to the School of Engineering
of Washington University in partial fulfillment of the
requirements for the degree of

DOCTOR OF SCIENCE

May 2012
Saint Louis, Missouri

ABSTRACT OF THE THESIS

Information Theoretic Methods for Biometrics, Clustering, and Stemmatology

by

Po-Hsiang Lai

Doctor of Science in Information Theory

Washington University in St. Louis, 2012

Research Advisor: Professor O'Sullivan

This thesis consists of four parts, three of which study issues related to theories and applications of biometric systems, and one which focuses on clustering.

We establish an information theoretic framework and the fundamental trade-off between utility of biometric systems and security of biometric systems. The utility includes person identification and secret binding, while template protection, privacy, and secrecy leakage are security issues addressed. A general model of biometric systems is proposed, in which secret binding and the use of passwords are incorporated. The system model captures major biometric system designs including biometric cryptosystems, cancelable biometrics, secret binding and secret generating systems, and salt biometric systems. In addition to attacks at the database, information leakage from communication links between sensor modules and databases is considered. A general information theoretic rate outer bound is derived for characterizing and comparing the fundamental capacity, and security risks and benefits of different system designs.

We establish connections between linear codes to biometric systems, so that one can directly use a vast literature of coding theories of various noise and source random processes to achieve good performance in biometric systems.

We develop two biometrics based on laser Doppler vibrometry (LDV) signals and electrocardiogram (ECG) signals. For both cases, changes in statistics of biometric traits of the same individual is the major challenge which obstructs many methods from producing satisfactory results. We propose a

robust feature selection method that specifically accounts for changes in statistics. The method yields the best results both in LDV and ECG biometrics in terms of equal error rates in authentication scenarios.

Finally, we address a different kind of learning problem from data called clustering. Instead of having a set of training data with true labels known as in identification problems, we study the problem of grouping data points without labels given, and its application to computational stemmatology. Since the problem itself has no "true" answer, the problem is in general ill-posed unless some regularization or norm is set to define the quality of a partition. We propose the use of minimum description length (MDL) principle for graphical based clustering. In the MDL framework, each data partitioning is viewed as a description of the data points, and the description that minimizes the total amount of bits to describe the data points and the model itself is considered the best model. We show that in synthesized data the MDL clustering works well and fits natural intuition of how data should be clustered. Furthermore, we developed a computational stemmatology method based on MDL, which achieves the best performance level in a large dataset.

Acknowledgments

I thank my long lasting academic advisor Professor Joseph A. O'Sullivan. In the past nine years, I was able to learn a wide range of exciting and important topics rigorously, to draft and carry out novel research ideas, and leading research seminars under the guidance and freedom provided by him. In addition, I was fortunate to work and interact extensively with a number of outstanding researchers: Dr. Robert Pless and Dr. Steve Smith in clustering projects, Dr. John Rohrbaugh and Dr. Erik Sirevaag in biometric projects, Dr. Kilian Q. Weinberger in machine learning discussions, and and Dr. Jr-Shin Li in system science. Each of them has brought me great experiences and inspirations. All of these cultivate me to be a scholar.

My deep appreciation also goes to my dissertation and qualify committee for their time and effort: Dr. Joseph A. O'Sullivan, Dr. Martin Arthur, Dr. Jr-Shin Li, Dr. Hiro Mukai, Dr. R. Robert Pless, Dr. John Rohrbaugh, and Dr. Heinz Schaettler.

I would like to thank Dr. Teemu Roos and Dr. Petri Myllymäki for their interests in my work and introducing me the computational stemmatology challenge, which I was grateful to contribute my effort in the area.

There are a number of students and colleagues with whom I worked and shared ideas, and from whom I learned a lot. A partial list includes Dr. Brandon Westover, Dr. Naveen Singla, Dr. Mei Chen, Dr. Alan D. Kaplan, Dr. Sean Kristjansson, Mr. Ikenna Odinaka, Mr. Michael Walker, Mr. Kenji Truman, and Ms. Amanda K. Sheffield.

I thank the tremendous support from my family: Wang Sun, Lai Zheng-Chi, Lai PoLin.

To those who sharing life with me, you have my special appreciation and thanks: Chen Eva, Pîrvu Oana Marina, Wang Tze-Hua, Wu Ching-Yi, Zhan Jiening, Zheng Ya-Jian, and the pathological optimists from HSNU.

Po-Hsiang Lai

Washington University in Saint Louis
May 2012

Dedicated to you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This dissertation consists of two themes. One theme concerns the theories and designs of identification systems and security issues around such systems, with a major application focus on biometrics. Template security and robustness against changes in data statistics are two unique challenges in biometric identification systems. The other theme is on the problem of automatically partitioning objects or data points, and the application is on reconstructing relations among ancient text documents, known as computational stemmatology. Two key challenges in computational stemmatology are missing text documents which are not discovered by historians, and damages in available documents such as missing or broken pages.

In this introduction, I outline the key concepts and problems involved in each theme, as well as the contributions in each chapter. On the other hand, there are a number of interesting real life situations that are in fact closely related to the problems studied in this dissertation. By presenting a small selection of them here, one may find it more inspiring or motivating to appreciate the results, even though one has other topics of interest.

1.1 Identification in Real Life and Biometrics

One important goal of biometrics is to recognize humans based on physical and behavioral characteristics. Classical biometrics includes fingerprint, face, voice, and iris. In most cases, the problem

of interest is either "who is this person?" or "Is this person genuine or an impostor?". Both cases are under the category of supervised learning problems.

Supervised learning is a major form of inference and a major research area across many fields from engineering, statistics, to neuroscience. Identification, recognition, classification, and hypothesis testing are examples of supervised learning. The abstract problem goes like the following. A supervised learning problem consists of two stages, the training phase and the testing phase. In the training phase, a learner, a machine or a living organism, is given a set of training data points, i.e. examples, (x_i, y_i), where x_i denotes the observation of the ith data point, and y_i is the true class label taking values from a discrete set, i.e. the names of different classes, provided by a supervisor or reliable past experiences. The index i goes from 1 to n where n is the number of training data. In the testing phase, the learner obtains new testing data points x_t associated with an unobserved label y_t where the relation, characterized by the joint distribution $p(x, y)$ is assumed to be the same as the examples. The task of the learner is to make a prediction $\hat{y}_t$ about the unobserved label. The learner makes an error if $\hat{y}_t \neq y_t$. A special case of identification problem is when the labels take only two values, such as "yes" and "no", or "accept" and "reject". This is called verification in biometrics or binary hypothesis testing in statistics.

There are abundant identification problems in life. In biometrics, the x_i may be a fingerprint image, a face image, a voice recording, or a heartbeat signal, and y_i is the identity or name of the person, or an indicator for whether this person is genuine or an impostor. Also, in the problem of recognizing words written in a document, the x_i here is the image perceived in one's eyes and y_i is the actual word. In identifying a speaker's voice, x_i is sound waves received by the ears, and y_i is the name of the speaker. In spam email filtering, x_i is the content, the timing, the sender, and the receivers of the email and y_i is "spam" and "not spam." In suggesting products to consumers, x_i may be the shopping behavior, living style and so on, and y_i is the product that the person may purchase.

In most machine learning systems or the brain, the x_i is processed and transformed into another domain. The transformed data in this domain is denoted as f_i, where the decision is taken place. The goals are to extract and select relevant and reliable information and to remove redundancy such that decisions can be made with higher accuracy or efficiency. Then a decision mechanism operates on f_i to predict y_i. In fingerprint recognition, the features may be the ending or bifurcation locations

of ridges. For visual object recognition, there are neurons called simple cells in the early visual cortex that react to specific simple patterns such as ring patterns with red inside and green outside at specific locations of the visual field. The information is then forwarded to higher level neurons which react to highly complex patterns such as faces. Such biologically motivated features can be used in face recognition systems. In real life identification problem such as choosing a relationship, the features may be a list of personal characteristics, behaviors, locations and so on. The performance of an identification system depends on both good feature extraction and selection, and the decision mechanism based on the features.

1.1.1 Evaluating Quality of Data and Learners

There is a value, or cost, associated with each correct inference and error made by the learner. Some error leads to much higher loss than others. For example, misrecognizing a puma as a cat may be disastrous, while misrecognizing a cat as a puma leads to less harm. In biometrics, letting an impostor get into an important facility usually bears a higher cost relative to rejecting and further questioning a genuine user. On the other hand, there is also a probability associated with each event the learner should also consider. Certain facilities are less targeted by attackers than others and so overly complicated screening procedures in fact are not necessarily justifiable. The goal of a learner is to maximize the total expected value, or minimize the expected cost. The expectation is taken over the underlying distribution of events weighted by the associated value and cost, while this distribution is in general not available. The evaluation of a learner is usually done by testing the learner on an unseen dataset with the same or similar properties as the training dataset.

In verification or binary hypothesis testing cases, if the learner rejects a data point that it should accept, it is called a false reject error, false non-match, false positive, or type I error. On the other hand, if the learner accepts a data point that it should reject, it is called a false accept error, false match, false negative, or type II error. The trade-off between these two types of error is critical to system design.

It is critical to understand the fundamental properties of the identification problem, from the data to the operational requirements and constraints before evaluation of a learner. In many problems,

the performance of a learner is said to be good in a relative sense, comparing to fundamental performance limits posed by the problem and the data. However, in problems such as biometrics, a specified level of performance must be met. Thus the data itself must allow the possibility of a good learner achieving the desired performance. It is often impossible to have any learner perform properly for a given set of data. For example, height measured at the accuracy of a centimeter may not be applicable to distinguish several thousands of individuals. Hence height alone is not a good biometric. In the following, I list four critical properties that an identification problem should have for a learner achieving a specified performance level to exist. In addition to these four properties, a biometric trait also needs to be universal in that every individual should have it.

1 **Uniqueness of each label class given the measurements** A good biometric trait must lead to data that distinguishes one individual from another with high probability. There are some problems in which the data is fundamentally impossible for any learner to achieve a desired performance because some classes are indistinguishable with high probability.

2 **Stability or manageable changes** In order for a learner to perform well, the statistical properties of data from one measurement occasion to another must stay the same, or change in a manageable manner. For example, the face of an individual changes across ages, while the changes are small in a short time frame and updates of new images are not required frequently. On the other hand, some changes are more rapid and harder to update constantly. The later case is discussed more in the next section and Chapter 3.

3 **Measurability and acceptability** The data must be measurable within the available resources and external constraints of the learner. For example, a DNA sequence is very unique and stable, but it is not measurable in a short enough period of time for access control of an entrance point or an ATM machine. In daily applications, the use of DNA for identification is also not acceptable to most individuals, because it has aspects of invasion and leakage of privacy.

4 **Circumvention** The data properties of different labels have to be hard to fake or mimic for an impostor. Artificial fingerprints or mimicking someone else's voice are attacks that threaten the security and usability of some biometrics.

It is worth noting that except measurability and acceptability, the other three properties are essential for selected features also. However, one should not confuse the requirements of a *set* of selected features as the requirements of a *single* feature. For example, height alone may not be enough to identify the gender of an individual, while combining height with other body measurements, which separately are also insufficient, can lead to high accuracy of gender identification.

In addition, there are three other aspects that are important to biometric systems.

- **Template security** The data stored in a biometric system must be protected. Unlike the classical encryption of a password, the biometric trait of an individual is hard to replace. Thus the data security, or more specifically template security, in biometrics focuses on the condition that if the database is compromised, there must be only a limited amount of information leaked. The information leakage may due to a database being compromised, or a communication link between the sensor and database being compromised.

- **Passwords** Passwords are another popular way for access control and identifying individual. It may be considered as a sequence of symbols to remember, or a physical object such as an identification card. Passwords can easily be chosen to be unique, easy to measure, and generally acceptable, while they are easier to steal or copy. One important question is what additional benefits a biometrics-password joint system provides over password systems or biometrics systems alone.

- **Secret embedding** In some systems, there are secrets embedded in biometric signals such that only the genuine user should be able to retrieve them. The biometric signals can be used as an enveloped signal to protect information about the secrets being leaked, as in watermarking systems. On the other hand, the secret signal can also be used to protect information about the biometric traits being leaked. Two important questions regarding secrets in biometric systems are how many different secrets can be embedded in a biometric trait and what benefits such as information leakage protection the secrets can provide to a biometric system.

1.2 Learning the Genuine Hearts: Biometrics Based on Cardiac Signals

In the past 10 years, the idea of establishing biometrics based on signals related to heartbeats has been studied. The signals are obtained using a laser Doppler vibrometor (LDV) remotely measuring skin vibrations at the neck due to arterial movements associated with each pulse, or obtained by attaching electrodes onto the skin measuring electrical changes of the skin due to heart muscle depolarization, known as electrocardiography (ECG). One benefit of such measurements based on heartbeats is that it is difficult to mimic, preventing impostors from breaking into the system.

In earlier studies, it has been shown that both LDV and ECG signals are unique when training data and testing data are obtained consecutively on the same day. However, when the testing data is obtained on another day, from weeks to months after collecting the training data, huge degradation in performance of learning algorithms is observed. A closer examination of the LDV cases suggests that the statistical properties of the impostors are the same, while statistical properties of the genuine cases change from session to session. The performance degradation is thus driven by an increase in the false reject rate.

The underlying mechanism is that for the same genuine individual, the cardiovascular outputs change from day to day, due to physical, psychological, and other factors. The changes are rapid and drastic as compared to face changes due to aging processes. The major challenge is then to overcome the changes in statistical properties of the genuine cases, under constraints on the cost of data collection.

1.3 Outlines and Contributions in the Identification Theme

In chapter 2, a general information theoretic framework of biometric systems is proposed. The framework covers major categories of biometric system designs in the literature, including the use of passwords along with biometric traits, and secrets embedded in biometric signals. An information theoretic outer bound is provided to characterize fundamental trade-offs between system utilities such as identification performance and secret capacity, and system constraints on the amount of

information allowed to be stored and communicated. This is the first work taking into account both database and communication leakage in biometric systems with both secrets and passwords involved. Existing literature concerns mostly only the database leakage, or in cases where communication leakage is also considered, secrets and passwords are not considered.

The contribution of chapter 3 is to provide links between biometric system designs to error correcting code designs and source coding. The theoretical performances of biometric system designs for two scenarios using linear codes are evaluated. One scenario consists of constraints on the number of bits allowed in the database and through the communication channel. The second scenario concerns constraints on information leakage about the biometric template and secrets from the database. In the second scenario, the proposed system design is proved to approach theoretical performance limits in many general statistical models of biometrics.

In chapter 4, robust feature selection concepts and methods are introduced for LDV and ECG biometrics to cope with changes in statistics of a biometric trait of the same individual across sessions. The methods use data from only two training sessions to largely reduce the equal error rates, leading to the best performance in the literature as of 2012 for for both LDV and ECG based biometrics.

1.4 Clustering and Computational Stemmatology

In the second theme, the problem of partitioning unlabeled data points is considered. Clustering is a type of unsupervised learning problem, where there are no true underlying labels available. The goal of the learner is to partition the data into a set of groups, or a hierarchy which "makes sense". It is not a well defined problem unless a measure of "makes sense" of a clustering is defined. However the goodness of such measures still has to match human intuitions, or the true labels on some supervised learning problems. In many clustering approaches, the measure of goodness of a clustering is defined only if the number of resulting clusters is specified. The critical question is how to compare results across different number of clusters, or results of a different data hierarchies.

Stemmatology is the study of reconstructing the relationship and transmission of different text variants of an original document. Throughout the copying process from one variant to the other, differences among parent documents, the ones being copied, and their children documents are introduced by copying error or deliberate changes. The true copying process is not available in most cases as it spans hundreds of years without reliable records. In many cases, the underlying unobservable true copying process is nearly a tree, meaning that if one connects every pair of parent and child variants, the resulting graph will be connected without cycles. A cycle exists when contamination occurs where a copier refers to two or more parent variants to generate the new variant. Thus, the problem of stemmatology is closely related to finding a hierarchical structure among data points, with two additional challenges:

- Some text variants are never recovered.
- In available text variants, there are missing parts in the variants due to physical damage.

The first challenge occurs also in phylogenetics. Phylogenetics is the study of reconstructing the evolution tree among species, where some species are not discovered due to extinction or rarity. Attempts to use methods in phylogenetics yield good performances in smaller datasets of stemmatology problems in which there are fewer missing parts in variants. However in a large dataset with a large number of missing parts in variants, the performances of most methods are degraded.

The minimum description length (MDL) principle is an information theoretic method for selecting models for available data in absence of the "true" model. The idea of the MDL principle is to choose the model which describes the data and the model itself using the least number of bits. In other words, the model should capture the regularity of the data, hence a short data description length, and yet still be simple enough, hence a short description length of the model itself. Albert Einstein had articulated a similar concept:

> "It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience."

In Chapter 5, we apply the MDL principle to clustering and computational stemmatology. The idea is to view the problem of finding a cluster or a hierarchy of data points as finding efficient descriptions of data points. The resulting algorithm performs the best on a large stemmatology dataset as of 2011.

Chapter 2

An Information Theoretic Rate Outer Bound for Identification, Secrets, and Passwords in Biometric Systems

2.1 Introduction

In recent years, advanced results have led to an increasing use of biometric systems for personal identification, authentication and other security applications such as secret binding. There are three research areas emerging from this trend: biometric template security and privacy, secret binding and secret generating using biometrics, and joint systems biometrics with passwords. Template security and privacy are unique challenges resulting from the use of biometric systems that biometric traits are hard to be replaced or canceled once information about biometric traits is stolen or compromised by attacks [39], different from classical information security in communication systems where keys and passwords can be replaced. The key goal of template security and privacy protection in biometric systems is to prevent attackers from obtaining information about the biometrics from databases or communication links between the sensors and databases for unauthorized activities. In secret binding biometric systems, one is also concerned about information leakage of the secret [62].

There are two major categories of biometric system designs addressing different template and secret security approaches versus system performance: biometric cryptosystems and cancelable biometrics [39, 62]. Biometric cryptosystems can be further divided into secret binding and secret generating systems, and cancelable biometrics can be divided into noninvertable systems and salting [39, 62]. In all cases, instead of storing the original biometric templates, helper data is stored in the system. The idea is that the helper data should contain a limited amount of information or be computationally intense to invert so that inferring the original templates are impossible or hard.

In secret binding biometric cryptosystems, the secret is generated independently of the biometrics. Both the secrets and the biometrics are used to generate the helper data. In secret generating cases, the secret is generated directly from the biometrics. For both cases, regeneration of the secret based on the query data and helper data is a major goal of such systems. The regenerated secret can be used as a way of secret based identification where the secret matching is done in an encrypted domain, and a list of encrypted keys has to be stored in the database also. Template protection is realized through not storing full template information but enough for secret regeneration in helper data [40, 39, 62]. In other cases, biometric cryptosystems offer an alternative way of steganography [25]. On the other hand, the ideas of cancelable biometrics specifically focus on template protection so that biometric templates are intentionally stored with distortion in the database such that inversion is not possible. Thus, the matching must be performed on the distorted domain.

Intuitively, to protect template and privacy, the amount of biometric related information stored and transmitted in the system should be small. However on the other hand, biometric system possessing more reliable identification of large number of users and larger secret capability shall require more biometric relation information. It is then a fundamental question to ask, what the tradeoffs between the security of the biometric system itself versus the security and utility provided by the biometric system are.

Recently, fundamental properties of biometric cryptosystems have been studied by Ignatenko and Willems [35] using information theoretic approaches. In their framework, biometric templates are used for secret binding or secret generating. A fundamental information theoretic relation among secret capacity, and information leakage about the biometric templates and secrets from the database are studied, while the use of passwords is also studied [35]. Furthermore in [36], a fundamental trade

off between identification capacity, secret capacity versus information leakage from the database is characterized. Meanwhile, fundamental information theoretic results on the trade off between recognition capacity and the amount of information allowed to be stored in and communicated across a recognition system has been studied extensively [86, 72, 81]. A biometric system is a specialized recognition system so these results can be interpreted as fundamental results on the trade off between biometric identification capacity and the amount of information about a biometric template that must be stored, which is also the template and privacy information that could be leaked if the system was compromised. Others have considered the cases where side information is available to the attacker in biometric cryptosystems [47]. Table 2.1 summarizes existing information theoretic results on biometric systems under different assumptions on locations of attack, utility, and additional information.

On the other hand, encryption is used in biometric cryptosystems, but not all cancellable biometric systems. Thus, by specifying information theoretic secured components of a system, one can then ask what is achievable under information theoretic security only, and what additional capacity and utility can be obtained by allowing additional components that are only cryptographically secure. An important fundamental question is what are the benefits or drawbacks provided by biometric cryptosystems in which identification is based on a key-matching process. In other words, the question one can ask is, under a given set of information theoretic security constraints, how many users can be registered to infer identity directly versus how many keys can be inferred for matching in an encrypt domain. If one can infer more keys than direct inference of user identities, this implies that the use of a hybrid system of information theoretic security and cryptographic security leads to system with larger biometric capacity.

This chapter is organized as follows. In section 2.2 we introduce major biometric system architectures and security issues in the literature that are considered in this paper, and information theoretic security for biometrics. Next, a formal problem definition and a general model is given in section 2.3. The main results is presented in section 2.4 and related to existing literatures. Conceptual interpretation of the results and the implication on practical methods and novel system designs are also discussed. Finally, the chapter is concluded in 2.5, and the proof is given in the Section 2.6.

Table 2.1: Information Theoretic Study on different considerations of biometric systems

	Leakage sources		Utility		Additional Information	
References	Database	Sensor-Database Communication Link	Direct Identity Control	Secret Recovery	Passwords	Side Information
[35]	O			O	O	
[36]	O		O	O		
[86, 81]	O	O	O			
[47]	O			O		O
This work	O	O	O	O	O	

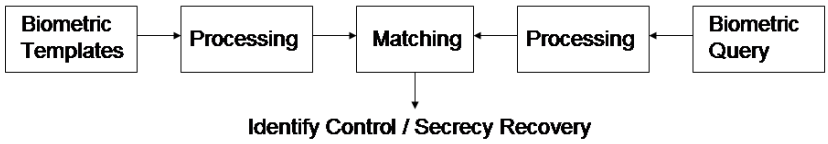

Figure 2.1: Basic biometric systems

2.2 Security and Utility of Biometric Systems

In this section, we introduce major architectures of biometric systems, and information theoretic concepts for measuring security of a biometric system. Consider the basic diagram shown in figure 2.1. For each individual registered in a biometric system, the biometric trait is obtained and processed to store in the database. Here the term process is general, covering processes such as filtering, feature selection, encryption ect.; and the stored data of an individual may be multiple data files while here we first group all together and call it as the data associated with an individual. More detailed models will be discussed later. The sensor is then evoked again to obtain a query data for one of the two identity control modes: authentication or identification. In the authentication mode, the sensor also obtained an input of claimed identity, and the goal is to decide if the sensed data is indeed coming from the claimed individual, e.g. is the data obtained from Alice? In the identification mode, the goal is to decide to whom the data belongs. Upon deciding the identity is genuine, the user or the operator is then allowed to proceed with legitimate activities, such as entering a place, accessing and operating a system, and so on.

Besides serving as an identity control front end, there are also designs of biometric systems that a secret is associated with each user. Such biometric systems are called biometric cryptosystems. The data stored in the database may be a function of both the biometric template and the secret in chosen secret, also known as chosen key, cases. In secret binding, the stored data is a function only of the biometric template. In both settings, the stored data is related to identity control and secret retrieval. When a query data comes in, the goal is either to perform identity control and key retrieval simultaneously, or key retrieval only where identification is based on key matching [62].

There are multiple parts and operational processes in a biometric system that are subjected to attack; for a comprehensive review see Jain, Nandakumar, and Nagar [39]. In particular, leakage

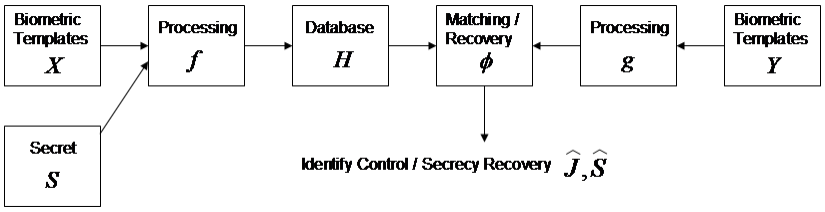

Figure 2.2: Key-Binding System: Information Hiding

of information about the template and secrecy are of primary concern in this work. Template leakage may occur when an attacker successfully enters the database, or an attacker eavesdrops the communication link between different modules, such as between the sensor and the database, [41, 42, 80, 1, 39]. There are two major categories of approaches for template and secret protection, namely biometric cryptosystems and cancelable biometrics.

2.2.1 Template and Secret Protection in Biometric Systems

In biometric cryptosystems, a helper data is stored in the database instead of the original template. There are two subcategories, secret binding systems and secret generating systems. In secret binding system, a key S is chosen for each individual, independently from biometric templates. The helper data H is then a function of the template X and the secret, denoted as $H = f(X, S)$. When a query data Y comes in, the system regenerates the secret S, or the individual identity, or both. Generally we can denote the identity control and secret regeneration process as a function $(\hat{S}, \hat{J}) = \phi(H, Y)$ shown in Figure 2.2. There are two applications of using secret in secret binding systems. One way is to use the secrets as a way for *information hiding* with access control through direct inference of identity using helper data and the query, as shown in 2.2. The other way is *secret based identity control*, as shown in Figure 2.3. In *secret based identity control*, secrets are encrypted and stored as a lookup list in the database, and the identity control is done by matching the lookup list encryption value of recovered secret from helper data and the query. Encryption and thus cryptographic security and computational hardness of inverting keys from the database is involved.

In generating secret systems, each helper data is a function of the corresponding biometric templates of an individual, that the secret and helper data are both functions of the template $(H, S) = f(X)$. As in secret binding system, regeneration of secrets and identity control can be done jointly as in

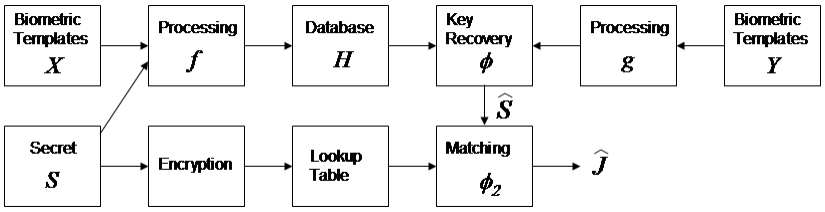

Figure 2.3: Key-Binding System: Key Based Identity Control

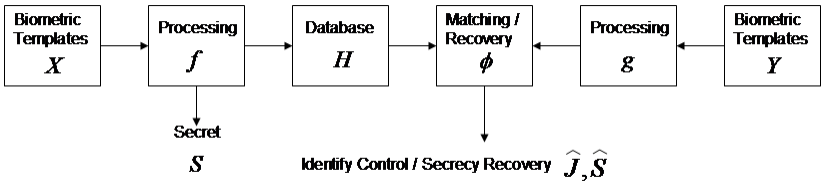

Figure 2.4: Key-Generating System

Figure 2.4 or use the regenerated secret for matching 2.5 with encrypted version of keys stored in the database. One application beyond identity control of secret generating biometric system is in cryptography, where generating identical secret based on the same biometric owner can be use as a way for key distribution.

Cancelable biometric systems can also be divided into two subcategories, noninvertable systems and salting. In noninvertable systems, shown in Figure 2.6, templates are transformed by a noninvertable function, regardless computing power, and the results are stored in the database that identification is done using the query and stored data, usually in the same transformed domain. Such systems are required to prevent template recovery by attackers even if the database and transform parameters are compromised.

Biometric salting, on the other hand, refers to systems using transform functions that are invertible information theoretically, but computationally hard. Template security is approached by choosing different transform parameters K for each template to generate helper data $H = f(X, K)$. Such parameters may be referred to as passwords, secret tokens, or private keys. The parameters have to be presented in the query alone with sensed biometrics that the inference is denoted as $J = \phi(H, Y, K)$.

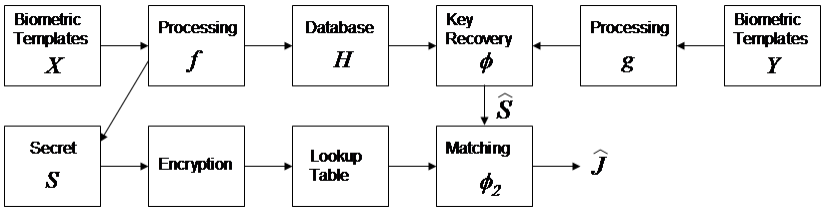

Figure 2.5: Key-Generating System

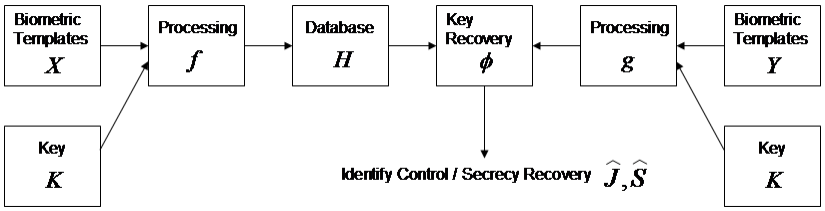

Figure 2.6: Cancelable biometric

2.2.2 Cryptographic Security and Information Theoretic Security

As in secret key based identity control and cancelable biometric, part of security requires a notion of computational hardness, i.e. Cryptographic security, of inverting or computing certain functions. Cryptographic security focuses on infeasibility for the attacker to compute variables of interest. In contrast to cryptographic security, information theoretic security does not assume the computation system or computing limits of the attacker. Information theoretic security depends on the intrinsic randomness existing in the system.

In information theoretic security, mutual information is a common way to measure leakage in biometric systems [36, 47] and communication [53, 22]. For two random variables X and Y, the information leakage about X from Y and vise versa is captured by the mutual information between them as $I(X;Y) = H(X) - H(X|Y)$. The term $H(X|Y)$ itself can be used as leakage too and called information *equivocation* in that context [53].

2.2.3 Biometrics and Information Theoretic Security

In the above mentioned four major categories of biometric systems, the major security concern is information leakage of biometric template information or secret. The leakage may result from

compromised database or communication link. This implies that the mutual information between the stored data, or processed query data, and the templates must be bound by a tolerated leakage level.

However, note that in a key based identification system, storing an encrypted version of each key is necessary, while encryption is not information theoretically secure. In fact, encrypted keys contain full information about the keys as long as the encryption function is reversible with unconstrained computational resource.

In this paper, we focus on information theoretic security, but not cryptographic modules in the systems. Note that above mentioned system architectures with encryption of secrets are not excluded. Secret recovery of such systems is prior to secret matching in encryption domain, in which the accuracy of the systems is bound by recovery accuracy, which can be studied by considering information theoretically secured modules along, assuming encryption function with low collision probability.

Table 2.1 summarizes fundamental results on capacity of biometric systems have been studied in the literature, as well as the contribution of this paper. Six key considerations are listed in the table: leakage of information from database, leakage of information from communication link, identity control, secret recovery, use of password, and side information. In most of the existing literature, up to four elements are considered. For example, in the work of [35], leakage from database, password, secret, and secret recovery are considered. In this paper, all six elements will be considered while deriving fundamental capacity of biometric systems.

2.3 Formal Problem Definition

In this section, we describe a general model which encompass identification, secret key binding, and private key in biometric systems, and the information theoretic framework for studying biometric system security and utility. A biometric system and its operational environment consists of ten parts: the set of subject indices, the distribution of indices, a biometric source, the source alphabet, a secret source, a secret set, a password source, a password set, a observation channel, and the observation alphabet.

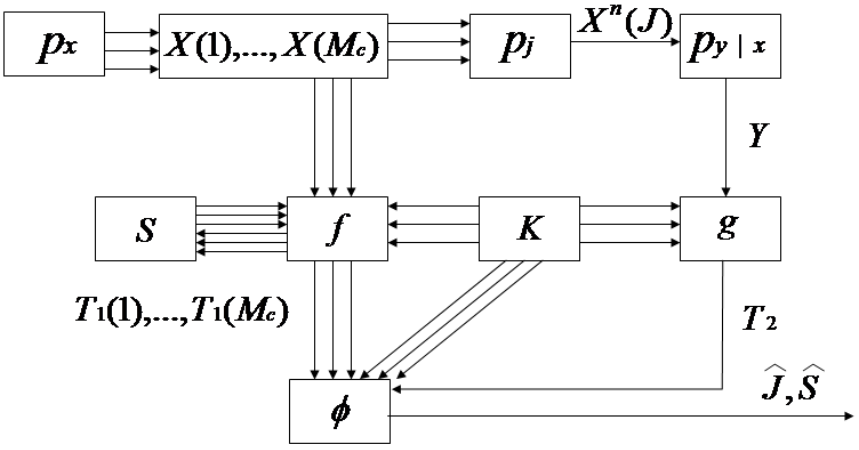

Figure 2.7: The information theoretic privacy protection framework for biometric systems

- The subject index, denoted as j takes value from the set $\mathcal{J} = \{1, \cdots, M_J\}$. The distribution of indices is $p_j, j \in \mathcal{J}$, and in this paper, it is assumed to be uniform.

- The biometric source is modeled as a random process of a collection of distributions p_{x^n} over a source alphabet $x_i \in \mathcal{X}$ for all i. For each individual j to be enrolled, the corresponding biometric template is modeled as a length n realization vector $x^n(j)$ of the biometric source random process. The templates of individuals are assumed to be independent, and are independent of the subject index j, and password $k(j)$.

- The secret source is modeled as an uniform probability mass function $P(\mathbf{S})$ over the set of secrets $\mathcal{S}$. For each individual j, the corresponding secret $s(j)$ is independent of the index j, the template $x^n(j)$, and the private key $k(j)$.

- The password source is modeled as an uniform probability mass function $P(\mathbf{K})$ over the set of private keys $\mathcal{K}$. For each individual j, the corresponding password $k(j)$ is independent of the index j, the template $x^n(j)$, and the secret $s(j)$.

- The query observation y^n is modeled as passing the templates x^n through a noisy channel with a sequence of conditional distributions $p_y^n|x^n$, whose output takes value in an observation alphabet $\mathcal{Y}$, independent of the index j, the secret $S(j)$, and the private key $K(j)$.

2.3.1 System Operation Overview

The biometric system is to identify M_c individuals. During *enrollment*, M_c templates are drawn i.i.d. from the biometric source $p_{\mathbf{x}}$ and presented to the biometric system. For each individual, a secret key $S(j)$ and a private key $K(j)$ are drawn from the probability mass functions $P(\mathbf{S})$ and $P(\mathbf{K})$, independent of everything else. The system encoder takes each template with the secret key and the private key and outputs an encoded data, a.k.a. the helper data, and stores only the helper data but not the templates, secret keys, nor the private keys. During *identification*, an index j is drawn from p_j, and a noisy observation $\mathbf{y}$ of the corresponding template $\mathbf{x}(j)$ is drawn based on the observation channel $p_{\mathbf{y}|\mathbf{x}}$ and presented to the system with the corresponding private key $k(j)$. The biometric system encodes the observed $\mathbf{y}$ along with $k(j)$. The identification algorithm, i.e. the decoder, uses the encoded sensor data, along with all the helper data and all private keys for identification and outputs an estimated identity $\hat{j}$ and an estimated secret $S(\hat{j})$. Note that the private keys K are assumed to be secure and available to both the enrollment encoder and the sensory encoder, as well as the identification algorithm.

2.3.2 Encoding and Decoding

For enrollment, the helper data of an individual j is generated based on template $\mathbf{x}(j)$, secret key $s(j)$, and private key $k(j)$. The helper data is denoted as $t_1(j)$ taking values from a set $\mathcal{T}_1$. $t(j)$ is the output of a *template encoder* f:

$$f(\mathbf{x}(j), s(j), k(j); p_{\mathbf{x},\mathbf{y}}, p_j, P(S), P(K)) : \mathcal{X}^n \times \mathcal{S} \times \mathcal{K} \to \mathcal{T}_1. \tag{2.1}$$

The notation means that the encoder may use all the distributions to encode a template, but not the realizations of other templates, secrets, private keys, nor the observation $\mathbf{y}$. This assumption is made in most information theoretic studies of biometric systems [86, 81, 35, 47]. The set of all encoded data is denoted as $\mathbf{T}_1 = \{t(1), \cdots, t(M_c)\}$.

Similarly, for observation encoding, the encoded data is denoted as t_2 taking values from a set $\mathcal{T}_2$. It is the output of an *observation encoder* g:

$$g(\mathbf{y}, k(j); p_{\mathbf{x},\mathbf{y}}, p_j, P(S), P(K)) : \\ \mathcal{Y}^n \times \mathcal{K} \to \mathcal{T}_2. \tag{2.2}$$

The observation encoder uses the observed $\mathbf{y}$ and $k(j)$ and all distributions, but not the realizations of $\mathbf{x}$.

The identification algorithm is a function ϕ:

$$\phi(t_2, \mathbf{T}_1, \mathbf{K}; p_{\mathbf{x},\mathbf{y}}, p_j, P(S), P(K)) : \\ \mathcal{T}_2 \times \mathcal{T}_1^{M_c} \times \mathcal{K}^{M_c} \to (\mathcal{J}, \mathcal{S}). \tag{2.3}$$

The identification algorithm uses the realizations of all helper data, the password observed at the sensor, and the distributions, but not the realizations of the templates nor the observation. It outputs the estimate of the identity $\hat{j}$ and the associates secret $s(\hat{j})$.

2.3.3 Definition of Achievability

A rate tuple $(R_M, L_1, L_2, R_S, R_K)$ is said to be achievable if there exists a sequence of encoder and decoder triples (f_n, g_n, ϕ_n) such that for any $\delta > 0$ and n large enough:

$$\begin{aligned} Pr\left((\hat{j}, S(\hat{j})) \neq (j, S(j))\right) &\leq \delta \\ n^{-1} I(S(j); T_1(j)) &\leq \delta \quad \forall j \\ n^{-1} I(S(j); T_2(j)) &\leq \delta \quad \forall j \\ n^{-1} I(X(j); T_1(j)) &\leq L_1 \quad \forall j \\ n^{-1} I(Y(j); T_2(j)) &\leq L_2 \quad \forall j \\ n^{-1} \log |\mathcal{J}| &\geq R_M \\ n^{-1} \log |\mathcal{S}| &\geq R_S \\ n^{-1} \log |\mathcal{K}| &\leq R_K. \end{aligned}$$

The set of all achievable rate tuples is defined as the achievable rate region, denoted as $\mathcal{R}$.

2.4 Main Results and Discussion

The main result is summarized in the following converse theorem.

$$\begin{aligned} \mathcal{R} \subseteq \mathcal{R}_{out} = \big\{ R_S + R_M - R_K &\leq I(\mathscr{X};\mathscr{U}) + I(\mathscr{Y};\mathscr{V}) - I(\mathscr{X},\mathscr{Y},;\mathscr{U},\mathscr{V}) \\ L_1 + R_K - R_M &\geq -I(\mathscr{Y};\mathscr{V}) + I(\mathscr{X},\mathscr{Y};\mathscr{U},\mathscr{V}) \\ L_2 + R_K - R_M &\geq -I(\mathscr{X};\mathscr{U}) + I(\mathscr{X},\mathscr{Y};\mathscr{U},\mathscr{V}) \big\} \end{aligned}$$

where $\mathscr{U}$ and $\mathscr{V}$ are auxiliary random variables:

$$\begin{aligned} U^n &= (X^{n-1}, T_1, S, K) \\ V^n &= (Y^{n-1}, T_2, K), \end{aligned}$$

satisfying the Markov conditions $U^n - X^n - Y^n$ and $X^n - Y^n - V^n$,

$$I(\mathscr{X};\mathscr{Y}) = H(\mathscr{X}) + H(\mathscr{Y}) - H(\mathscr{X},\mathscr{Y}), \tag{2.4}$$

and $H(\mathscr{X})$ denotes the entropy rate of the random process X.

2.4.1 Special Cases

The outer bound presented in this paper is tight in the sense that the outer bounds or exact achievable rate region in many special cases published in the literature can be directly derived from it.

Zero Privacy leakage and passwords

The theorem states that

$$L_1 + R_K - R_M \geq -I(\mathscr{Y}; \mathscr{V}) + I(\mathscr{X}, \mathscr{Y}; \mathscr{U}, \mathscr{V}), \tag{2.5}$$

where the right hand side is nonnegative. Thus it is impossible to avoid potential privacy leakage in direct identification biometric systems using a biometric trait alone. Even in secret based identification systems, one has to assume that the lookup list of encrypted secrets will not be compromised, which is theoretically and practically difficult to validate. However, with the use of a password along with a biometric, one can develop zero privacy leakage and zero secrecy leakage systems.

Recognition with constraints

There are a number of results on information theoretic biometrics or recognition, where the focus is solely on direct identification, or recognition, without secrets or passwords. The constraints are focused on the amount of information allowed to be leaked, or allowed to be stored. Note that these two constraints are closely related but different. The former is measured by mutual information, the latter is measured by bits per stored helper data. Early studies yield fundamental insights on systems whose database is constrained, while the sensory capacity and communication link are not [71]. More comprehensive results on recognition with both database and sensing, or communication, constrains were obtained by Westover and O'Sullivan [86]. The constraints are

$$\log \mathcal{T}_1 \leq R_1 \tag{2.6}$$

$$\log \mathcal{T}_2 \leq R_2. \tag{2.7}$$

The achievable rate region $\mathcal{R}^{wot}$ is defined as the set of all achievable rate tuples (R_1, R_2, R_M) such that for any $\epsilon > 0$ and n sufficiently large, there exist (f, g, ϕ) such that

$$\begin{aligned} \Pr(\hat{w} \neq w) &\leq \epsilon \\ n^{-1} \log |\mathcal{T}_1| &\leq R_1 \end{aligned}$$

$$
\begin{aligned}
n^{-1}\log|\mathcal{T}_2| &\leq R_2 \\
n^{-1}\log|M_c| &\geq R_c.
\end{aligned}
$$

The outer bound $\mathcal{R}_{out}^{rec}$ [86, 81], rewritten with some algebra, is

$$
\begin{aligned}
\mathcal{R}_{out}^{rec} = \big\{ R_M &\leq I(\mathscr{X};\mathscr{U}) + I(\mathscr{Y};\mathscr{V}) - I(\mathscr{X},\mathscr{Y},;\mathscr{U},\mathscr{V}) \\
R_1 - R_M &\geq -I(\mathscr{Y};\mathscr{V}) + I(\mathscr{X},\mathscr{Y};\mathscr{U},\mathscr{V}) \\
R_2 - R_M &\geq -I(\mathscr{X};\mathscr{U}) + I(\mathscr{X},\mathscr{Y};\mathscr{U},\mathscr{V}) \big\}.
\end{aligned}
$$

Notably, any achievable rate 3-tuple (R_1, R_2, R_c) for these conditions, also satisfies the achievable conditions described in section 2.3 since

$$
\begin{aligned}
R_1 &> \frac{1}{n}\log|\mathcal{T}_1(j)| && \forall j \\
&\geq \frac{1}{n}H(T_1(j)) && \forall j \\
&= \frac{1}{n}\left(H(T_1(j)) - H(T_1(j)|\mathbf{X})\right) && \forall j \\
&= \frac{1}{n}I(\mathbf{X};T_1(j)). && \forall j
\end{aligned}
$$

Thus, $\mathcal{R}_{out}^{rec}$ is a subset of $\mathcal{R}_{out}$. In fact the proposed outer bound $\mathcal{R}_{out}$ is exactly $\mathcal{R}_{out}^{rec}$ when $R_S = R_K = 0$. Thus the outer bound is tight in this case, and constraints on storage capacity in terms of bits and constraints on leakage in terms of mutual information are closely related.

Special Cases in the Literature

A comprehensive study of biometric privacy and secrecy was conducted by Ignatenko and Williams [35]. Even though direct identification and sensor communication leakage are not considered, fundamental information theoretic results on secret capacities with and without passwords in biometric systems were thoroughly studied, including both chosen secret cases and generated secret cases. The outer bound derived in this paper matches the special cases presented in their work as four theorems: theorem 1 on generated secret cases, theorem 3 on chosen secret cases, theorem 5 on zero leakage generated secret with password cases, and theorem 7 on zero leakage chosen secret with password

cases. Each can be derived from our results by removing the modules or constraints that are not considered. These results were also obtained by Lai, Ho, and Poor [47], while in [47] attackers with side information are also considered. Note that in [35], the achievable rate regions are exact that both achievability and converse yields the same bound. This also indicates that our outer bound is tight in the sense that in many special cases it matches the achievable regions.

Identification and secrets in biometrics were considered simultaneously later also by Williams and Ignatenko [89], while passwords and sensor communication leakage were not included. The exact achievable rate region found in Ignatenko [89] is a special case of the outer bound in this paper, thus our bound is tight in this case.

2.4.2 Direct Identity Control versus Key Based Identity Control

In key based identity control, as we mentioned, the system has a risk of leaking key information if the list of encrypted keys are compromised. The question is then the following: is there any benefit offered by key based identity control that may justify this risk? In other words, what is the benefit of combining information theoretic secured modules in conjunction with computationally secured modules in biometric systems?

To see this, consider direct identity control systems with identification rate R_M^d without secret hiding. The corresponding key based identity control system will need to have $R_S^k = R_M^d$ to serve the same number of users, as each user will be identified by each individual encrypted key. Note that the direct identification rate $R_M^k = 0$ in key based identification system. Thus we have $\mathcal{R}_{out}^d$, converse of achievable rate region of direct identity control systems, as

$$\begin{aligned}
\mathcal{R}_{out}^d = & \\
\big\{R_M^d - R_K \leq & \quad I(\mathscr{X};\mathscr{U}) + I(\mathscr{Y};\mathscr{V}) \\
& - \quad I(\mathscr{X},\mathscr{Y},;\mathscr{U},\mathscr{V}) \\
L_1 + R_K - R_M^d \geq & - \quad I(\mathscr{Y};\mathscr{V}) + I(\mathscr{X},\mathscr{Y};\mathscr{U},\mathscr{V}) \\
L_2 + R_K - R_M^d \geq & - \quad I(\mathscr{X};\mathscr{U}) + I(\mathscr{X},\mathscr{Y};\mathscr{U},\mathscr{V})\big\},
\end{aligned}$$

while $\mathcal{R}_{out}^{sk}$, the converse of the achievable rate region for secret key based identity control systems, is

$$\begin{aligned}
&\mathcal{R}_{out}^{sk} = \\
\big\{R_S^{sk} - R_K \leq \quad & I(\mathscr{X};\mathscr{U}) + I(\mathscr{Y};\mathscr{V}) \\
- \quad & I(\mathscr{X},\mathscr{Y},;\mathscr{U},\mathscr{V}) \\
L_1 + R_K \geq \quad - \quad & I(\mathscr{Y};\mathscr{V}) + I(\mathscr{X},\mathscr{Y};\mathscr{U},\mathscr{V}) \\
L_2 + R_K \geq \quad - \quad & I(\mathscr{X};\mathscr{U}) + I(\mathscr{X},\mathscr{Y};\mathscr{U},\mathscr{V})\big\}.
\end{aligned}$$

Thus when both systems serve the same number of users, $R_S^{sk} = R_M^d$, the secret key base system requires less leakage of template information. This means *if* one has confidence in, or is willing to risk, security of the encryption list, there is a gain in protection of information leakage of the biometric templates.

2.4.3 Insights for System Designs

A classical idea of biometric template protection system called fuzzy commitment scheme by Jeols and Wattenberg. The framework considers identification through secret recovery, while no leakage constraint was considered. The idea is to store the helper data $f(x) = x - w$ and encryption $f_2(w)$, where w is a codeword of some error correcting code. Note the w then can be think of as the secret in our framework. When a query data comes in, the difference $w' = y - x - w$ is computed. If x and y are from the same individual and hence close, w' is then close to w. If w and w' are close enough to be within the error correcting capacity of the code, w and $f_2(w)$ can then be recovered and used as identification. Note that since $x - w$ is stored in the data base, recover w is equivalent to recovering x. Later similar ideas are proposed by O'Sullivan and Lai [60] and Martinian et al. [55], known as template protection by syndrome or distributed source coding, that $f(x)$ and $f_2(x)$ are stored, and then the query y is used with helper data $f(x)$ to recover x and hence $f_2(x)$. Here the function f is usually the low density generator matrix of a low density parity check code.

Note that, however, recovering x requires $\min(L_1, L_2) \geq R_M + H(X|Y)$ [60, 48, 55]. However, from the discussion in 2.4.1, we know that recognition is possible even when L_1 and L_2 are smaller than

$H(X|Y)$. Thus a new design which beyond identification through direct or indirect reconstruction of the template is need. In fact, simple codes are proposed and proven to work very close to the presented theoretical outer bound for simple theoretical examples [48].

2.5 Conclusion and Remarks

In this chapter, we develop a general information theoretic framework for biometric secrecy systems. This framework captures the fundamental property of major biometric system designs, including key-binding, key-generating, biometric cryptosystems, and cancelable biometrics. In addition to considering information leakage due to the database being compromised, we also consider cases where the communication link between the database and sensor is compromised. An outer bound on system capacity is presented; the bound is tight that it coincides with results in the literatures in many special cases.

The results also lead to a theoretical comparison of direct identification systems versus key based identification systems. When one builds a key based identification system, there is an additional risk of encrypted key database being compromised. However, assuming the encryption is intact, one gains more protection of the biometric privacy, in the event of attacks on helper data in the database or communication link between sensor and database.

The obtained bound also provides insights on system designs. An interesting consequence of the presented outer bound and its tightness is that identification is possible while reconstruction of the original signal is not possible. This is different from the idea of identification through reconstruction such as the use of Slepain-Wolf coding. However it does not rule out the possibility of modifying and extending the way which Slepain-Wolf coding could be used, such as partial reconstruction or rate distortion.

An important category of security concern whose fundamental result is not obtained in this work is conditional leakage. In conditional leakage cases, one assumes that the secret may be stolen and may be used along with the helper data to gain more information about the biometrics. Conditional

leakage in some biometric systems is studied by Ignatenko and Willems [35], while only the database is assumed to be vulnerable, and only secret reconstruction is considered.

2.6 Proof of the Converse Outer Bound

2.6.1 Identification and Secrecy Rates

$$
\begin{aligned}
& n(R_S + R_M) \\
\leq\ & \log|\mathcal{S}(\mathcal{J})| + \log|\mathcal{J}| \\
=\ & H(S(J), J) \\
=\ & I(S(J), J; \mathbf{T}_1, T_2(J), K(J)) + H(S(J), J|\mathbf{T}_1, T_2(J), K(J)) \\
\leq\ & I(S(J), J; \mathcal{T}_1, T_2(J), K(J)) + \mathcal{F}, \\
=\ & I(S(J), J; \mathbf{T}_1) + I(S(J), J; K(J)|\mathbf{T}_1) + I(S(J), J; T_2(J)|\mathbf{T}_1, K(J)) + F \\
\leq\ & I(S(J), J; \mathbf{T}_1) + nR_K + I(S(J), J; T_2|\mathbf{T}_1, K(J)) + F \\
=\ & I(J; \mathbf{T}_1) + I(S(J); \mathbf{T}_1|J) + I(S(J), J; T_2(J)|\mathbf{T}_1, K) + nR_K + F \\
\leq\ & 0 + n\delta + I(S(J), J; T_2(J)|\mathbf{T}_1, K) + nR_K + F \\
=\ & I(S(J), J; K, T_2(J)|\mathbf{T}_1, K) + nR_K + F + n\delta \\
=\ & I(S(J), J, \mathbf{T}_1; T_2(J)|K(J)) - I(\mathbf{T}_1; T_2(J)|K(J)) + nR_K + F + n\delta \\
\leq\ & I(S(J), J, \mathbf{T}_1; T_2(J)|K(J)) + nR_K + F + n\delta \\
\overset{(a)}{=}\ & I(S(J), T_1(J); T_2(J)|K) + nR_K + F + n\delta \\
\overset{(b)}{=}\ & I(X^n(J); T_1, S(J)|K(J)) + I(Y^n(J); T_2|K(J)) \\
& -I(X^n(J), Y^n(J); T_1(J), S(J), T_2(J)|K(J)) + nR_K + F + n\delta \\
\overset{(c)}{=}\ & I(X^n(J); T_1(J), S(J), K(J)) + I(Y^n(J); T_2(J), K(J)) \\
& -I(X^n(J), Y^n(J); T_1(J), S(J), T_2(J), K(J)) + nR_K + F + n\delta \\
\overset{(d)}{\leq}\ & I(X^n(J); U^n) + I(Y^n(J), V^n) - I(X^n(J), Y^n(J), ; U^n, V^n) + nR_K + F + n\delta,
\end{aligned}
$$

where

$$U^n = (X^{n-1}, T_1, S, K) \tag{2.8}$$
$$V(n = (Y^{n-1}, T_2, K).$$

Divide both sides by n, move nR_K to the left hand side, and take the limit as n gets large, we have

$$R_s + R_m - R_k \leq I(\mathscr{X}; \mathscr{U}) + I(\mathscr{Y}; \mathscr{V}) - I(\mathscr{X}, \mathscr{Y}, ; \mathscr{U}, \mathscr{V}), \tag{2.9}$$

where

$$I(\mathscr{X}; \mathscr{U}) \triangleq H(\mathscr{X}) + H(\mathscr{U}) - H(\mathscr{X}, \mathscr{U}), \tag{2.10}$$

and $H(\mathscr{X})$ is the entropy rate of the random process X^n. In particular, for X and Y being i.i.d., we have the last being

$$\begin{aligned} & I(X^n; T_1, S, K) + I(Y^n, T_2, K) - I(X^n, Y^n; T_1, S, K, T_2) + nR_K + F + \delta \\ \stackrel{(e)}{=} \; & nI(X; U) + nI(Y; V) - nI(X, Y; U, V) + nR_K + F + \delta. \end{aligned}$$

By dividing n on both sides and taking limit as n gets large, we have

$$R_s + R_m - R_k \leq I(X; U) + I(Y; V) - I(X, Y; U, V). \tag{2.11}$$

It is clear from definitions that $U^n - X^n - Y^n$ and $X^n - Y^n - V^n$. This completes the proof.

For (a), define $\mathbf{T}_1^* = \mathbf{T}_1 \setminus T_1(J)$, we have

$$\begin{aligned} & I(S(J), J, \mathbf{T}_1; T_2(J)|K(J)) \\ = \; & I(S(J), J, \mathbf{T}_1^*, T_1(J); T_2|K(J)) \\ = \; & I(S(J), T_1(J); T_2(J)|K(J)) + I(J, \mathbf{T}_1^*; T_2(J)|S(J), K(J), T_1(J)) \\ = \; & I(S(J), T_1(J); T_2(J)|K(J)) + I(J; T_2(J)|S(J), K(J), T_1(J)) + I(\mathbf{T}_1^*; T_2(J)|S(J), K(J), T_1(J)) \\ = \; & I(S(J), T_1(J); T_2(J)|K(J)) + 0 + 0. \end{aligned}$$

Equality (b) utilizes lemma 6 of Westover and O'Sullivan [86], which states that

$$I(\alpha;\beta|C) \geq I(A;\alpha|C) + I(B;\beta|C) - I(A,B;\alpha,\beta|C), \tag{2.12}$$

with equality if and only if $I(A,\alpha;B,\beta|C) = I(A,B|C)$. Equality (b) follows from substituting (A,B,C,α,β) with $(X^n,Y^n,K,(S,T_1),(T_2))$ and then showing that equality condition holds. We have

$$\begin{aligned}
& I(X^n,S,T_1;Y^n,T_2|K) \\
= \;& I(X^n,S,T_1;Y^n,K) + I(X^n,S,T_1;T_2|Y^n,K) \\
\overset{(ba)}{=} \;& I(X^n,k,S,T_1;Y^n|K) + 0 \\
= \;& I(X^n;Y^n|K) + I(S;Y^n|K,X^n) \\
\overset{(bb)}{=} \;& I(X^n;Y^n),
\end{aligned}$$

where (ba) results from T_2 is a function of (Y^n,K), and (bb) follows that K and S are independent of X^n and Y^n, and K and S are independent of each other.

The inequality (d) is based on the fellowing Lemma 1. *Lemma 1* Let A^n be a random processes such that its entropy rate exists and B be a random variable and $C_i = (A^{i-1},B)$, we have

$$I(A^n;B) \leq I(A^n;C^n) \tag{2.13}$$

By substitute (A,B,C) three times to $(X,(T_1,S,K,),U)$, $(Y,(T_2,K),V)$, and $((X,Y),(T_1,T_2,S,K),(U,V))$, we get the inequality (c).

proof

$$\begin{aligned}
& I(A^n;C^n) \\
= \;& I(A^n;A^{n-1},B,\cdots,B) \\
\geq \;& I(A^n;B)
\end{aligned}$$

For cases where X and Y are i.i.d., (c) can be derived from writing the mutual information terms in the form of $I(X^n; T_1, S, K)$ into telescoping sum

$$\begin{aligned} I(X^n; T_1, S, K) &= \sum_{i=1}^{n} I(X_i; T_1, S, K | X^{i-1}) \\ &\leq \sum_{i=1}^{n} I(X_i; T_1, S, K, X^{i-1}) \\ &= \sum_{i=1}^{n} I(X_i; U_i). \end{aligned}$$

Then using standard argument through defining a time sharing parameter T and variable $U = (U_T, T)$, we get the last inequality.

2.6.2 Privacy and Secrecy leakage

$$\begin{aligned} & nL_1 \\ \geq\ & \frac{1}{M_c} \sum_{j=1}^{M_c} I(X^n(j); T_1(j)) \\ =\ & I(X^n(J); T_1(J) | J) \\ \overset{(a)}{=}\ & I(J, X^n(J); T_1(J)) \\ \geq\ & H(X^n(J), J) - H(X^n(J), J | T_1(J)) - H(S(J) | X^n(J), J, T_1(J), K(J)) \\ =\ & H(X^n(J), J) - I(X^n(J), J; K(J) | T_1(J)) - H(X^n(J), J | T_1(J), K(J)) \\ & - H(S(J) | X^n(J), J, T_1(J), K(J)) \\ \geq\ & H(X^n(J), J) - H(K) - H(X^n(J), J, S(J) | T_1(J), K(J)) \\ =\ & H(X^n(J), J) - H(K) - H(J, S(J) | T_1(J), K(J)) - H(X^n(J) | J, S(J), T_1(J), K(J)) \\ =\ & H(X^n(J), J) - H(K) - H(J, S(J) | T_1(J), K(J), T_2) - I(T_2; J, S(J) | T_1(J), K(J)) \\ & - H(X^n(J) | J, S(J), T_1(J), K(J)) \\ \overset{(b)}{\geq}\ & H(J) - H(K) + H(X^n) - H(X^n(J) | J, S(J), T_1(J), K(J)) - I(T_2; J, S(J) | T_1(J), K(J)) - \mathcal{F} \\ \geq\ & H(J) - H(K) + I(X^n(J); J, S(J), T_1(J), K(J)) - I(T_2; J, S(J), T_1(J) | K(J)) - \mathcal{F} \\ =\ & nR_M - nR_K + I(X^n(J); J, S(J), T_1(J), K(J)) - I(T_2; J, S(J), T_1(J) | K(J)) - \mathcal{F} \end{aligned}$$

$$
\begin{aligned}
&\overset{(c)}{\geq} && nR_M - nR_K + I(X^n(J); S(J), T_1(J)|K(J)) - I(T_2; J, S(J), T_1(J)|K(J)) - \mathcal{F} \\
&\overset{(d)}{\geq} && nR_M - nR_K - I(Y^n(J); T_2, K(J)) + I(X^n(J), Y^n(J); T_1, S(J), T_2, K(J)) - \mathcal{F} \\
&\overset{(e)}{=} && nR_M - nR_K - I(Y^n; V^n) + I(X^n, Y^n; U^n, V^n) - \mathcal{F}
\end{aligned}
$$

where (a) comes from J being independent of $T_1(J)$; (b) follows from Fano's inequality of the term $H(J, S(J)|T_1(J), K(J)) \leq H(\hat{J}, \hat{S(J)})$, and $X^N(J)$ and J are independent; (c) follows from $X^n(J)$ being independent of J and $K(J)$, and also J and $K(J)$ being independent of each other; (d) and (e) both follow the same arguments as in the identification and secrecy rate proof. Dividing both sides by n, we have

$$L_1 + R_K \geq R_M - I(\mathscr{Y}; \mathscr{V}) + I(\mathscr{X}, \mathscr{Y}; \mathscr{U}, \mathscr{V}). \tag{2.14}$$

Similar, for L_2, we have

$$
\begin{aligned}
& && nL_2 \\
&\geq && \frac{1}{M_c} \sum_{j=1}^{M_c} I(Y^n(j); T_2(j)) \\
&= && I(Y^n(J); T_2(J)|J) \\
&\overset{(a)}{=} && I(J, Y^n(J); T_2(J)) \\
&\geq && H(Y^n(J), J) - H(Y^n(J), J|T_2(J)) - H(S(J)|Y^n(J), J, T_2(J), K(J)) \\
&\geq && H(Y^n(J), J) - H(K) - H(Y^n(J), J, S(J)|T_2(J), K(J)) \\
&= && H(Y^n(J), J) - H(K) - H(J, S(J)|T_2(J), K(J)) - H(Y^n(J)|J, S(J), T_2(J), K(J)) \\
&= && H(Y^n(J), J) - H(K) - H(J, S(J)|T_2(J), K(J), T_1) - I(T_1; J, S(J)|T_2(J), K(J)) \\
& && -H(Y^n(J)|J, S(J), T_2(J), K(J)) \\
&\overset{(b)}{\geq} && H(J) - H(K) + H(Y^n) - H(Y^n(J)|J, S(J), T_2(J), K(J)) - I(T_1; J, S(J)|T_2(J), K(J)) - \mathcal{F} \\
&\geq && H(J) - H(K) + I(Y^n(J); T_2(J), K(J)) - I(T_2; J, S(J), T_1(J)|K(J)) - \mathcal{F} \\
&\overset{(c)}{=} && H(J) - H(K) + I(Y^n(J); T_2(J), K(J)) - I(T_2; T_1(J)|K(J), J, S(J)) - \mathcal{F} \\
&\leq && nR_M - nR_K + I(Y^n(J); T_2(J)|K(J)) - I(T_2; J, S(J), T_1(J)|K(J)) - \mathcal{F} \\
&= && nR_M - nR_K - I(X^n; U^n) + I(X^n, Y^n; U^n, V^n) - \mathcal{F}.
\end{aligned}
$$

Dividing both sides by n, and let n go large, we have

$$L_2 + R_K - R_M \leq -I(X^n; U^n) + I(X^n, Y^n; U^n, V^n). \tag{2.15}$$

Inequality (c) is from the following lemma: $I(A; f(B,C)|B) = 0$ if $B \perp A, B \perp C$, and $C \perp (A,B)$, or equivalently, $H(A|B) = H(A|B, f(B,C))$ We have

$$\begin{aligned} H(A|B, f(B,C)) &\geq H(A|B,C) \\ &= H(A,B,C) - H(B,C) \\ &= H(A,B) + H(C) - H(B) - H(C) \\ &= H(A|B). \end{aligned}$$

On the other hand, $H(A|B, f(B,C)) \leq H(A|B)$. Thus $H(A|B) - H(A|B, f(B,C)) = I(A; f(B,C)|B) = 0$ given the conditions holds. Notice that

$$I(T_2; J, S(J), T_1(J)|K(J)) = I(T_2; J, S(J)|K(J)) - I(T_2; T_1(J)|K(J), J, S(J)), \tag{2.16}$$

and substitute $(A, B, C, f(A,B)$ with $((J, S(J)), K(J), X^n(J), T_1(J))$, we have equality (c).

Chapter 3

Secured Biometric System Designs Using Linear Codes

In the previous chapter, an information theoretic framework is presented along with an outer rate bound which characterizes the fundamental trade off between biometric system utility and security. In this chapter, we look at the system design and coding theoretic aspects of the problem. In coding theory for communication problems, one focuses on the actual design and algorithm of codes that can be implemented, and attempts to approach the limits established from information theoretic results. Here we bring a similar concept of coding for practical system designs into biometric systems.

Instead of developing new kinds of error correcting codes or source codes, we focus on the concept of translating biometric system design problems into coding problems. The results in this chapter have the following flavor: if one has an error correcting code, or a source code, for a noise or source random process, then one also has a good biometric system with the identical random process as noise model or biometric source model. Also, if one has an decoding algorithm for the code, it can be used as the corresponding identification or verification problem. Thus, this chapter introduces several ways to turn biometric system design problems into code design problems that the vast amount of coding theory literature can then be used as powerful tools in biometrics.

In this chapter, we translate two biometric problems into coding problems by proposing the actually system design using linear codes, and then establishing the theoretical performances of the resulting biometric systems. In both problems, we focus on discrete cases. The first problem is to design an

identification system with constraints on the number of bits of the helper data per individual, and the number of bits communicating from sensor to database. Despite the theoretical performance of the proposed system design is suboptimal in this case, it still shows the potential of linear codes in identification problems. Also in a simple theoretical case, we illustrate that near optimal identification performance can be achieved while signal reconstruction is not possible, and it is robust against uncertainty in the noise model.

In the second case, we study the problem of secret binding in biometric system. In this problem, the information leakage of template and secret from the database is considered, but not from the communication channel. We propose a simple system design that can achieve optimal identification and secret capacity trade off shown in the last chapter and in other literature. In addition, we also propose an alternative system which is suboptimal but reduces the potential computational complexity of the problem.

3.1 Identification System with Finite Storage and Communication Constraints

Three aspects defining the first identification problem are the *environment* under which identification takes place, the *identification system* itself, and *measures of performance*. These follow the problem setting in [88], which is a slight modification of the general model described in the previous chapter.

Environment

The environment consists of six elements, denoted as

$$\mathcal{E} = (P_x, \mathcal{X}, P_{y|x}, \mathcal{Y}, M_c, P_j). \tag{3.1}$$

$M_c = 2^{nR_c}$ is the total number of individuals to be registered in the system, and R_c is defined as the identification rate. Each pattern template is a length n sequence with each element taking values over a discrete set $\mathcal{X}$. Each pattern is drawn independently from a distribution P_x, denoted as

$\mathbf{x}(j), i \in \{1, 2, \cdots, M_c\}$. The set of all M_c patterns to be recognized is denoted as $\mathcal{C}$. In the training phase, we assume that the biometric system observes all template patterns $x(j)$.

In the testing phase, an individual index j is drawn from $\{1, 2, \cdots, M_c\}$ based on an index distribution P_j which we assume to be uniform here. The corresponding object sequence $\mathbf{x}_j$ is then presented to the identification system through a noisy channel whose transition probability is $P_{\mathbf{y}|\mathbf{x}}$, where each element of $\mathbf{y}$ takes values over the set $\mathcal{Y}$.

In this section, the noisy channel is assumed to be additive that the input and our are both integers $\mathcal{X} = \mathcal{Y}$ which are finite subset of $\mathbb{Z}$. The noise is denoted as $\mathbf{z}$ which is a length n sequence drawn from a distribution $P_{\mathbf{z}}$, independent of $\mathbf{x}(j), \forall j$, and any part of identification systems. Hence

$$P_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) = P_{\mathbf{z}}(\mathbf{y} - \mathbf{x}), \tag{3.2}$$

and the identification system observes data

$$\mathbf{y} = \mathbf{x} + \mathbf{z}. \tag{3.3}$$

The Biometric Identification System

An identification system consists three parts: a database compression function f, a sensor compression function g, and an identification algorithm ϕ.

The database compression f maps each object sequence $\mathbf{x}_j$ to a compressed sensor data $t_1(j) \in \mathcal{T}_1$, where L_1 is the database compression rate. Notice that the definition of L_1 is different from the leakage definition presented in the previous chapter measured by mutual information between $\mathbf{X}$ and T_1.

Similarly, the sensor compression function g maps an observed $\mathbf{y}$ to a compressed sensor data $t_2 \in \mathcal{T}_2$, where $L_2 = \log |\mathcal{T}_2|$ is defined to be the sensor compression rate.

In cases of system design using linear codes, the sensor compression and database compression are carried out by using two matrices H of size nT_1 by n for the database and G of size nT_2 by n for

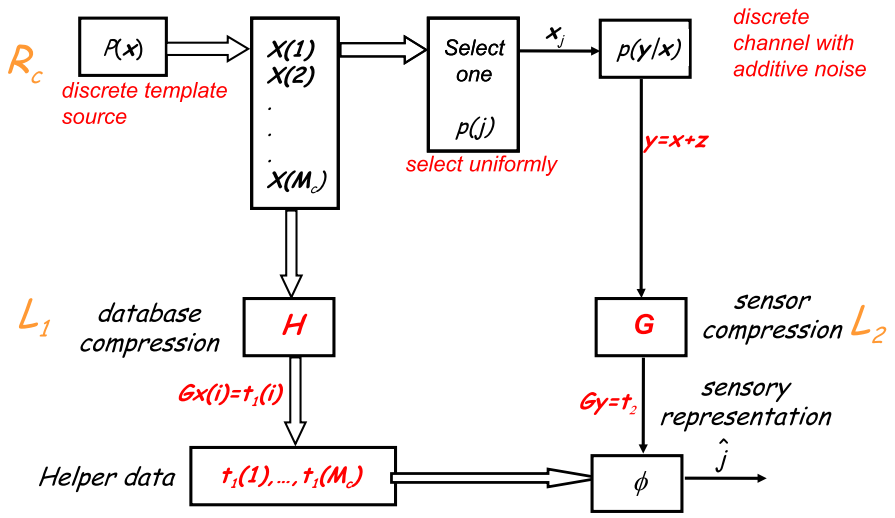

Figure 3.1: Identification system design using linear codes

the sensor, such that

$$t_1(j) = H\mathbf{x}(j) \tag{3.4}$$

is the compressed database data of the template with index j, and

$$t_2 = G\mathbf{y} \tag{3.5}$$

is the compressed sensor data. The set of all database data $t_1(j), j \in \{1, 2, \cdots, M_c\}$ is denoted as $\mathbf{T}_1$. In designing database compression H and sensor compression G, we assume that the probability mass functions $P_{\mathbf{x}}$ and $P_{\mathbf{z}}$ are given.

We are interested in designing good identification systems given $(R_c, L_1, L_2, P_x, P_z)$. The identification algorithm ϕ takes $\mathbf{T}_1$ and t_2 as inputs and computes an estimate $\hat{j}$ of the true object index, assuming the probabilities $p_{\mathbf{x}}$ and $p_{\mathbf{z}}$ are given. It consists of a noise estimation algorithm and an

index estimation algorithm. The noise estimation algorithm is denoted as

$$d(t_1(i), t_2) : \mathcal{T}_1 \times \mathcal{T}_2 \to \mathcal{X}^n \cup \{e\}, \tag{3.6}$$

that for each individual index i, it computes an estimated noise under the hypothesis that the query in the testing phase belongs to the ith individual. The estimated noise of the ith object is denoted as

$$\hat{\mathbf{z}}(i) = d(t_1(i), t_2). \tag{3.7}$$

If the algorithm fails for the ith index, subject to some criteria of failure depending on the system design, $d(\cdot, \cdot)$ outputs an error e. After the identification system completes noise estimation for all individual indexes, it proceeds to index estimation. Since an index j is chosen uniformly in the testing phase, for index estimation, the index estimation algorithm simply selects the index estimate $\hat{j}$ to be the index associated with the largest $P_{\mathbf{z}}(\hat{\mathbf{z}}(i))$. We define $P_{\mathbf{z}}(e) = 0$, so that the identification system always rejects indexes with noise estimation error. From now on in this paper, j always denotes the true object index selected in the test phase, and $i \in \{1, \cdots, M_c\} \setminus \{j\}$.

Note that $P_{\mathbf{z}}(\hat{\mathbf{z}}(i))$ is used to select the estimated index, instead of the joint probability mess of $P_{\mathbf{x},\mathbf{z}}$. This is because $\mathbf{x}$ is not directly stored in the database, estimating $\mathbf{x}$ may be difficult or impossible when the database compression rate is small.

Performance Measure

An identification system makes an error if $\hat{j} \neq j$. The average probability of error of an ensemble of identification system design is defined to be

$$P_e^n = \sum_{f,g,\mathcal{C},\mathbf{z}} P(\hat{j} \neq j | \mathcal{C}, \mathbf{z}, f, g) P_{\mathcal{C}}(\mathcal{C}) P_{\mathbf{z}}(\mathbf{z}) P_{f,g}(f, g), \tag{3.8}$$

which is averaging over all realizations of $\mathcal{C}$, $\mathbf{z}$, and the identification system. Note that $P_{f,g}(f, g)$ is necessary to consider when parts of the system are randomly generated, such as system designs in which random codes are employed, that the density is specified when the ensemble of identification

system designs is defined. Due to the i.i.d. assumption of templates, we have

$$P_{\mathcal{C}}(\mathcal{C}) = \prod_{\mathbf{x}\in\mathcal{C}} P_{\mathbf{x}}(\mathbf{x}). \tag{3.9}$$

Probability of error depends on the pattern length n, while we focus on the limiting cases where n is large.

Definition 3.1 A three rate tuple (R_c, L_1, L_2) is said to be achievable in an environment $\mathcal{E}$ if there exists a sequence of identification systems (f^n, g^n, ϕ^n) such that P_e^n goes to zero as n goes to infinity.

Definition 3.2 The set of all achievable rate tuple of a system is defined as the achievable rate region of this system.

We are interested in system designs that yields large achievable rate region. An inner and an outer achievable rate region bounds of this framework was obtained by Westover and O'Sullivan [86] and the previous chapter.

3.1.1 Truncation Encoding

In this section, we describe a simple encoding strategy called truncation encoding to illustrate four ideas:

1 The concept of "noise estimation" and how it can be used for identification.

2 A performance level that is very close to theoretical limit can be achieved in a simple method that per individual computation complexity is very low.

3 Good identification performance is possible even when reconstruction of the original template given *both* helper data and the query is impossible.

4 Robust identification against modeling error of the noise is possible.

In this sub-section, it is assumed that each element of a pattern sequence is independent and identically distributed (i.i.d.) drawn from a distribution Q_x on $GF(r)$, where $GF(r)$ denotes a Golid Field

of r elements. Similarly, each element of the noise vector is i.i.d. drawn from Q_z on the same $GF(r)$. Let $H = [I_{nT_1} 0]$ and $G = [I_{nT_2} 0]$, where I_{nT_1} and I_{nT_2} are identity matrices of size nT_1 and nT_2 respectively. Thus $t_1(j)$ is the first nT_1 elements of $\mathbf{x}(j)$, and t_2 is the first nT_2 elements of $\mathbf{y} = \mathbf{x}_j + \mathbf{z}$. Let $n_{\min} = \min(nT_1, nT_2)$. For any length n sequence a, $a_{n_{\min}}$ denotes the sequence of the first $n_{\min}$ elements of a, and $a_{n^c_{\min}}$ denotes the rest of a. By definition, we know that $T_{1,n_{min}(i)} = \mathbf{x}_{n_{min}}(i)$ and $t_{2,n_{min}} = \mathbf{y}_{n_{min}}$.

The noise estimation algorithm works as follows. For each pair of $(t_1(i), t_2)$, the algorithm checks if $(t_{n_{min}}(i), t_{2,n_{min}})$ is in the jointly typical set $T^{xy,\epsilon}_{n_{\min}}$, where the jointly typical set $T^{xy}_{n_{\min}}$ is defined as

$$\begin{aligned} T^{xy,\epsilon}_{n_{\min}} \quad = \quad & \{(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^{n_{\min}} \times \mathcal{X}^{n_{\min}} : \\ & \left| -\frac{1}{n_{\min}} \log P(\mathbf{x}) - H(Q_x) \right| < \epsilon \\ & \left| -\frac{1}{n_{\min}} \log P(\mathbf{y}) - H(Q_x * Q_{\mathbf{z}}) \right| < \epsilon \\ & \left| -\frac{1}{n_{\min}} \log P(\mathbf{x}, \mathbf{y}) - H(Q_x) - H(Q_z) \right| < \epsilon \}, \end{aligned} \tag{3.10}$$

where $Q_x * Q_z$ denotes the output distribution of a noisy channel with input distribution Q_x and additive noise distribution Q_z. It proceeds if $(t_{1,n_{min}(i)}, t_{2,n_{min}}) \in T^{xy,\epsilon}_{n_{\min}}$, otherwise it outputs an e indicating an error. The algorithm computes

$$\hat{z}_{i,n_{\min}} = t_{2,n_{\min}} - t_{1,n_{\min}}(i) = x_{i,n_{\min}} - x_{j,n_{\min}} + z_{n_{\min}}, \tag{3.11}$$

and then concatenates it with $n - n_{min}$ zeros to get the estimated noise $\hat{z}(j)$. Finally, the system selects the index

$$\hat{j} = \arg \max_{k \in \{1,2,\cdots,M_c\}} P_z(\hat{z}_k) = \arg \max_{k \in \{1,2,\cdots,M_c\}} P_z(\hat{z}_{k,n_{\min}}) \tag{3.12}$$

as its estimated index.

To see how truncation encoding works for identification, consider the following example. Let template pattern A and template Pattern B be realization of a binary Bernoulli 1/2 source as shown in Figure 3.1.1. The additive noise elements are assumed to be drawn i.i.d. from Bernoulli 1/10. Let $L_1 = L_2$ thus the truncation encoding cuts off the right side of each pattern and stores the result in the

database. Then it subtract each pattern from the truncated sensor input and obtain the results. As we can see, even though half of the images are not available, one can still infer that the query data is from individual A, as the estimated noise is a more "typical" pattern of Bernoulli 1/10, and on the other hand if the query is from individual B, it is a very unusual or "atypical" situation. The following theorem provides the quantitative side of this intuition.

Theorem 3.1 The probability of $\hat{j} \neq j$ goes to zero as n goes to infinity if

$$R_c < \min(T_1, T_2)(H(Q_x * Q_z) - H(Q_z) - 3\epsilon). \tag{3.13}$$

Proof

There are two situations under which the truncation encoding identification system makes an error:

- The first situation is when $(t_{1,n_{min}}(j), t_{2,n_{min}})$ is not in $T^{xy,\epsilon}_{n_{\min}}$.
- The second situation is when $(t_{1,n_{min}}(j), t_{2,n_{min}}) \in T^{xy,\epsilon}_{n_{\min}}$ but there exists at least one other object index i such that $(t_{1,n_{min}}j), t_{2,n_{min}}) \in T^{xy,\epsilon}_{n_{\min}}$ and $P(\hat{z}(j)) \geq P(\hat{z}_j)$.

The probability of the first situation goes to ϵ as n goes large, and ϵ can be chosen to be arbitrarily small because of the standard properties of jointly typical set and the law of large number. The probability of the second situation can be upper bounded by the probability that there exists at least one other individual index i with $(t_{1,n_{min(i)}}, t_{2,n_{min}}) \in T^{xy,\epsilon}_{n_{\min}}$, regardless if the resulting noise estimate is more likily than the true index or not. Hence the probability of the second condition is bounded by

$$\begin{aligned}
& \sum_{j\in 1,2,\cdots,M_c} P(j) \sum_{\mathbf{z}\in\{0,1\}^n} P(\mathbf{z}) \\
& \qquad P\left(\exists i : (\mathbf{x}_{n_{\min}(i)}, \mathbf{y}_{n_{min}}) \in T^{xy,\epsilon}_{n_{\min}} | z, i\right) \qquad (3.14) \\
\overset{(a)}{=} & \sum_{\mathbf{z}\in\{0,1\}^n} P(\mathbf{z}) P(\exists i : (\mathbf{x}_{n_{\min}}(i), \mathbf{y}_{2,n_{min}}) \in T^{xy,\epsilon}_{n_{\min}} | z) \\
= & \sum_{\mathbf{z}\in\{0,1\}^n} P(\mathbf{z}^c_{n_{\min}}) P(\mathbf{z}_{n_{\min}}) P(\exists i : (t_{1,n_{\min}}(i), t_{2,n_{min}}) \in T^{xy,\epsilon}_{n_{\min}} | \mathbf{z}_{n_{\min}}) \\
= & \sum_{\mathbf{z}^c_{n_{\min}}\in\{0,1\}^{n-n_{\min}}} P(\mathbf{z}^c_{n_{\min}}) \sum_{\mathbf{z}_{n_{\min}}\in\{0,1\}^{n_{\min}}} P(\mathbf{z}_{n_{\min}})
\end{aligned}$$

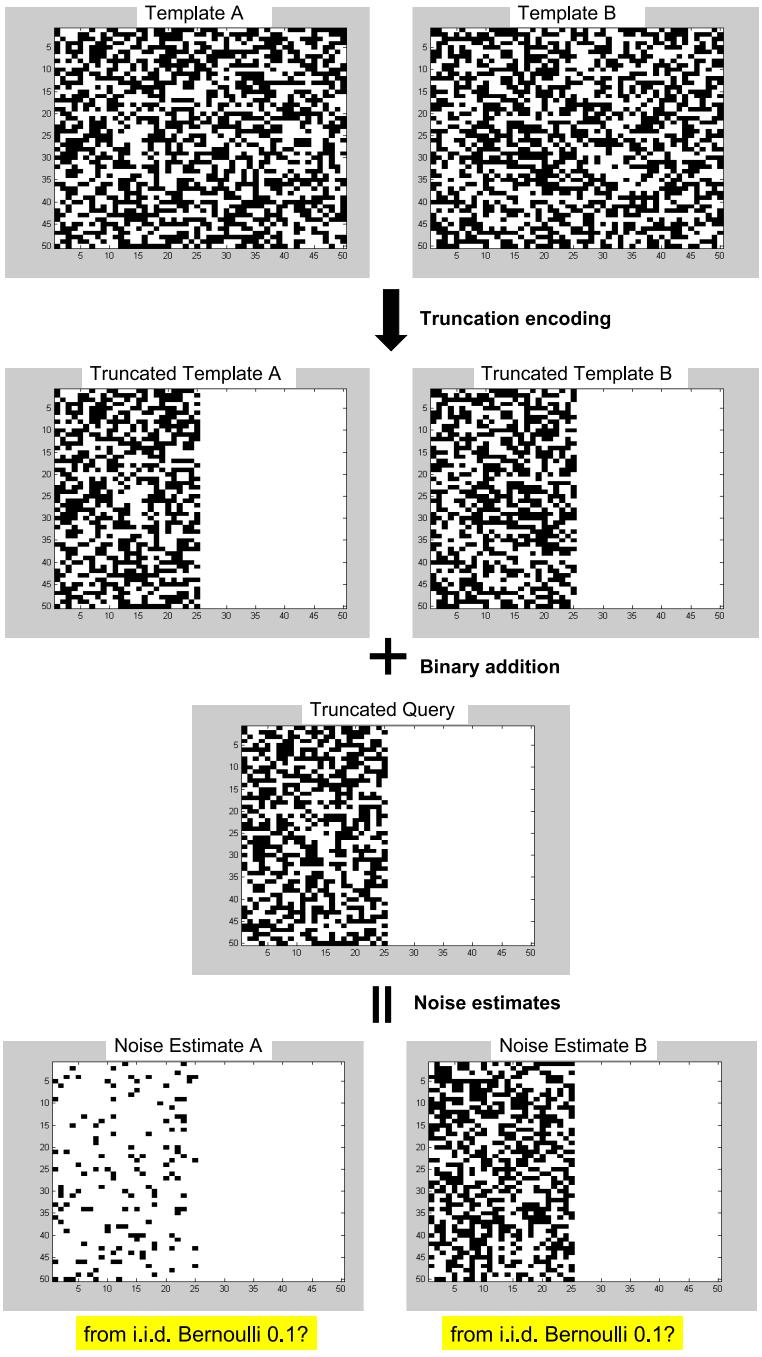

Figure 3.2: Two Bernoulli 1/2 template patterns at the top, left as A and right as B and their truncated versions at the second row. Both stored helper data add with the truncated query pattern resulting in two noise estimates. The last step is to check if resulting noise estimates are typical to the known noise distribution.

$$P(\exists i : (t_{1,n_{\min}}(i), t_{2,n_{min}}) \in T^{xy,\epsilon}_{n_{\min}} | \mathbf{z}_{n_{\min}}) \tag{3.15}$$

$$= \sum_{\mathbf{z}^c_{n_{\min}} \in \{0,1\}^{n-n_{\min}}} P(\mathbf{z}^c_{n_{\min}}) P(\exists i : (t_{1,n_{\min}}(i), t_{2,n_{min}}) \in T^{xy,\epsilon}_{n_{\min}})$$

$$\overset{(b)}{<} \sum_{\mathbf{z}^c_{n_{\min}} \in \{0,1\}^{n-n_{\min}}} P(\mathbf{z}^c_{n_{\min}}) \left(\sum_{i=\{2,3,\cdots,M_c\}} P\left((\mathbf{x}_{n_{\min}}(i), \mathbf{y}_{n_{min}}) \in T^{xy,\epsilon}_{n_{\min}}\right) \right)$$

$$\overset{(c)}{=} (M_c - 1)\, P\left((\mathbf{x}_{n_{\min}}(i), \mathbf{y}_{n_{min}}) \in T^{xy,\epsilon}_{n_{\min}}\right) \tag{3.16}$$

$$\overset{(d)}{<} (1+\epsilon) 2^{nR_c} 2^{-n_{\min}(I(X;Y)-3\epsilon)} \tag{3.17}$$

$$\overset{(e)}{=} (1+\epsilon) 2^{-n(\min(T_1,T_2)(H(Q_x * Q_z) - H(Q_z)) - R_c - 3\epsilon)}, \tag{3.18}$$

where

(a) follows from that j is uniformly distributed and all $\mathbf{x}(j)$ is independently drawn from the same distribution;

(b) follows from taking the union bound;

(c) follows from that the terms inside the parenthesis of (3.16) is independent of $\mathbf{z}^c_{n_{\min}}$;

(d) follows from the property of jointly typical set under the condition that if $\mathbf{x}_{n_{\min}}(i)$ and $\mathbf{y}_{n_{\min}}(j)$ are independent with the same marginals as $P(\mathbf{x}_{n_{\min}}(j), \mathbf{y}_{n_{\min}}(j))$, then the probability that $(\mathbf{x}_{n_{\min}}(i), \mathbf{y}_{n_{\min}}) \in T^{xy,\epsilon}_{n_{\min}} \leq 2^{-(I(X^{n_{\min}};Y^{n_{\min}})-3\epsilon)}$ [20], and elements of $\mathbf{x}_{n_{\min}}(j)$ and $\mathbf{z}_{n_{\min}}$ are i.i.d. hence so are elements of $\mathbf{y}_{n_{\min}}$;

(e) follows from

$$I(X;Y) = H(X+Z) - H(X+Z|X) \tag{3.19}$$

$$= H(Q_x * Q_z) - H(Q_z). \tag{3.20}$$

Thus if

$$R_c < \min(T_1, T_2)(H(Q_x * Q_z) - H(Q_z)) - 3\epsilon, \tag{3.21}$$

The probability of identification error goes to zero as n goes to infinity.

Corollary In particular, if elements of $\mathbf{x}(j)$ are drawn from i.i.d. Bernoulli $\frac{1}{2}$, and noise is from i.i.d.

Bernoulli q, we have the lower bound of possible R_c to be

$$R_c < \min(T_1, T_2)(1 - H(q)) - 3\epsilon, \tag{3.22}$$

where $0 \leq \min(T_1, T_2) \leq 1$ and $H(q) \leq 1$.

Notice that for $T_1 = T_2 = R$, the bound (3.22) of R_c is very close to the theoretical inner and outer bounds computed by Westover [87] and Westover and O'Sullivan [88], where they have shown an outer bound which is a concave function of R and is slightly above but very close to the straight line $R(1 - H(q))$.

Robustness of Truncation Encoding

Here we discuss another interesting example where the noise distribution Q_z is partially known. We assume that each element of $\mathbf{x}(j)$ is i.i.d. drawn from the uniform distribution over $GF(r)$. We assume that each element of $\mathbf{z}$ is i.i.d. drawn from a distribution Q_z, but only $Q_z(0) = 1 - q$ is known that each element of $\mathbf{z}$ takes value 0 with probability $1 - q$ but the probability mess of other values are unknown. We want to find the least upper bound on R_c among all such distributions given $R = \min(T_1, T_2)$ using truncation encoding. This is a constrained optimization problem

$$\max_{Q_z} H(Q_z) \text{ subject to } \sum_{k \in GF(r)} q_k = q, q_k \geq 0 \ \forall k \tag{3.23}$$

where $q_k = Q_z(k)$. The maximum can easily be shown to be achieved for $q_k = \frac{q}{r-1} \ \forall k \neq 0$. The least upper bound of R_c is then

$$R\left(\log r + (1-q)\log(1-q) + q\log\left(\frac{q}{r-1}\right)\right), \tag{3.24}$$

where all logarithms are taken base 2.

Identification versus Reconstruction

Note that there are noticeable differences between identification and lossless source coding with side information. The bits useful in identification systems are different from bits useful for lossless source coding. Also even if a joint lossless source code is available, it might not be good for identification. Given two correlated sequences $\mathbf{x}$ and $\mathbf{y}$, the achievable rate region of lossless source codes with side information obtained by Ahlswede and Körner [5] is

$$R_x \geq H(X|V), \tag{3.25}$$

$$R_y \geq I(Y;V), \tag{3.26}$$

where V is an auxiliary random variable and $X-Y-V$ is a Markov chain. For $\mathbf{x}$ being Bernoulli $\frac{1}{2}$, and $\mathbf{y}=\mathbf{x}+\mathbf{z}$ where $\mathbf{z}$ is Bernoulli q, $R_y=1$ and $R_x=H(q)$ is an achievable rate pair to reconstruct $\mathbf{x}$ and hence reconstruct z. However, Theorem 1 shows that it is not always necessary to reconstruct entire $\mathbf{x}$ or $\mathbf{z}$ for identification. Also theorem 1, [87], and [88] all show that even if lossless coding is possible for a given identification system with $T_1=R_x, T_2=R_y$, it is not good for identification if the compression rates are below the required bounds. A large sensor compression rate $T_2=R_y=1$ alone does not yield good performance because even if it is sufficient to reconstruct the true noise $\mathbf{z}$, it is not sufficient to suppress the probability that there exists another pattern which is jointly typical with a sequence matching the compressed database and sensor data. From a linear coding point of view with G for encoding $\mathbf{x}$, the above argument means that the cardinality of each coset of G is too large to prevent that for all the $2^{R_c}-1$ false objects, the coset $G(\mathbf{x}(j)+\mathbf{y})$ does not contain a sequence which is jointly typical with $\mathbf{x}(j)$.

3.1.2 Identification System Designs Using Linear Codes for General Additive Noise Models

Although the truncation encoding works well for i.i.d. Bernoulli patterns under i.i.d. Bernoulli noise conditions, we shall see that identification systems using linear codes or ensemble of linear codes, can be proven to work reasonably well in general noise models. To see this, let us assume that elements of patterns are i.i.d. drawn from the uniform distribution over $GF(r)$, denoted as $\bar{Q}_x$. The additive

noise sequence is drawn from a distribution whose mean entropy is nR_z for some $0 < R_z < 1$. Under this loose constraint which allows nonstationary noise distributions, it might not be sufficient to have good statistical properties for identification by simply computing the first $n_{\min}$ elements of the noise sequence. Notice that when an low density parity check (LDPC) matrix is used for compression, the codes used are viewed as low density generator matrix (LDGM) codes, which are also known to have good performance for source coding and channel coding [32] [95]. Under the pattern and noise assumptions stated above, one can use LDPC codes to design good identification systems as stated in the following Theorem 2.

By *good* ensemble for generating LDPC matrices, we mean that the ensemble and noise average block decoding error goes to zero as n goes to infinity. By *good identification system design* we mean that the ensemble and noise average identification error goes to zero as n gets large.

Theorem 3.2: If there exists a *good* ensemble for generating LDPC matrices of rate $R = \min(T_1, T_2)$, alone with a syndrome decoding algorithm under a noise distribution with entropy nR_z, then there exists a *good identification system design* using the same LDPC matrix ensemble and syndrome decoding algorithm for all $R_c < \min(T_1, T_2) - R_z$.

The proof is omitted since it follows directly from the following Theorem 3, as LDPC codes being a special case of linear codes.

Theorem 3.3 If there exists a good ensemble of linear codes of rate $R = \min(T_1, T_2)$ and a decoding algorithm for a noise distribution with entropy nR_z. Then for all $R_c < \min(T_1, T_2) - R_z$, there exists a good pattern identification system design using the generator matrix of the linear block code, and the decoding algorithm as noise estimation algorithm under the same noise distribution.

Proof To prove theorem 3, we start by constructing a system design that utilizes a good linear code and the corresponding decoder, and then we show that the system achieves the stated performance. Without loss of generality, let us assume that $T_1 \leq T_2$. Database compression is done by using H, denoting a parity check matrix generated by the "good" linear code ensemble, such that $t_1(j) = H\mathbf{x}(j)$. sensor compression is done by a matrix $G = [H^T 0]^T$, where 0 is simply the zero matrix. Let $d(\cdot, \cdot)$ denotes the syndrome decoding algorithm associated with the linear code ensemble. For

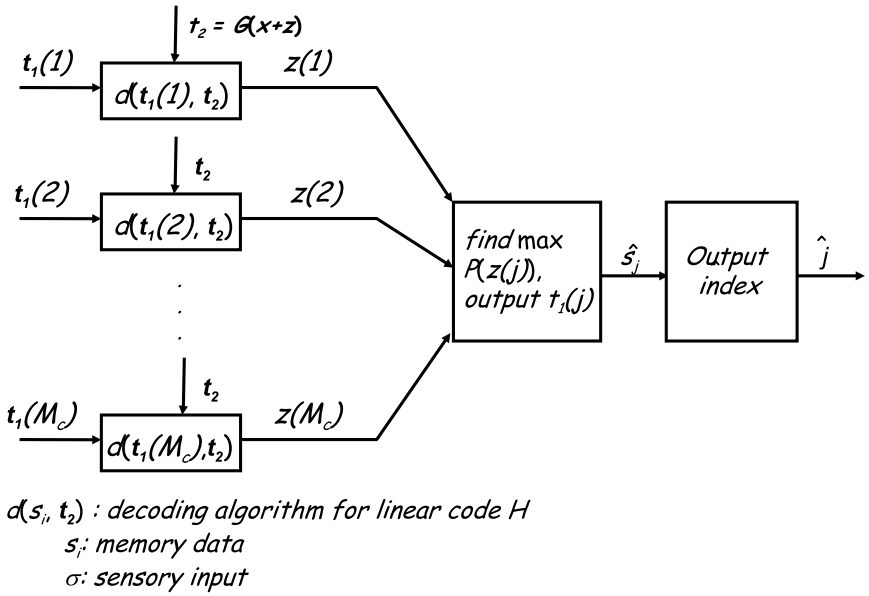

Figure 3.3: Identification using a linear code and its decoder

each individual index i, the identification computes $\hat{\mathbf{z}}(i) = d(t_1(i), t_2)$, and selects i that maximizes $P(\hat{\mathbf{z}}(i))$. The system makes an error if for the true index j, there exist some i such that $P(\hat{\mathbf{z}}(i)) \leq P(\hat{\mathbf{z}}(j))$.

Similar to the proof of Theorem 1, we bound the probability of error of this system from above by the probability that there exists an index i, the resulting noise estimate $\hat{\mathbf{z}}(i)$ is typical, or for the true index j, $\hat{\mathbf{z}}(i)$ is not typical. The typical set $T_n^{z,\epsilon}$ is defined as

$$T_n^{z,\epsilon} = \{\mathbf{z} : \left| \frac{1}{n} \log P(\mathbf{z}) - R_z \right| < \epsilon\} \tag{3.27}$$

Because the probability of $\mathbf{z}(j) \notin T_n^{\mathbf{z}(j),\epsilon}$ is ϵ which can be chosen to be arbitrarily small as n gets large, and the decoding algorithm for inferring $\hat{\mathbf{z}_j}$ is good that $\hat{\mathbf{z}_j} = \mathbf{z}(j)$ almost surely, we focus on the probability of index estimation error that there exist some $\hat{\mathbf{z}}(i) \in T_n^{z,\epsilon}$. The probability of index

estimation error is less than

$$\sum_{j=1}^{M_c} P(j) \sum_{\mathbf{z}\in T_n^{z,\epsilon}} P(\mathbf{z}) P\left(\exists i : \hat{\mathbf{z}}(j) \in T_n^{z,\epsilon} | \mathbf{z}, i\right) \tag{3.28}$$

$$= \sum_{\mathbf{z}\in T_n^{z,\epsilon}} P(\mathbf{z}) P\left(\exists i : \hat{\mathbf{z}}(j) \in T_n^{z,\epsilon} | \mathbf{z}\right) \tag{3.29}$$

$$= \sum_{\mathbf{z}\in T_n^{z,\epsilon}} P(\mathbf{z}) P\left(\exists i : d(t_1(j), t_2)) \in T_n^{z,\epsilon} | \mathbf{z}\right) \tag{3.30}$$

$$\stackrel{(a)}{=} \sum_{\mathbf{z}\in T_n^{z,\epsilon}} P(\mathbf{z}) P(\exists i : d(0, H(\mathbf{x}(j) - \mathbf{x}(1) + \mathbf{z})) \in T_n^{z,\epsilon} | \mathbf{z})$$

$$\stackrel{(b)}{=} \sum_{\mathbf{z}\in T_n^{z,\epsilon}} P(z) P\left(\exists i : d(0, H(\tilde{\mathbf{x}})) \in T_n^{z,\epsilon}\right) \tag{3.31}$$

$$\stackrel{(c)}{\leq} 2^{nR_c} \sum_{z\in T_n^{z,\epsilon}} P(\mathbf{z}) P\left(d(0, H(\tilde{\mathbf{x}})) \in T_n^{z,\epsilon}\right) \tag{3.32}$$

$$\stackrel{(d)}{\leq} 2^{nR_c} P\left(d(0, H(\tilde{\mathbf{x}})) \in T_n^{z,\epsilon}\right) \tag{3.33}$$

$$\leq 2^{nR_c} \sum_{\tilde{\mathbf{z}}\in T_n^{z,\epsilon}} P(H\tilde{\mathbf{x}} = H\tilde{\mathbf{z}} | \tilde{\mathbf{z}}) \tag{3.34}$$

$$= 2^{nR_c} \sum_{\tilde{\mathbf{z}}\in T_n^{\mathbf{z},\epsilon}} 2^{-nT_1} \tag{3.35}$$

$$\stackrel{(e)}{\leq} 2^{-n(T_1 - R_z - R_c - \epsilon)}, \tag{3.36}$$

where

(a) follows from the construction of G based on H.

(b) is because elements of $\mathbf{x}(j)$ and $\mathbf{x}(1)$ both are i.i.d. from the uniform distribution over $GF(r)$, and $\mathbf{x}(j)$ and $\mathbf{x}(1)$ are independent of each other and independent of $\mathbf{z}$, so that elements of $\mathbf{x}(j) - \mathbf{x}(1) + \mathbf{z}$ are also i.i.d. and uniformly distributed, denoted as $\tilde{\mathbf{x}}$;

(c) follows from union bound and there are totally $2^{nR_c} - 1$ terms in the sum;

(d) follows from that $\tilde{\mathbf{x}}$ is independent of $\mathbf{z}$, see (b);

(e) The cardinality of $T_n^{\mathbf{z},\epsilon}$ has upper bound $2^{n(R_z+\epsilon)}$). Hence the probability of index estimation error goes to zero as n goes to infinity if

$$R_c < \min(T_1, T_2) - R_z - \epsilon. \tag{3.37}$$

Note that clearly if the complexity of the decoding algorithm is $O(f(n))$, the complexity of the identification system per object is also $O(f(n))$. Hence Theorem 3 not only connects good linear code design to good identification system design, it also connects low complexity algorithms for decoding linear code to noise estimation in identification systems.

LDPC codes can be used for non-i.i.d. noise. For example, Eckford, Kschischang, and Pasupathy [24] analyzed LDPC codes for Gilbert-Elliot Channels, which are binary symmetric channels with crossover probability depending on Markov processes, and Nicola, Alajaji, and Linder [58] developed decoding algorithms for LDPC codes with a queue-based channel. Based on Theorem 3 and [60], LDPC codes with the algorithms by [24, 58] can be used for good identification system design for those noise models in identification systems.

3.2 Optimal Trade-off Between Identification and Secrecy-Key Binding Using Linear Codes

We now consider the trade-off between biometric identification capacity and secret binding capacity. As discussed in the previous chapter, this is an important template protection design under the catagory of biometric cryptosystems. We focus on designing systems that can be proved to achieve good performance utilizing good linear codes. In this problem, for each individual enrolled, a template of a biometric trait is measured and a *secret* is selected, independent of the identity and measured template. Both the template and the secret are used to generate the *helper data* which is stored in the database. Note that the template and secret are themselves not stored in the database. During identification, the biometric system receives a query signal. The goals of the system are to infer the identity to which this query belongs, and recover the secret of this individual. The system requirements are that the helper data contains a negligible amount of information about the secret, and the identification error should be small. The questions of interest are how many individuals can be enrolled in the systems, how many secrets can be generated, and what the trade-off between them are. The overall system is illustrated in Fig. 3.4. Note that only attacks at the database are considered in this section, which is a degenerated case of the model discussed in the previous chapter.

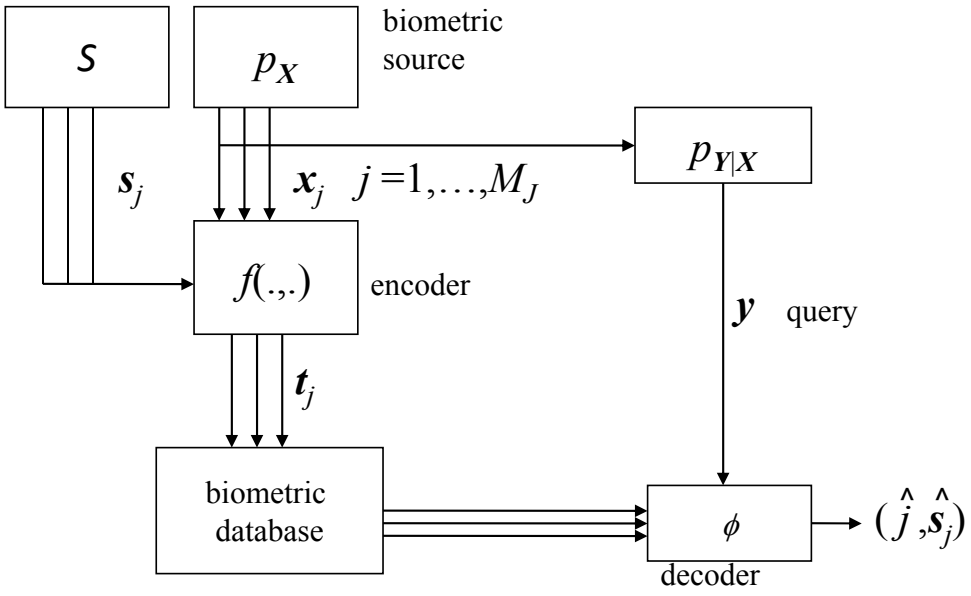

Figure 3.4: The identification and secret binding system

In the following sections, capital case denotes the random variables, lower case denotes realizations of random variables, and bold case denotes sequences and vectors. The biometric template of each individual is denoted as a length n sequence $\mathbf{x}_j = (x_{j1}, \cdots, x_{jn}) \in \mathcal{X}^n$ where $j \in \{1, \cdots, M_J\}$ is the index of the individual and M_J is the total number of individual enrolled in the system. Each template $\mathbf{x}_j$ is modeled as a realization drawn independent and identically from a biometric source random process governed by a sequence of distributions $p^n_{\mathbf{X}}$, whose entropy rate is assumed to exist. Note that $\mathbf{X}$ is independent of the index j. The query signal $\mathbf{y} \in \mathcal{Y}^n$ is modeled as a noisy version of a biometric template drawn from a sequence of distributions $p^n_{\mathbf{Y}|\mathbf{X}}$, that the query signal $\mathbf{y}$ can be thought of as the output of a template $\mathbf{x}$ passing through a noisy channel. When the noise is additive and independent of the source, we denote the additive noise as $\mathbf{z}$ drawn from a sequence of distribution $p^n_{\mathbf{Z}}$ such that $\mathbf{y} = \mathbf{x} + \mathbf{z}$. Throughout this paper, we focus on discrete cases, and $\mathcal{X}$ and $\mathcal{Y}$ are fields. In practice, they may be integers as a result of quantization. Each individual is assumed to be equally likely to present the query input that $p(i) = \frac{1}{M_s}$.

For each individual j, one template is measured and a secret is uniformly drawn from a set $\mathcal{S}$ independent of one another, the individual indexes j, and the templates $\mathbf{X}(j)$. The cardinality of $\mathcal{S}$ is denoted as M_s. A large M_s suggests that it is harder to guess the secret of an individual and hence the system is safer. The measured template and selected secret are used to generate the helper data which will be enrolled and stored in the database with the index. This process is modeled as the encoding part of the system: an encoder function f takes an input $\mathbf{x}(j)$ and a secret $\mathbf{s}(j) \in \mathcal{S}$ to produce the helper data

$$\mathbf{t}(j) = f(\mathbf{s}(j), \mathbf{x}(j)). \tag{3.38}$$

The identification process is modeled as a decoder mapping g which takes the query $\mathbf{y}$ along with $\mathcal{S}$ to produce an estimate of subject index $\hat{i}$ and an estimate of the secret $\mathbf{s}(\hat{j})$

$$(\hat{j}, \hat{\mathbf{s}_j}) = \phi\left(\mathbf{y}, \mathbf{t}(1), \cdots, \mathbf{t}(n)\right). \tag{3.39}$$

Note that the decoder does not have any knowledge about what are the realizations of the secrets, but only the set $\mathcal{S}$ from which they are drawn. The system requirements of small information leakage about the secret from helper data, small identification error, and design for a large number of individuals and secret choices can then be cast as an information theoretic problem.

Definition 1 Achievability: A pair of identification rate and secrecy key rate (R_J, T_2) is achievable in the biometric identification and secret binding setting if for all $\delta > 0$ and for n large enough, there exists an encoder $f(\cdot)$ and a decoder $\phi(\cdot)$ such that:

$$\begin{aligned}
\Pr\left((\hat{j}, \mathbf{s}(\hat{j})) \neq (j, \mathbf{s}(j))\right) &\leq \delta && (3.40)\\
\frac{1}{n}\log M_J &\geq R_J - \delta && (3.41)\\
\frac{1}{n}\log M_S &\geq T_2 - \delta && (3.42)\\
\frac{1}{n} I(\mathbf{S}(j); \mathbf{T}(j)) &\leq \delta \quad \forall j. && (3.43)
\end{aligned}$$

The achievable rate region is the set of all achievable rate pairs and is denoted as $\mathcal{R}$. In the work of Ignatenko and Willems [36, 89], the achievable rate region of i.i.d. sources with independent i.i.d.

additive noise is proved to be:

$$\mathcal{R} = \{(R_J, T_2) : 0 \leq R_J + T_2 \leq I(X;Y)\}. \tag{3.44}$$

Note that the X and Y are not in bold case, as the source and noise are assumed to be i.i.d. in their derivation. In particular, they also show that for i.i.d. Bernoulli $\frac{1}{2}$ source with i.i.d. Bernoulli noise q independent of the source, a simple linear code is able to achieve the optimal rate region [89].

3.2.1 Summary of Results

We summarize the results and contribution of this section as three main theorems. The entropy rates of distributions $p^n_{\mathbf{X}}, p^n_{\mathbf{Y}}, p^n_{\mathbf{Z}}, p^n_{\mathbf{X},\mathbf{Y}}$, and the conditionals $p^n_{\mathbf{X}|\mathbf{Y}}, p^n_{\mathbf{Y}|\mathbf{X}}$ are assumed to exist. Before we present the key theorems, we introduce the following three definitions.

Definition 3.3 A sequence of source codes is said to be *good* for a sequence of source $p^n_{\mathbf{X}}$ if it asymptotically achieves the optimal lossless source coding rate $H(\mathcal{X})$.

Definition 3.4 A sequence of decoders is said to be *good* for a good sequence of source codes of $p^n_{\mathbf{X}}$ if it can reconstruct the typical sequences of the sequence of $p^n_{\mathbf{X}}$ with probability $1 - \epsilon(n)$, where $\epsilon(n)$ approaches zero as n gets large.

Definition 3.5 A good sequence of universal source codes for two sources $p^n_{\mathbf{X}}$ and $p^n_{\mathbf{Z}}$ asymptotically achieves optimal lossless source coding rates for both sources.

The main results are as follows.

Theorem 3.4 For arbitrary $p^n_{\mathbf{X},\mathbf{Y}}$ with all entropy rates assumed to exist, the achievable rate region of identification and secret binding trade-off is

$$\mathcal{R}_a = \left\{(R_J, T_2) : R_I + T_2 \leq \frac{1}{n} I(\mathbf{X};\mathbf{Y})\right\}. \tag{3.45}$$

Comment: The converse proof is a direct generalization of Ignatenko and Willems [35, 89], and a special case of the result in the previous chapter.

Theorem 3.5 If there exists a *good* code for a source $p^n_{\mathbf{X}}$ with a *good* decoder, a system can be designed to achieve optimal identification and secret binding trade-off as in eq. (3.45) for arbitrary $p^n_{\mathbf{Y}|\mathbf{X}}$ if the entropy rates exist.

Comment: Theorem 5 requires a good decoder for source $\mathbf{X}$, whose entropy rate may be high. In practice, there may not be a low complexity decoder available. Thus one may choose a suboptimal system design that only requires a simpler decoder, which motivates the next result.

Theorem 3.6 Assume the query noise is additive and independent of the template source. If there exists a *good universal* code for sources $p^n_{\mathbf{X}}$ and $p^n_{\mathbf{Z}}$ with *good* decoder only for $p^n_{\mathbf{Z}}$, then a system can be designed to achieve the rate region

$$\mathcal{R}_2 = \{(R_J, T_2) : R_J + T_2 \leq H(\mathcal{X}) - H(\mathcal{Z})\}, \tag{3.46}$$

where $H(\mathcal{X})$ and $H(\mathcal{Z})$ denote the entropy rates of $p^n_{\mathbf{X}}$ and $p^n_{\mathbf{Z}}$.

Comment: This Theorem requires only a decoder which is good for $\mathbf{Z}$, whose entropy rate is usually smaller than that of $\mathbf{X}$ in practice. It is easier to build a low complexity practical decoder for a smaller entropy rate distribution. However there is a small loss in the performance. Several results on universal linear codes are available, such as the seminal paper by Csiszár [21]. As shown by Theorem 1, we know that $R_2 \subseteq R_a$. For example, for binary symmetric Bernoulli source p and binary symmetric noise q, it is easy to show that R_2 is strictly contained in R_a.

3.2.2 Converse for Theorem 4

The converse proof presented here is a generalization of the converse proof for i.i.d. source and noise models proved derived by Willems and Ignatenko [89]. We observe that their proof can be readily applied to arbitrary source and noise distributions as long as their entropy rates exist. Despite the

result is a special case of the general result in Chapter 2, we here present a specific proof for the converse of Theorem 4, while is easier to understand.

We are interested in the error probability $P_e = P\left((\hat{j}, \mathbf{s}(\hat{j})) \neq (j, \mathbf{s}(j))\right)$, where $(\hat{j}, \mathbf{s}(\hat{j})) = \phi\left(\mathbf{y}, \mathbf{t}(1), \cdots, \mathbf{t}(n)\right)$. From Fano's inequality we know

$$\begin{aligned} H(j, \mathbf{s}(j)|\hat{J}, \hat{s_J}) &\leq 1 + P_e \log |M_J||M_S| \\ &= 1 + P_e n(R_J + T_2) \\ &\leq 1 + \delta n(R_J + T_2), \end{aligned} \tag{3.47}$$

where the last inequality follows from the error constraint of system. The converse proof is derived as follows:

$$\begin{aligned} H(J, \mathbf{S}(j)) &\overset{(a)}{=} I(J, \mathbf{S}(j); \mathbf{T}(1), \cdots, \mathbf{T}(M_J), \mathbf{Y}) + H(J, \mathbf{S}(j)|\mathbf{T}(1), \cdots, \mathbf{T}(M_J), \mathbf{Y}, \hat{J}, \hat{S_J}) \\ &\overset{(b)}{\leq} I(J, \mathbf{S}(j); \mathbf{T}(1), \cdots, \mathbf{T}(M_J), \mathbf{Y}) + H(J, \mathbf{S}(j)|\hat{J}, \hat{S_J}) \end{aligned}$$

The second term $H(J, \mathbf{S}_J|\hat{J}, \hat{S(J)})$ can be bounded using Fano's inequality. We then focus on the first term:

$$\begin{aligned} & I(J, \mathbf{S}(j); \mathbf{T}(1), \cdots, \mathbf{T}(M_J), \mathbf{Y}) \\ = \; & I(J, \mathbf{S}(j); \mathbf{T}(1), \cdots, \mathbf{T}(M_J)) + I(J, \mathbf{S}(j); \mathbf{Y}|\mathbf{T}(1), \cdots, \mathbf{T}(M_J)) \\ \overset{(c)}{\leq} \; & I(J, \mathbf{S}(j); \mathbf{T}(1), \cdots, \mathbf{T}(M_J)) + H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}, J, \mathbf{S}(j), \mathbf{T}_1, \cdots, \mathbf{T}(M_J)) \\ \overset{(d)}{=} \; & I(J, \mathbf{S}(j); \mathbf{T}(1), \cdots, \mathbf{T}(M(j))) + I(\mathbf{X}; \mathbf{Y}) \\ \overset{(e)}{=} \; & \frac{1}{M_J} \sum_{j=1}^{M(j)} I(\mathbf{S}(j); \mathbf{T}(j)) + I(\mathbf{X}; \mathbf{Y}) \\ \overset{(f)}{\leq} \; & n\delta + I(\mathbf{X}; \mathbf{Y}), \end{aligned} \tag{3.48}$$

where

(a) By the definition of mutual information, which is assumed to exist. Note that $\hat{J}, \hat{S(j)}$ is the decoder output which is determined by $H_1, \cdots, H_{M_J}, \mathbf{Y}$.

(b) Conditioning reduces entropy.

(c) Rewriting mutual information into entropy, and adding an $\mathbf{X}$ term where the inequality follows because conditioning reduces entropy. Assuming all terms exist.

(d) Given $\mathbf{X}$, $\mathbf{Y}$ is independent of the remaining terms.

(e) Only $\mathbf{T}(j)$ may contain information about $\mathbf{S}(j)$.

(f) From the system requirement that $I(\mathbf{S}(j); \mathbf{T}(j)) \leq \delta$.

Because $H(j, \mathbf{S}(j)) = \log M_J M_S = n(R_J + T_2)$, along with equations (3.47) and (3.48), we have

$$R_J + T_2 \leq \frac{1}{1-\delta}(\frac{1}{n} I(\mathbf{X}; \mathbf{Y}) + \delta + \frac{1}{n}). \tag{3.49}$$

When n gets large and δ gets to zero, we have

$$R_J + T_2 \leq \frac{1}{n} I(\mathbf{X}; \mathbf{Y}). \tag{3.50}$$

3.2.3 Concepts of Linear Code Designs for Identification and Secret Binding

We describe two linear coding approaches. The two approaches utilize the same encoding strategy; the difference lies in the decoding capacity requirements. The approach which achieves the optimal trade-off bound between identification and secret capacity would require a decoding method with higher computational complexity. The other approach requires less decoding capacity so that more practical decoders may be used for this system design, but it leads to a suboptimal trade-off.

To design a system preventing information about the secret from leaking through the helper data, the idea is to use the biometric template signal as a surrogate noise in a additive channel between the secret and the helper data, while the channel capacity should be zero. The idea is shown in Figure 3.5. Consider the secret and the helper data as two sequences of the same length. The secret is the channel input and the helper data is the channel output while the channel is the encoding function f whose noise is characterized by the biometric template. The best channel in this security context has

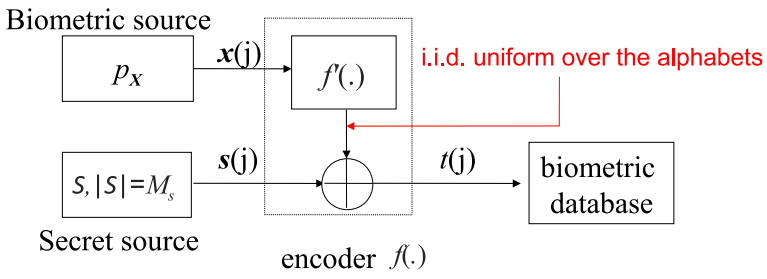

Figure 3.5: Biometric templates are used as worst case noise of an additive channel to eliminate mutual information between the secrets and the helper data.

the mutual information between the input and output is zero. One such channel is a channel of i.i.d. uniform additive noise. Thus to achieve small leakage, one should map the biometric templates to i.i.d. uniform sequences with a function f', and add the result to the input secret sequences. Mapping a typical sequence of a random process is analogous to optimal lossless source coding [94].

To identify individuals and estimate the secret, the system utilizes the correlation between the correct template and the query, and the performance can be proved through standard joint-typicality arguments and union bounds.

3.2.4 Encoding and Decoding

Based on the design concept described above, the encoding consists of two steps:

- Map each secret $s(j)$ onto sequences taking values over $mathcalX$. Since this does not change the properties of the problem, the notation $s(j)$ is kept to represent the resulting secret sequences.
- For each $\mathbf{x}(j)$, the helper data is computed as $\mathbf{t}(j) = G\mathbf{x}(j) + \mathbf{s}(j)$, where G is a $\frac{nH(\mathbf{X})}{\log|\mathcal{X}|}$ by n generator matrix with certain requirements described later and $H(\mathbf{X})$ is the entropy rate of $p^n_{\mathbf{x}}$.

There are two decoding approaches proposed in this section.

- *Optimal trade-off achieving approach*: the decoder checks if there exists a helper data $\mathbf{t}(\hat{j})$ and a sequence $\mathbf{s}(\hat{j}) \in \mathcal{S}$ such that from $\mathbf{t}(\hat{j}) - \mathbf{s}(\hat{j})$, the reconstructed template $\hat{\mathbf{x}}$ is *jointly typical* with the input query $\mathbf{y}$. The decoder then outputs the satisfying $\mathbf{t}(\hat{j})$ and $\mathbf{s}(\hat{j})$. If there are multiple candidates, the decoder picks one randomly.
- *Suboptinal trade-off with lower decoder computational complexity*: the decoder takes the query $\mathbf{y}$ and checks if there exists a helper data $\mathbf{t}_{\hat{j}}$, a sequence $\mathbf{s}(\hat{j})$ and sequence $\hat{\mathbf{z}}$ that is a *typical sequence* of $P^n_{\mathbf{Z}}$ and satisfies $G\hat{\mathbf{z}} = G\mathbf{y} - \mathbf{t}_{\hat{j}} - \mathbf{s}(\hat{j})$. The decoder outputs the corresponding $\hat{j}$ and $\mathbf{s}(\hat{j})$. If there are multiple candidates, the decoder picks one randomly.

We assume that the source has higher entropy rate than the noise that $H(\mathbf{X}) \geq H(\mathbf{Z})$. The conditions for the matrix G, or an ensemble of generating G, to achieve the optimal trade-off or near optimal trade-off with lower decoder complexity are

(1) it achieves near optimal lossless source coding of source $\mathbf{X}$, and

(2.a) most typical sequences $\mathbf{x}$ can be recovered from $G\mathbf{x}$.

(2.b) most typical sequences $\mathbf{z}$ can be recovered from $G\mathbf{z}$.

The first condition is needed to ensure that there is only a negligible amount of information leaked from $\mathbf{t}$ about $\mathbf{s}$, i.e. $I(\mathbf{T};\mathbf{S})$ approaches zero as n gets large. The first condition is required by both approaches. Condition (2.a) is only required by the optimal trade-off approach where the decoded $\mathbf{x}(\hat{j})$ has to be computed and checked for joint typicality with the query $\mathbf{y}$. Condition (2.b) is only

needed for the near optimal trade-off with lower computational demand, where computing the noise vector $\mathbf{z}$ is needed.

3.2.5 Secrecy Leakage Analysis

We have to check that $\mathbf{T}$ leaks negligible information about $\mathbf{S}$, and conditions when the probability of identification and secret estimation error approaches zero. For both approaches, the secrecy leakage analysis is the same. We show the achievable region and error analyses of both approaches separately.

By condition (1) of the previous subsection, each sequence of the set

$$G\mathbf{X} = \left\{ \mathbf{w} : \mathbf{w} = G\mathbf{x}, \forall \mathbf{x} \in \mathcal{T}^n_{P(\mathbf{X})} \right\}. \tag{3.51}$$

is nearly drawn from the i.i.d. uniform distribution over $\mathcal{X}$. This is because each typical sequence is almost equally likely and there are around $2^{nH(\mathbf{X})}$ typical sequences, which is the number of possible outcomes of G. Thus we have

$$\begin{aligned} H(\mathbf{S}|\mathbf{T}) &= H(\mathbf{S}|G\mathbf{X}+\mathbf{S}) \\ &\overset{(a)}{=} H(\mathbf{S}) + H(G\mathbf{X}) - H(\mathbf{S}+G\mathbf{X}) \\ &\overset{(b)}{\geq} H(\mathbf{S}) - n\delta \end{aligned}$$

where (a) is because $\mathbf{S}$ and $G\mathbf{X}$ are independent, and (b) is because $G\mathbf{X}$ is almost i.i.d. uniform over $\mathcal{X}$. Hence we have $\frac{1}{n}I(\mathbf{S};\mathbf{T}) \leq \delta$. This means that the helper data leaks a negligible amount of information about the secret.

3.2.6 Achievable Region of the Optimal Trade-off Approach

The system error comes from the following two conditions:

- For the correct individual j with all possible choices of $\mathbf{s}(\hat{j})$, there does not exist a reconstructed sequence $\hat{\mathbf{x}}(j)$ which is jointly typical with the query $\mathbf{y}$.
- There is another combination of individual l and secret $\mathbf{s} \in \mathcal{S}$ such that the reconstructed $\hat{\mathbf{x}}(l)$ is jointly typical with the input query $\mathbf{y}$.

The first type of error is bounded by the probability of the following events:

$$P\left((\mathbf{x}(j), \mathbf{y}) \notin \mathcal{T}^n_{p(\mathbf{X},\mathbf{Y})}\right) + P\left(G\mathbf{x} \notin \mathcal{D} | \, (\mathbf{x}(j), \mathbf{y}) \in \mathcal{T}^n_{p(\mathbf{X},\mathbf{Y})}\right), \tag{3.52}$$

where $\mathcal{D} \subseteq G\mathbf{X}$ is the set of decodable sequences by the decoding algorithm. The first term approaches zero as n gets large. The second is negligible because the decoder is good, so condition $(2.a)$ described in subsection 3.2.4 is satisfied.

The second type error e_2 is that there exists another $t_{\hat{l}}$ with a secret $\mathbf{s} \in \mathcal{S}$ so that $\hat{\mathbf{x}} = \phi(\hat{\mathbf{t}}(j), \mathbf{s}_k)$ is jointly typical with the query $\mathbf{y}$, where $\phi(\cdot, \cdot)$ is the decoding algorithm. We have:

$$\begin{aligned}
P(e_2) &\overset{(a)}{\leq} \sum_{(l,\mathbf{s}) \neq (j,\mathbf{s}(j))} P\left((\phi(\hat{\mathbf{t}}(l), \mathbf{s}), \mathbf{y}) \in \mathcal{T}^n_{p(\mathbf{X},\mathbf{Y})}\right) \\
&\overset{(b)}{=} \sum_{(l,\mathbf{s}) \neq (j,\mathbf{s}(j))} P\left((\mathbf{x}', \mathbf{y}) \in \mathcal{T}^n_{p(\mathbf{X},\mathbf{Y})}\right) \\
&\overset{(c)}{=} \sum_{(l,\mathbf{s}) \neq (j,\mathbf{s}(j))} 2^{-I(\mathbf{X};\mathbf{Y})+n\delta} \\
&= 2^{-n(\frac{1}{n}I(\mathbf{X};\mathbf{Y})-R_J-T_2)-\delta},
\end{aligned}$$

where $\mathbf{x}'$ is a sequence with distribution $p(\mathbf{X})$ but independent of $\mathbf{y}$, and

(a) the union bound is applied;

(b) follows from condition (1) in subsecition 3.2.4 and each typical sequence is almost equally likely;

(c) is due to the fact that if $\mathbf{x}'$ has the same distribution as the marginal of $\mathbf{x}$ but independent of $\mathbf{y}$, the probability that $\mathbf{x}'$ and $\mathbf{y}$ is jointly typical is $2^{-I(\mathbf{X};\mathbf{Y})}$.

Thus the second type of error goes to zero as long as $R_J + T_2 < \frac{1}{n} I(\mathbf{X}, \mathbf{Y})$ and we obtain Theorem 2 which matches the converse result.

3.2.7 Achievable Region of the Suboptimal Approach with Lower Decoder Complexity

The probability of error comes from

- For the correct individual j, the decoder can not find a $\hat{\mathbf{z}}$ which is a typical sequence of $P_{\mathbf{Z}}^n$ and satisfies $G\hat{\mathbf{z}} = G\mathbf{y} - \mathbf{t}(\hat{j}) - \mathbf{s}(\hat{j})$.
- There is another combination of individual and secret such that there exists a $\hat{\mathbf{z}}$ which is a typical sequence of $P_{\mathbf{Z}}^n$ and satisfies $G\hat{\mathbf{z}} = G\mathbf{y} - \mathbf{t}(\hat{l}) - \hat{\mathbf{s}}$.

The first type of error happens when the noise $\mathbf{z}$ is not typical or $G\mathbf{z}$ is not decodable when z is typical, and both cases have negligible probability as the decoder is assumed to be good for $p_{\mathbf{Z}}^n$. The second condition happens if there exists a false pair of identity and secret so that $G\mathbf{y} - \mathbf{t}(l) - \mathbf{s}$ is the syndrome of a typical sequence of $p_{\mathbf{Z}}^n$. There are $nH(\mathbf{Z})$ out of $nH(\mathbf{X})$ syndromes of G that contain one and only one typical sequence of $p_{\mathbf{Z}}^n$ when the source code is good for $p_{\mathbf{Z}}^n$. Since $\mathbf{h}(j)$ and $\mathbf{s}_k$ are i.i.d. uniform over $\mathcal{X}$, the probability of each pair of $(l, \mathbf{s})$ resulting in a syndrome of a typical sequence of $p_{\mathbf{Z}}^n$ has probability $2^{n(H(\mathbf{Z})-H(\mathbf{X}))}$. The union bound of the second type of error e_2 is

$$\begin{aligned} P(e_2) &\leq \sum_{(l,\mathbf{s})\neq(j,\mathbf{s}(j))} 2^{-n(H(\mathbf{X})-H(\mathbf{Z}))} \\ &\overset{(a)}{=} (M_S M_J - 1) 2^{-n(H(\mathbf{X})-H(\mathbf{Z}))} \\ &\leq 2^{-n((H(\mathbf{X})-H(\mathbf{Z})-R_J-T_2)}, \end{aligned}$$

which goes to zero when $R_J + T_2 < H(\mathbf{X}) - H(\mathbf{Z}|\mathbf{X})$. This proves Theorem 3.

3.2.8 Conclusions

We proposed two system designs using linear codes for identification and secret binding trade-off. One of the designs is proved to achieve the optimal trade-off for general source and noise distributions. This design may require higher decoder complexity, while some practical applications may require low computational complexity. Thus we also proposed a second system design which uses decoders with lower complexity and prove its performance for general sources with additive noise. Also we generalize previous achievable rate region results to general source and noise distributions. A future direction is to design systems which consider information about the templates leaked from helper data, and its trade-off with identification and secret capacity.

Chapter 4

Robust Informative Feature Selection for LDV and ECG Biometrics

Laser Doppler Vibrometry (LDV) measures vibrations on the surface using the Doppler shift. LDV signals are recorded in a non-contact fashion and the unobtrusiveness is a major benefit of this technique as a biometric. LDV is targeted at the skin above the carotid artery due to arterial wall movements associated with heartbeat. In contrast to the electrocardiogram (ECG), which measures electrical activity of the heart through electrodes directly attached to the skin, the LDV signal is derived from mechanical movements. Both signals have been proposed for biometric applications [12, 13, 14, 9, 37, 46].

The LDV and ECG pulse signals provide information of the coarse aspects of the cardiac signal, including heart rate and heart rate variability, as well as more fine-grained and advanced features reflecting extremely detailed aspects of cardiovascular system. The LDV and ECG signals are nearly impossible to mimic. In addition, liveness and stress information provided by both signals are also useful against forgery.

Both the LDV and ECG signals of the same individual change from one occasion to the next, affected by factors such as physical exercise, mental stress, and perhaps other unobservable states. Indeed, several identification protocols based on the LDV signal produce a low equal error rate (EER) in

the range of 0.5% to 3% if training and testing data are recorded consecutively on the same day, but performance degrades to 19% or worse if testing data is collected one week to six months after the training session [13]. One reason for this performance degradation is that the physiological properties of the individual change gradually, and that these changes become more appreciable as time between the training and testing sessions become large. Similar performance degradation has also been observed in ECG biometrics, as discussed later in section 4.6. One way to overcome this issue is to use training data collected from multiple sessions such that gradual change can be modeled; however, this approach may not always be practical in actual use as the cost for repeated controlled measures is high and is intrusive to the users' activities. Methods that require only a few sessions to achieve proper performance are needed.

We propose a new robust feature selection method that takes into account effects of changes in statistics from training data to testing data for LDV and ECG biometrics. The idea is to jointly consider how well a feature distinguishes an individual from others and how stable this feature is. The idea of robust feature selection has evolved from an information-theoretic approach for dimensionality reduction [71] and its applications [13], as well as literatures on feature selection and sample size effects on recognition system performance [63, 31]. Extended from classical works, the proposed approach attempts to encompass scenarios where probability densities of some features vary from session to session gradually. When two training sessions are available, this method reduces EER to single digits with testing data collected in a period as short as 4 s. When 12 s are available for collecting testing data, EER can be further reduced to 7.4%. The EER is reduced by at most 21% relative to no selection, and 57% relative to the best approach based on two training sessions described in [12]. Further optimization on the feature extraction and selection methods leads to a cross session performance level of 6.5% for LDV and ECG biometrics, which are the leading results as of January 2012.

The rest of this chapter is organized as follows. Section 4.1 introduces basic procedures for LDV signal and ECG signal acquisition and preprocessing, training and testing data set, and experimental conditions. Section 4.2 provides an example of using a single training session. Section 4.3 describes the concepts of robust feature selection, and the implementation of a robust feature selection algorithm for the LDV based biometric system. Section 4.4 summarizes simulation results of the LDV

signal. The robust feature selection method for LDV biometric is concluded with discussion in 4.5. Section 4.6 presents a comparative study of methods in ECG biometrics.

4.1 Laser Doppler Vibrometry Signal Acquisition and Preprocessing

LDV data were obtained from 191 individuals who were asked to sit quietly for 5 min on three separate occasions. A Polytec PSV400 Laser Doppler Vibrometer, positioned at a distance of 91 cm from the recording location, was targeted at a site overlying the carotid artery, at a level approximately 1 to 2 cm below the right carotid sinus. Photographic records were obtained to ensure cross-session targeting constancy. For this study, target locations were marked with a small patch of retroreflective tape, and we have since confirmed that comparable data can be obtained from untreated skin. In the later comparison study, ECG data was obtained from 285 individuals, including the previous 191 individuals.

Recorded signals are sampled at 10 kHz and digitized with a Biopac MP 150 recording system. Signals are processed to suppress speckle dropout artifacts, and segments of signals related to individual heartbeats are extracted and downsampled to 1 kHz. Extracted carotid pulse signals are set to 700 ms in duration with the major velocity peak of each carotid pulse signal aligned at 200 ms. Figure 1 shows LDV pulse signal signals from two different individuals from two sessions. LDV pulse signals from the same individual have the same color, with solid line and dash line indicating different sessions. Wave form differences between individuals can be seen as well as differences between pulse signals from the same individual but different sessions. The pulse signals are normalized so that each has zero mean and unit energy.

The training data set consists of LDV signals from 191 individuals from two recording sessions, separated by at least one week up to a month. In each session, there are 150 LDV pulse signals, each corresponding to a single heartbeat, for each individual. A length 1102 feature vector is obtained from each LDV pulse signal by a prolate spheroidal based time-frequency decomposition method described later in this section. Hence, for each of the 191 individuals, the training data set consists

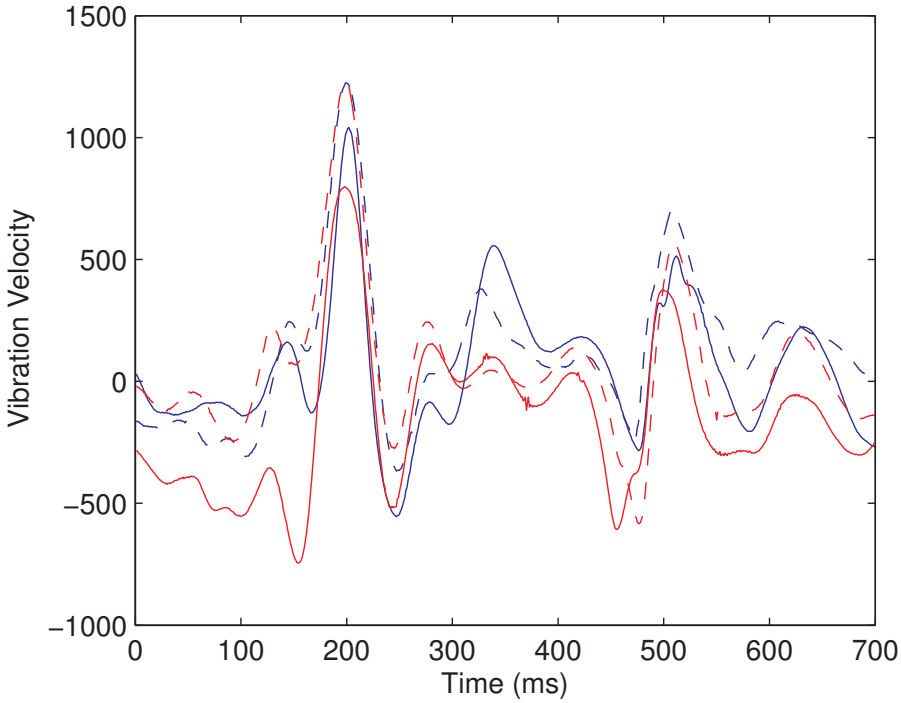

Figure 4.1: LDV carotid pulse signals from two individuals from two sessions.

of 150 length 1102 vectors corresponding to 150 LDV pulse signals from each of the two sessions. The testing data set consists of LDV signals from the same 191 individuals from a third recording session, separated by one week to six months from either of the two training sessions. We use up to 150 LDV pulse signals for each individual. Feature vectors are by the same method used in the training data.

Each LDV pulse signal abbreviated to 688 ms is parsed into 38 short time fragments, 96 ms each with 80 ms of overlap between consecutive short time fragments. We downsampled the LDV signals to 1000 Hz so that each small time fragment is a length 96 vector. The short time fragments are then projected onto a 29 dimensional space whose bases are discrete time prolate spheroidal functions with complex coefficients. In our simulations, discrete time prolate spheroidal basis functions are singular vectors associated with the largest 29 singular values of a matrix W whose (u, v) element is

$$W_{u,v} = e^{-j\pi \frac{(u-1)(v-1)}{L}}, \quad u \in 1, \cdots, U; v \in 1, \cdots, V, \tag{4.1}$$

where $U = 96$ is the length of each small time fragment, $V = 29$ is total number of prolate spheroidal basis functions, and L selected to be $2U$. The log of magnitudes of coefficients are taken such that each LDV pulse signal is represented as a length 1102 real number feature vector.

In practice, there may be constraints on data acquisition such that only a few heartbeats are available from a given individual as training data, or from an actual identity verification opportunity. Biometric performance under time constraints are studied by using subsets of the training and testing data sets. We simulated training data constraints of 6, 12, 37, and 150 consecutive heartbeats, and testing data constraints of 1, 4, 16, and 150 consecutive heartbeats. Overlapping sequences of heartbeats are used for both training and testing. Assuming an average heart rate of 75 beats per minute, this corresponds to 5, 10, 30, and 120 s for training and 1, 4, 12, and 120 s for testing. Actual acquisition times depended on each individual's heart rate.

4.2 LDV Biometrics Based on Single Training Session

We here illustrate the challenge of cross-session authentication in LDV and ECG biometrics and the key observations. A scenario of training on one session and testing on another is presented for the LDV biometrics. Authentication performance of a normal model is used [14]. This method yields the best performance up to December 2007, prior to the robust feature selection method was developed [12]. In this model, each feature (bin) of a individual is assumed to have the same variance as a nominal model. The empirical mean for each feature is obtained using maximum likelihood estimation. For the ith individual, the mean of the kth feature is calculated as $m_{k,i} = \frac{1}{150}\sum_{n=1}^{150} x_{n,k,i}$, where $x_{n,k,i}$ denotes the k-th feature of the n-th training LDV pulse of the i-th individual. The nominal, population, mean and variance are calculated for the kth feature as $m_{k,0} = \frac{1}{191}\sum_{i=1}^{191} m_{k,i}, \sigma_{k,0}^2 = \frac{1}{191}\sum_{i=1}^{191}(m_{k,i} - m_{k,0})^2$. The decision-making is based on the plug-in hypothesis test and the null-hypothesis is chosen to be the nominal mean feature vector. During testing, the testing pulse signal is decomposed into a feature vector $\bar{x}$ with the same method as training. The log-likelihood ratio becomes normalized mean square error, which is calculated as

$$S_i = \sum_{k=1}^{1102} \left[-(\bar{x}_k - m_{k,i})^2 + (\bar{x}_k - m_{k,0})^2 \right]. \tag{4.2}$$

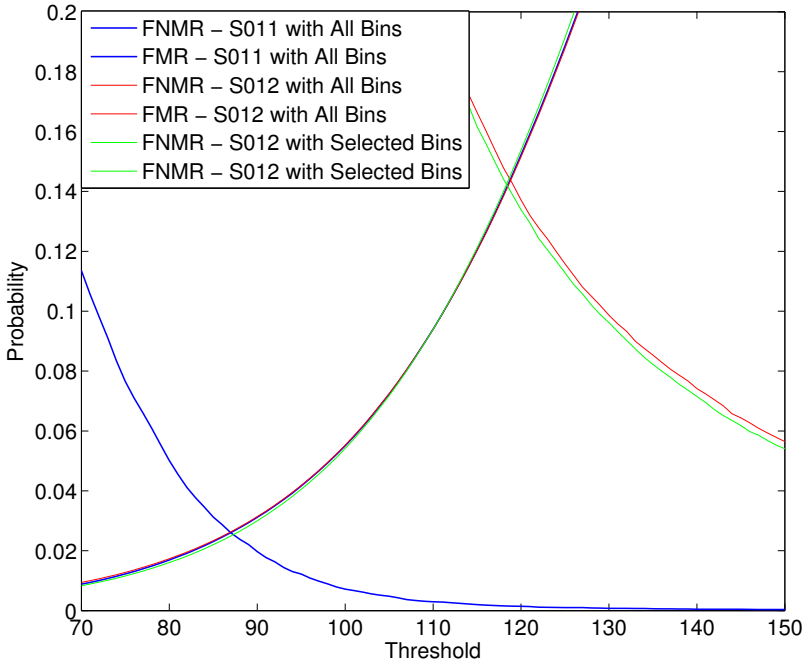

Figure 4.2: FMR and FNMR of training on session 1 and testing on sessions 1, 2 with normal model using a single heartbeat.

The thresholding dimensionality reduction method can increase recognition performance by selecting features which have distances to the nominal model greater than a threshold $\kappa > 0$ [71]. The distance is calculated with a thresholding function $d(\cdot, \cdot)$. The nominal model is used as the null hypothesis, which represents the population distribution. The selection criterion is to keep a feature for a particular individual only if the relative entropy, i.e. the Kullback-Leibler (KL) divergence, between the individual model and the nominal model is greater than a threshold. For Gaussian distributions with equal variances, the thresholding function becomes $d(p^0, p^i) = D(p^0||p^i) = (m_{k,0} - m_{k,i})^2/\sigma_{k,0}^2$, for the kth feature. Using the selected features, a distinct model for each individual is created. The modified log-likelihood ratio is calculated as

$$S_i = \sum_{k=1}^{1102} \left[-(\bar{x}_k - m_{k,i})^2 + (\bar{x}_k - m_{k,0})^2 \right] I_{d(p_k^0, p_k^i) > \kappa}. \tag{4.3}$$

Figure 4.2 shows the performance of the log-normal model using a single testing heartbeat training on single session. The EER is 2.6% for the within session test, 14.4% for inter-session testing with all bins selected. Results indicate that for a threshold that leads to the selection of 71% of the bins

on average, the EER for inter-session testing is reduced to 14.1%, and 12.5% using 4 heartbeats. This represents a marginal improvement over the case when all the bins are used. Note that the false non-match rate (FNMR) is very consistant across experimental settings, and what drives the overall performance down is the drastic increase in false reject rate (FRR). This indicates that the instability, or varibility, of the statistics of the LDV features of the same individual across different sessions is the primary source of error and a key challenge in developing the LDV biometrics. Similar perfromance degredation is also observed in ECG biometrics in cross session studies.

4.3 Robust Feature Selection against State Uncertainty

The results in Figure 4.2 can be improved if an approach is developed to account better for the instability of the statistics of the features within an individual across sessions. In this section, we present the concept of robustness in feature selection for LDV signal. A computational method for quantifying robustness and realizing this robust feature selection concept for LDV and ECG as biometrics is described. We evaluate this approach under several training and testing time constraints in terms of number of heartbeats available on LDV biometrics.

4.3.1 Motivation and Concepts

Assuming that we have data from two sessions with a given individual, the goal of robust feature selection is to choose features that provide more information to distinguish an individual from others against the uncertainty caused by variation of the features. Note that this does not mean that the features selected need to have both high distinguishability and stability. Consider six hypothetical examples of a two dimensional feature of data collected from two individuals in two sessions shown in Figure 4.3 to illustrate varying degrees of distinguishability and stability. The data points in the same color are from the same individual.

(a) In this best case scenario, the feature provides both high distinguishability and stability. The data points from two sources separated apart and form a single cluster for each source. Features of this type are very rare in LDV and ECG biometrics.

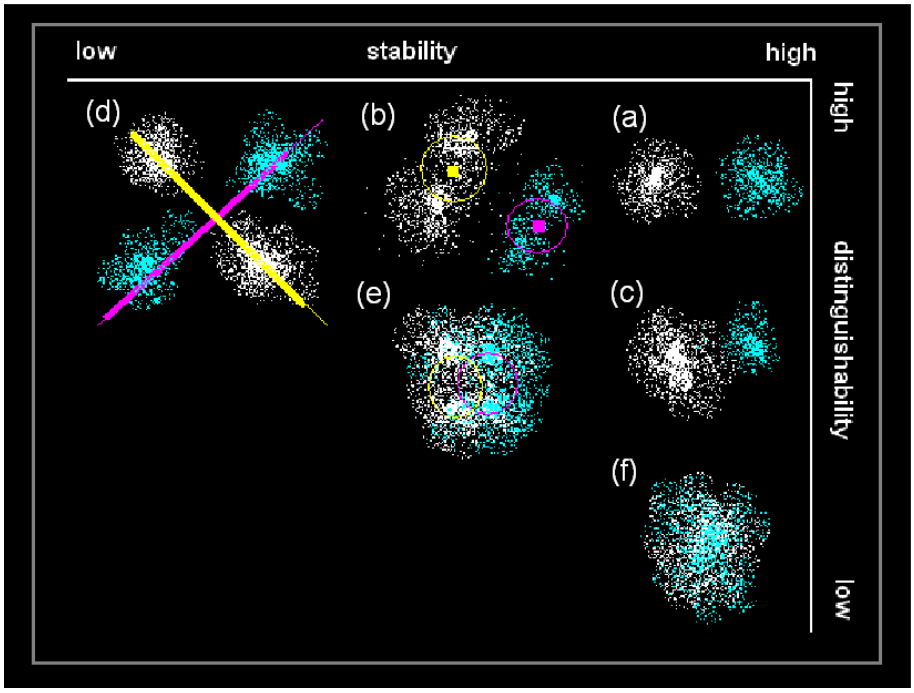

Figure 4.3: Illustration of features with different degrees of distinguishability and stability.

(b) This feature provides high distinguishability, but has lower stability. The data points from different sources are still separated widely, but for a given individual, the data from different sessions appear to form two clusters. However, by considering the empirical centers of two data sources, marked as yellow and purple dots in the figure, and drawing two circles at the centers that enclose the two means of data of different sessions, if the mean of each source in a testing session lies within the circle, two sources can still be distinguished by using this feature. This is a situation in which distinguishability is "stronger than" instability. By adjusting the radii of the circles, one can have a more restricted or loosened selection criterion.

(c) This feature is stable, but provides slightly less distinguishability. A quantitative selection rule can be developed by choosing a distinguishability measure and adjusting a threshold.

(d) The data points form four clusters, each corresponding to a source at a session, in a feature with high empirical distinguishability but notably low stability. This is an important example since if possible variations of states that govern data statistics are ignored, an overtrained classifier can be found by finding a decision boundary that separates data from different sources. However, physiological states are usually continuous and observed data from different states of

a single source should look connected. In this case, consider the purple and yellow lines that join two means of data from the same source but different sessions. Data from the two sources in a third session may have means anywhere alone these lines, resulting in highly overlapped data clusters. As argued in (b), the distinguishability of this feature is diminished by its acute instability.

(e) The distinguishability is already weak in this case, so that when taking its low stability into account, the usefulness of this feature is diminished.

(f) This is a clear case of a useless feature that while stable, exhibits no distinguishability.

In summary, when selecting features with slowly varying states that remain nearly constant within a session but change across sessions, neither distinguishability nor stability alone can be used to determine the biometric utility of a feature, as illustrated in cases (e) and (f) respectively. However, as illustrated in cases (b) and (c), distinguishability and stability must be considered jointly.

4.3.2 Computational Aspects of Robust Feature Selection

The algorithm consists of three steps: (1) learning how data are distributed across sessions and individuals, (2) quantifying distinguishability and instability of each feature for each individual, and then (3) selecting robust features. For identity verification, the binary hypothesis testing problem is to decide if the incoming LDV pulse signal is from the claimed identity j, or from others, i.e. the general population. Thus the goal of robust feature selection is to select features that can consistently provide information to distinguish a given individual from the population.

Given an individual, the steps can be computationally realized as shown in Figure 4.4 and the following.

(1) For the i-th feature, probability densities for each session denoted as $f_{1,i}$ and $f_{2,i}$ are learnt. Also each feature at each session has a nominal model representing the population density, denoted as $f_{1,p}$ and $f_{2,p}$. Moreover, the densities of fusing data from two sessions is denoted as $f_{12,i}$ and $f_{12,p}$ as the two centers in Figure 4.3.2. The fusion densities can be learnt from averaging the resulting density, or direct estimation from two session data. The probability

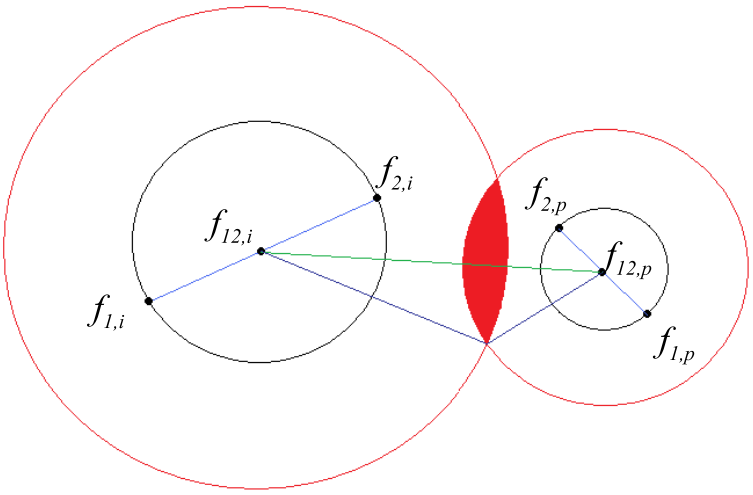

Figure 4.4: Accessing distinguishability and instablility for robustness. The green line represents the distinguishability of the feature between the individual model $f_{12,i}$ and the population model $f_{12,p}$. The red balls represents the instability of the features, whose radii are proportional to the difference between densities of two sessions $f_{1,i}$ and $f_{2,i}$. The red overlap region is inversely propotional to the robustness of the feature.

densities can be assumed from a parametric family of distributions and can be learnt from parameter estimation. Also nonparametric density estimates can also be used that are more robust to model assumptions [76].

(2) The distinguishability of a feature can be measured by how different the fusion densities $f_{12,i}$ and $f_{12,p}$ are. The difference measure $d(\cdot,\cdot)$ can be a metric such as the L_1 norm, or a distortion measure where symmetry is not required, such as the relative entropy, i.e. Kullback-Leibler (KL) divergence. The instablility, on the other hand, can be measured by the difference of two densities between the same individual across sessions [23]. We denote distinguishability and instablility as $D(i)$ and $I(i)$ respectively.

(3) Distinguishability and instablility can then be used to determine the robustness of a feature. For example, if both are quantified by the same difference measure, a simple rule as $D(i)-cI(i)$ can be used, where c is a constant. This is shown in Figure 4.3.2 that the blue radii represent the instability of the feature weighted by constant c, the green line is the distinguishability, and the red circles are the uncertainty balls whose overlap region is inversely proportional to the robustness of this feature.

4.3.3 A Robust Feature Selection Algorithm

The features we obtained represent 1102 coefficients resulting from the prolate spheroidal based time-frequency decomposition. The data distribution is estimated through nonparametric kernel density estimation. N denotes the number of the LDV pulse signals available in each training session under various time constraints. Also, $x_{s,n,k,i}$ denotes the value of the kth time frequency component of the nth LDV pulse signal from individual i from session s. For each feature of each individual, we obtain an empirical probability density estimate for each of the two sessions by using Gaussian kernel density estimation[76] with N training values. The densities are denoted as $f_{1,k,i}(x)$ for session 1, and $f_{2,k,i}(x)$ for session 2, where k is the index of the time frequency components ranging from 1 to 1102, and i is the index of the individual ranging from 1 to 191, such that

$$f_{s,k,i}(x) = N^{-1} \sum_{n=1}^{N} \frac{e^{-\frac{(x-x_{s,n,k,i})^2}{\sigma_{s,k,i}^2}}}{\sqrt{2\pi}\sigma_{s,k,i}}, \quad s \in \{1,2\} \tag{4.4}$$

$$\sigma_{s,k,i} = 0.9\hat{\sigma}_{s,k,i} N^{-\frac{1}{5}}, \tag{4.5}$$

where $\hat{\sigma}$ is the standard deviation of training data $x_{s,n,k,i}, n \in \{i, \cdots, N\}$. The mixture density, $f_{m,k,i}$, is computed as

$$f_{m,k,i}(x) = (2N)^{-1} \sum_{s=1,2} \sum_{n=1}^{N} \frac{e^{-\frac{(x-x_{s,n,k,i})^2}{\sigma_{m,k,i}^2}}}{\sqrt{2\pi}\sigma_{m,k,i}}, \tag{4.6}$$

$$\sigma_{m,k,i} = 0.9\hat{\sigma}_{m,k,i}(2N)^{-\frac{1}{5}}, \tag{4.7}$$

where $\hat{\sigma}$ is the standard deviation of training data $x_{s,n,k,i}, s = 1,2; n \in \{i, \cdots, N\}$, and m indicates that it is a mixture density of the two sessions. Also, for each feature, we use $191k$ LDV pulse signals to obtain a population density of the feature for each session. The densities are denoted as $f_{1,k,p}$ for session 1, and $f_{2,k,p}$ for session 2, where p indicates population densities. Also, overall mixture densities, $f_{m,k,p}$, are computed as

$$f_{s,k,p}(x) = (191N)^{-1} \sum_{i=1}^{191} \sum_{n=1}^{N} \frac{e^{-\frac{(x-x_{s,n,k,i})^2}{\sigma_{s,k,p}^2}}}{\sqrt{2\pi}\sigma_{s,k,p}}, \quad s \in \{1,2\} \tag{4.8}$$

$$f_{m,k,p}(x) = (382N)^{-1} \sum_{s=1,2} \sum_{i=1}^{191} \sum_{n=1}^{N} \frac{e^{-\frac{(x-x_{s,n,k,i})^2}{\sigma^2_{m,k,p}}}}{\sqrt{2\pi}\sigma_{m,k,p}}, \tag{4.9}$$

$$\sigma_{s,k,p} = 0.9\hat{\sigma}_{s,k,p}(191N)^{-\frac{1}{5}}, \tag{4.10}$$

$$\sigma_{m,k,p} = 0.9\hat{\sigma}_{m,k,p}(382N)^{-\frac{1}{5}}, \tag{4.11}$$

where $\hat{\sigma}_{s,k,p}$ is the standard deviation of kth value of all LDV pulse signals from all individuals from the s session, and $\hat{\sigma}_{m,k,p}$ is the standard deviation of the kth time frequency component of all LDV pulse signals from all individuals from both sessions. Hence, each individual has a total of 3306 densities for 1102 features from two sessions and one mixture, and so does the population.

The stability and distinguishability of a feature of an individual can be quantified by comparing the densities of the individual and densities of the population across different sessions. As discussed in part 4.3.1, the instability of a feature reflects differences in the data from the same individual or the population across multiple sessions. Hence, the instability can be quantified by the L_1 distance between $f_{1,k,i}$ and $f_{2,k,i}$

$$I(k,i) = \sum_{s=1,2} \int_{-\infty}^{\infty} |f_{m,k,i}(x) - f_{s,k,i}(x)|\, dx, \tag{4.12}$$

and the L_1 distance between $f_{1,k,p}$ and $f_{2,k,p}$

$$I(k,p) = \sum_{s=1,2} \int_{-\infty}^{\infty} |f_{m,k,p}(x) - f_{s,k,p}(x)|\, dx. \tag{4.13}$$

Here, $I(k,i)$ denotes the instability of feature k of individual i, and $I(k,p)$ denotes the instability of the densities of feature k of the population. Similarly, the distinguishability can be quantified as

$$d(k,i) = \int_{-\infty}^{\infty} |f_{m,k,i}(x) - f_{m,k,p}(x)|\, dx, \tag{4.14}$$

which measures the distance between the densities of the kth time frequency component of two sources, the individual i and the population.

For each individual i and each feature k, we compute objective scores

$$o(k,i,t) = d(k,i) - t(I(k,i) + I(k,p)), \tag{4.15}$$

where t is an adjustable parameter from 0 to 2. Then $o(k,i,s)$ is compared with zero and if $o(k,i,s) > 0$, feature k is selected for individual i. Thus, higher values of parameter t require greater levels of distinguishability versus instability for a given feature to be included into the biometric discrimination model. The set of selected feature indices for individual i is denoted as $\mathcal{F}_i$.

4.3.4 Identity Verification

In the testing phase, extracted LDV pulse signals are decomposed though the same prolate spheroidal time frequency decomposition used in the training phase, denoted as $\bar{x}$. For the claimed identity $\bar{i} \in \{1, \cdots, 191\}$, we extract the feature indexes selected for individual $\bar{i}$ and their associated mixture densities, and also the population mixture densities of those features. For each testing LDV pulse signal at each selected feature, optimal binary hypothesis testing is performed using trained mixture densities; the score of a selected feature is 0 if the feature rejects claimed identity, and 1 if the feature accepts the claimed identity:

$$score(\bar{x}, \bar{i}, k) = \mu\left(f_{m,k,\bar{i}}(\bar{x}_k) - f_{m,k,p}(\bar{x}_k)\right), k \in \mathcal{F}_{\bar{i}}, \tag{4.16}$$

where μ is the unit step function. The score of a testing LDV pulse signal is computed through normalized voting so that it is between 0 and 1:

$$score(\bar{x}, \bar{i}) = |\mathcal{F}_{\bar{i}}|^{-1} \sum_{k \in \mathcal{F}_{\bar{i}}} score(\bar{x}, \bar{i}, k). \tag{4.17}$$

Then the score is compared to a threshold for the final decision if it is tested based on a single heartbeat, i.e. under 1 s testing data acquisition time constraint. For 4 and 16 heartbeats based verification, the simple sum rule is used to fuse consecutive individual LDV pulse signal scores, and then compare to a threshold for decision.

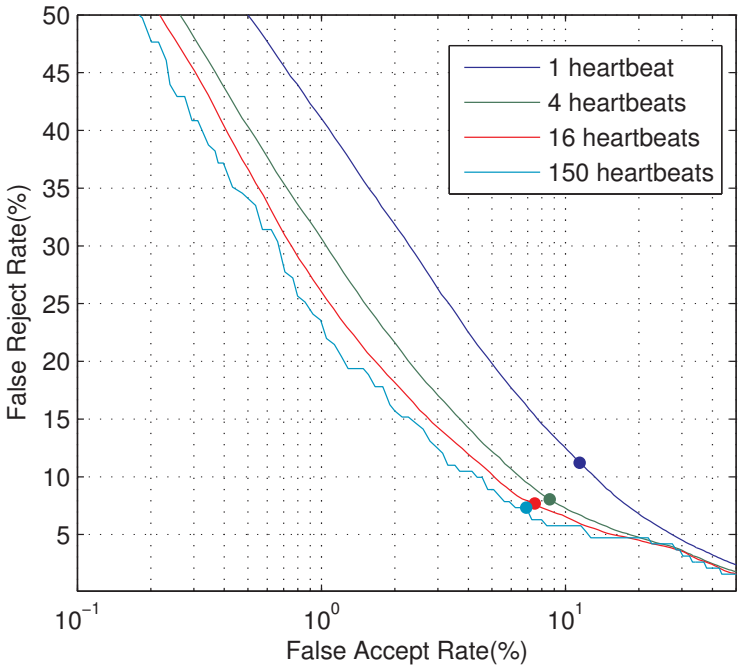

Figure 4.5: ROC curves for training on 37 heartbeats and testing on 1, 4, 16, and 150 heartbeats with feature selection. Dots mark the EER: 11%, 8.3%, 7.6%, and 7.0% for 1, 4, 16 and 150 heartbeats.

4.4 LDV Biometrics Results

Identity verification performance in terms of equal error rate (EER) is summarized in Table 1, under 16 training and testing constraints with or without the robust feature selection described in Section 4.3. The relative EER reduction by feature selection is shown in the last four columns of Table 1. The top row indicates training time allowed for collecting LDV signals, with units in s (and the associated number of heartbeats available). The first column indicates the testing time allowed and associated number of heartbeats available. Figure 3 shows the receiver operating characteristic (ROC) curves based on training on 37 heartbeats and testing on 1, 4, 16, and 150 heartbeats.

The "with feature selection" columns of Table 1 show the 16 EERs that are the best results of different feature selection thresholds t, simulated as described in (4.15). The "without feature selection" columns are EERs obtained without feature selection. It is clear that with the robust feature selection, as long as 4 or more heartbeats are available, the EER is under 10% (even for a training condition which allows for only 5 s to extract 6 heartbeats for each individual). The last four columns indicate the EER reduction gained by feature selection relative to systems without

feature selection under different training and testing conditions. For example, when training with 30 s and testing with 12 s of LDV data, EER is reduced from 8.8% (without feature selection) to 7.6% (with feature selection), such that a $\frac{8.6-7.6}{8.8} = 11\%$ relative reduction of EER is obtained. The relative reduction from training on two sessions with robust feature selection over training on a single session described in section 4.2 is 37%.

4.5 Conclusion of LDV Biometrics

We are able to obtain EER performance of less than 10% under various time constrained scenarios using a robust LDV method which emphasizes the analysis of variations of feature statistics among training and testing sessions. Robustness is achieved through jointly considering distinguishability and stability of features across sessions. The results support the utility of LDV measurements as a novel source of biometric information. Further optimization on the feature generation, i.e. the decomposition, methods and quantitative measures of robustness lead to a cross session performance level of 6.5% by Ikenna Odinaka.

Theoretical foundations for the proposed robust feature selection coupled with proposed feature fusion need to be further studied, as do other distance measures and feature fusion methods. The chosen L_1 distance measure between probability distributions is closely related to the Bayes risk of binary hypothesis test and the Kolmogorov variational distance [23]. For any two probability densities $h(x)$ and $g(x)$ over support $\mathcal{X}$, we have

$$\begin{aligned} & \int_{x\in\mathcal{X}} |g(x) - h(x)|dx \\ = & \int_{x\in\mathcal{X}, g>h} g(x) - h(x)dx + \int_{x\in\mathcal{X}, h>g} h(x) - g(x)dx \\ = & \; 2\left(1 - \int_{x\in\mathcal{X}} \min\left(g(x), h(x)\right) dx\right), \end{aligned} \tag{4.18}$$

where the integral of last equality of (4.18) is equal to the probability of error. Thus, voting through features selected by using the L_1 distance to measure distinguishability with a threshold q means that with k test data sets, the distributions of vote counts of data drawn from $h(x)$ and $g(x)$ should be distinguished at least as well as distinguishing a binomial with parameter q from a binomial

Table 4.1: EER (%) under 12 training and testing time constraints(s) and associated number of heartbeats (hb) available with or without robust feature selection, and relative EER reduction. Two significent digits are reported.

	With feature selection Training (s/hb)				Without feature selection Training (s/hb)				Relative EER reduction Training (s/hb)			
Testing (s/hb)	5/6	10/12	30/37	120/150	5/6	10/12	30/37	120/150	5/6	10/12	30/37	120/150
1/1	12	12	11	11	13	13	12	11	7.7	7.7	8.3	0.0
4/4	9.7	9.4	8.3	8.2	10	11	9.3	8.8	3.0	15	11	6.8
12/16	9.1	8.3	7.6	7.4	9.5	10	8.8	7.8	4.2	17	14	5.1
120/150	8.6	7.7	7.0	7.0	10	9.8	8.6	7.9	14	21	19	13

with parameter $1-q$, under the assumption of features being independent. Besides the congruence with the L_1 distance measure, another possible benefit of using voting for feature fusion lies in its robustness against outliers among selected features, such as few dominant likelihood ratios from some features toward the wrong decision. When relative entropy is used as a distortion measure between densities with the sum of loglikelihood ratios of features as the score, the EER performance degraded to above 10%.

4.6 Comparative Study of Methods in ECG Biometrics

As the robust feature method yield good results in LDV biometrics, two questions need to be addressed:

- How well is the performance of the robust feature selection method compare to other methods?
- How do the robust feature methods perform on related biometrics where instability due to state changes also plays an important role?

Both questions are studied by applying the proposed robust methods to electrocardiogram (ECG) based biometrics. The use of ECG as a biometric is an emerging field started from the studies by Biel et al. in 1999 [9], and Irvine et al. [37] and Kyoso and Uchiyama in 2001 [46]. Until May 2012, there have been over 100 publications on methods for ECG biometrics. This provides a great opportunity to compare the performance of different methods. Special thanks to our colleague Ikenna Odinaka who implemented a large number of methods in the literature, optimized the robust feature method specific for ECG, and carried out a comprehensive study. Note that methods based on fiducial features were not implemented and not included in the comparison due to following four reasons [38]:

- No consensus on standards for detection of characteristic points.
- Location of some characteristic points are disproportionately affected by the presence of noise, even using a fixed fiducial detector.

- Difficulty in defining the boundaries and peaks of atypical heartbeats usually leads to an increased failure to enroll.
- Problems with generalizability to larger databases, when the number of features are limited.

4.6.1 A Robust Feature Selection Method for ECG Biometrics

From each ECG pulse signal, we compute a spectrogram which is the logarithm of the square of the magnitude of the short-time Fourier transform of a normalized ECG heart pulse. In computing the short-time Fourier transform (STFT), we use a Hamming window of size 64ms, with a step size, which is defined as the distance between the beginnings of two consecutive windows of 10ms. Thus, there is an overlap of size 54ms between consecutive time frames. This window size was chosen empirically so that it yields robust and good single-heartbeat authentication performance in terms of equal error rate (EER). After computing the STFT, the frequency content was truncated at 250Hz to reduce boundary effects. The spectrogram is then computed as the logarithm of the squared-magnitude of the truncated STFT. We refer to the index of each point of the spectrogram as a time-frequency bin. Thus each ECG heart pulse can be represented by $L = 2048$ time-frequency components denoted as $Y(l)$. To build a generative classifier, we use independent normal distributions to model the time-frequency bins of each subject. During training, only the means and variances have to be estimated. For each bin l of subject i, we use the maximum likelihood (ML) estimates which are the sample means and variances denoted as $\hat{\theta}_i(l) = (\mu_{il}, \sigma^2_{il})$.

We use a robust informative feature selection method to select informative time-frequency bins for verification and recognition for ECG. This method is very similar to the one used for LDV biometrics. The two key elements considered in our feature selection method are distinguishability and stability. The feature should help distinguish the subject from a reasonably large subset of other subjects, and it should be stable across sessions. The l-th feature of the i-th subject is selected if the symmetric relative entropy, i.e. the symmetric Kullback-Leibler divergence, between $\mathcal{N}(\mu_{il}, \sigma^2_{il})$ and the nominal distribution $\mathcal{N}(\mu_{0l}, \sigma^2_{0l})$ is larger than a threshold $\kappa > 0$. The relative entropy between two densities p and q is defined by

$$D(p\|q) = \int p \log \frac{p}{q} \tag{4.19}$$

where the integral is taken over the support set of p. The symmetric relative entropy between the two densities is defined as

$$d(p,q) = D(p\|q) + D(q\|p) \tag{4.20}$$

For the Gaussian distributions used in our model, the symmetric relative entropy between $\mathcal{N}(\mu_{il}, \sigma_{il}^2)$ and $\mathcal{N}(\mu_{0l}, \sigma_{0l}^2)$ is

$$d(\hat{\theta}_i(l), \hat{\theta}_0(l)) = \frac{\sigma_{il}^2 + (\mu_{il} - \mu_{0l})^2}{2\sigma_{0l}^2} + \frac{\sigma_{0l}^2 + (\mu_{il} - \mu_{0l})^2}{2\sigma_{il}^2} - 1 \tag{4.21}$$

where the nominal model is obtained by using the spectrograms of all the subjects in the database. Using the symmetric relative entropy for feature selection ensures that only those bins whose distributions are far from the nominal are selected for each subject, thereby ensuring distinguishability. Moreover, stability of features is enforced by the variance of the subject's bin σ_{il}^2. It follows from the construction of the nominal model that for the most part, $\sigma_{il}^2 < \sigma_{0l}^2$, so that the second term in equation (4.21), with σ_{il}^2 in the denominator, increases as σ_{il}^2 decreases; for subject bins with small variances, the symmetric relative entropy tends to be large.

The score of a test heartbeat using the i-th subject's model is given by the log-likelihood ratio (LLR):

$$\Lambda_i = \sum_{l=1}^{L} \log \left[\frac{p_i(Y(l)|\hat{\theta}_i(l))}{p_0(Y(l)|\hat{\theta}_0(l))} \right] I_{d(\hat{\theta}_i(l), \hat{\theta}_0(l)) > \kappa} \tag{4.22}$$

where $I_{\{.\}}$ is the truth function indicating which time-frequency bins are selected; l is the index of the bins. For verification, the LLR given in equation (4.22) is compared with a threshold τ, so that if $\Lambda_i > \tau$, the heartbeat with the claimed identity is accepted, otherwise the heartbeat is rejected. For recognition, the LLR is computed for every subject model, and the subject whose model gives the largest score is declared. For rank-k recognition, subjects with models yielding the top k scores are declared. For across session verification, the score function was modified so as to disregard the role of the variances of the time-frequency bins. That is, we set $\hat{\theta}_i(l)$ and $\hat{\theta}_0(l)$ to a constant θ.

In recognition, to ensure that a variable number of time-frequency bins can be selected for each subject's model, the score obtained from comparing a test heartbeat to a subject's model is normalized by a score obtained from comparing the heartbeat to the nominal model. This normalization

ensures that there is no direct relationship between the number of bins used in a subject's model, and the value of the computed score [71]; more selected bins does not mean higher scores.

4.6.2 Comparative Results

The within-session analysis results are given in Table 4.2, which shows each algorithm, the authentication performance reported in the cited paper, if available, and its performance using our database. In the table, FS and NFS stand for "feature selection" and "no feature selection" respectively [59]; FS and NFS correspond to the cases where relative entropy based feature selection is or is not used, respectively. Moreover, "train 8, test 8" represents using 8 heartbeats, or 8 s, for the cases of Agrafioti *et al.* and Wang *et al.*. for training and the same number for testing. From the table, we can see that most algorithms do a decent job in modelling the ties within an individual and discriminating between individuals. However, for some algorithms there are noticeable differences between the authentication performance reported in the literature and what we obtained using our database. The original algorithm proposed by Molina *et al.* [57] uses a morphological baseline wander removal technique during preprocessing, which introduces distortions in the ECG recording; When bandpass filtering was used for preprocessing instead, the authentication performance improved. Also, the polynomial-based algorithm proposed by Sufi *et al.* [79, 78] suffers from performance deficiencies compared to what was reported in the literature. This is likely due to the large sample size we used for this study; only 15 individuals were used in the original study performed by the authors. When the first 15 individuals from our database were used for the biometric study, an equal error rate of 0.95% was obtained. The same phenomenon holds true for the algorithm proposed by Coutinho *et al.* [19, 18]. The original study performed by the authors used ECG data obtained from 26 individuals. When the first 26 individuals from our database were used for the biometric study, an equal error rate of 0% was obtained, in comparison to the much higher rates, in the range of 35% observed when applied to our full database of 265 individuals. The algorithm proposed by Yao and Wan [92] doesn't perform as well as some of the other methodologies. One possible reason for this is that only a single principal component was used for classification. The principal component approach may not be adequate to completely separate overlapping classes in the feature space. In general, when training and testing data come from the same session, most algorithms are good at accepting a true identity and rejecting a false one, as evidenced by their within-session authentication performance.

However, when training and testing are on different days, all the algorithms suffer a deterioration in performance, as is reflected in Table 4.3. In the table, "train 32, test 16" represents using 32 heartbeats (or 32 s) from session 1 for training and using 16 heartbeats (or 16 s) from session 3 for testing.

The results for across-session testing, when the training data are obtained from two different days is given in Table 4.4. In the table, "train (8+8), test 16" represents using 8 heartbeats (or 8 s) each from sessions 1 and 2 for training and using 16 heartbeats (or 16 s) from session 3 for testing. Cross-session training is vital to the improvement of biometric performance as it accomodates variability across different measurement times in the model.

Based on the across-session performance when fusion is used, we can see that a few of the methodologies provide the framework to capture the variability across time and provide for an improvement in authentication performance. By comparing the last columns in Table 4.3 and Table 4.4, where a total of 32 heartbeats (or 32 s) are used for training, and 16 heartbeats (or 16 s) are used for testing, we can see the effect of fusing data from more than one session during training, on the authentication performance; With the exception of the algorithm by Fang and Chan [26], all the algorithms show a varied degree of improvement in performance, which can be attributed to data fusion. The most remarkable improvement in performance can be seen in the algorithm by Odinaka *et al.* [59] and Wan and Yao [84], where data fusion accounts for about a 50% and 65% drop in EER, respectively.

Figures 4.6 and 4.7 show the detection error tradeoff curves for the top three (based on EERs) methodologies in the within-session and across-session (with fusion) analysis, respectively. In the figures, the red, blue, and green lines represent the detection error tradeoff curves for the first, second, and third methods (in terms of EER), respectively.

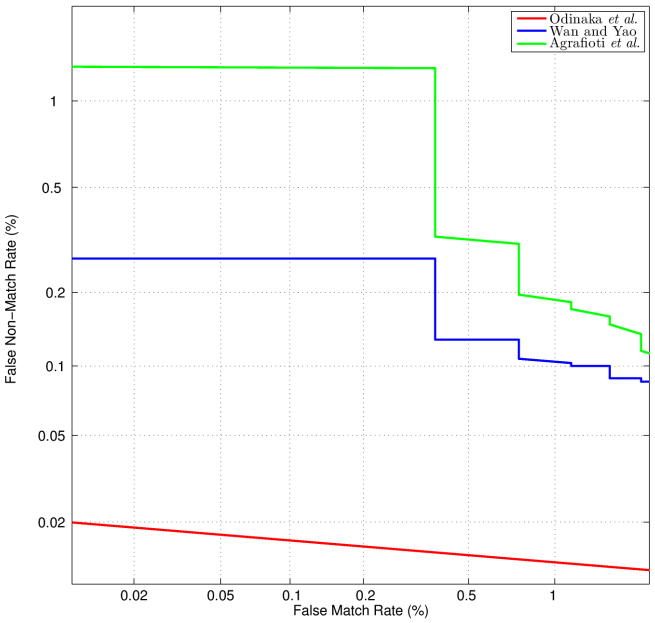

Figure 4.6: Detection error tradeoff (DET) curve for the top three methodologies in the within-session analysis

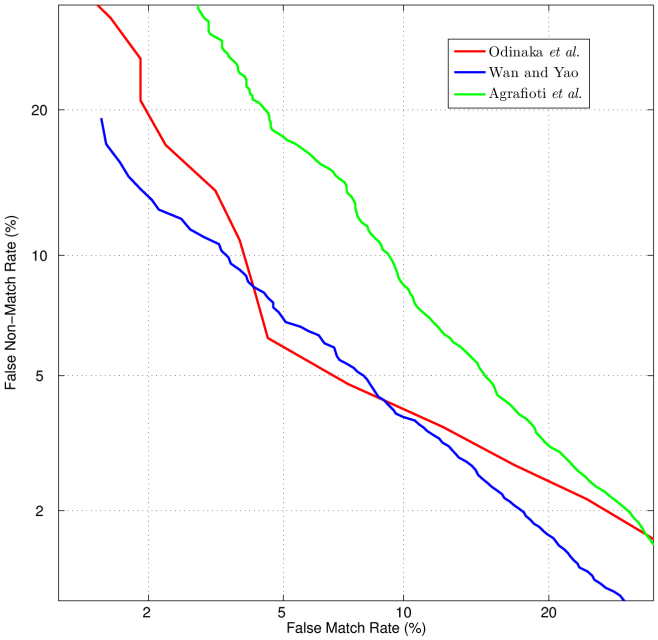

Figure 4.7: Detection error tradeoff (DET) curve for the top three methodologies in the across-session (with fusion) analysis

Table 4.2: Authentication performance for within-session analysis

Researchers	Equal Error Rates (%)				
	Literature	Train 8, Test 8	Train 16, Test 16	Train 32, Test 32	Train 64, Test 64
Agrafioti *et al.* [2]	0.6	3.88	0.85	0.57	0.38
Chan *et al.* [11]	-	5.82	3.84	3.02	2.26
Chiu *et al.* [15]	0.83 - 0.86	4.15	2.64	1.76	1.01
Coutinho *et al.* [18]	-	42.54	38.92	35.14	33.34
Fang and Chan (SC) [26]	-	19.81	19.1	19.03	18.82
Fatemian and Hatzinakos [27]	-	8.69	5.99	4.37	2.26
Irvine *et al.* [38]	-	2.25	1.74	1.26	0.69
Khalil and Sufi [44]	-	5.28	2.64	1.56	1.13
Li and Narayanan (HPE+SVM) [52]	0.55	2.19	1.24	1.17	0.96
Lourenço *et al.* [54]	13	12.01	9.3	6.56	5.25
Molina *et al.* [57]	2	19.99	16.31	16.27	15.98
Molina(M) [57]	-	13.71	7.53	6.16	5.57
Odinaka *et al.* (FS) [59]	0.02	1.89	0.93	0.38	0.03
Odinaka *et al.* (NFS)[59]	-	1.93	1.04	0.51	0.06
Sufi *et al.* [79, 78]	-	27.39	21.97	17.14	13.42
Wan and Yao *et al.* [84]	-	7.98	2.15	0.75	0.27
Wang *et al.* (DCT) [85]	-	3.9	2.22	1.74	1.36
Wübbeler *et al.* [90]	-	1.08	0.57	0.57	0.38
Yao and Wan [92]	-	24.46	22.32	20.63	18.49
Ye *et al.* [93]	-	5.11	2.84	1.64	1.13

Table 4.3: Authentication performance for across-session (without fusion) analysis

Researchers	Equal Error Rates (%)				
	Train 8, Test 8	Train 16, Test 16	Train 16, Test 32	Train 32, Test 16	Train 32, Test 32
Agrafioti *et al.* [2]	17.95	11.73	11.64	10.48	10.36
Chan *et al.* [11]	16.83	15.37	14.99	14.91	14.64
Chiu *et al.* [15]	26.56	26.28	26.20	26.38	26.36
Coutinho *et al.* [18]	47.14	45.77	44.67	44.59	43.93
Fang and Chan [26]	29.53	29.54	29.59	29.57	29.71
Fatemian and Hatzinakos [27]	20.40	20.30	20.18	19.42	19.31
Irvine *et al.* [38]	22.25	21.93	21.87	21.66	21.57
Khalil and Sufi [44]	24.16	22.89	22.13	21.94	21.13
Li and Narayanan (HPE+SVM) [52]	19.94	19.20	19.16	18.33	18.16
Lourenço *et al.* [54]	26.11	25.29	25.33	24.85	24.58
Molina *et al.* [57]	37.07	31.75	30.77	30.20	29.42
Molina(M) [57]	34.61	28.65	27.13	25.92	24.67
Odinaka *et al.* (FS) [59]	12.30	11.29	11.13	11.11	11.30
Odinaka *et al.* (NFS)[59]	21.46	20.66	20.37	19.80	20
Sufi *et al.* [79, 78]	35.34	33.17	31.98	32.23	31.64
Wan and Yao [84]	16.93	18.44	16.82	21.65	19.22
Wang *et al.* (DCT) [85]	17.94	17.69	17.63	17.72	17.61
Wübbeler *et al.* [90]	16	15.6	15.79	15.59	15.77
Yao and Wan [92]	33.33	31.79	31.38	30.92	30.13
Ye *et al.* [93]	22.98	19.19	20.17	20.17	18.55

Table 4.4: Authentication performance for across-session (with fusion) analysis

Researchers	Equal Error Rates (%)			
	Train(8+8), Test 8	Train(8+8), Test 16	Train(16+16), Test 16	Train(16+16), Test 32
Agrafioti *et al.* [2]	10.68	10.53	9.56	9.51
Chan *et al.* [11]	12.30	11.91	11.57	11.22
Chiu *et al.* [15]	21.34	21.18	21	20.97
Coutinho *et al.* [18]	46.17	45.64	44.41	43.77
Fang and Chan [26]	30.18	29.85	30.22	30.18
Fatemian and Hatzinakos [27]	17.13	16.66	16.92	16.35
Irvine *et al.* [38]	19.65	19.43	19.35	19.22
Khalil and Sufi [44]	18.91	18.58	18.87	18.53
Li and Narayanan (HPE+SVM) [52]	17.40	17.38	17.06	17.09
Lourenço *et al.* [54]	23.22	22.48	22.46	21.97
Molina *et al.* [57]	27.62	27.21	26.12	26.19
Molina(M) [57]	22.24	21.46	20.89	20.48
Odinaka *et al.* (FS) [59]	6.12	6.04	5.64	5.47
Odinaka *et al.* (NFS)[59]	16.08	15.85	14.91	14.73
Sufi *et al.* [79, 78]	33.49	31.03	31.35	29.95
Wan and Yao [84]	9.31	9.45	6.23	6.28
Wang *et al.* (DCT) [85]	16.16	15.92	15.85	15.93
Wübbeler *et al.* [90]	14.62	14.29	14.11	13.98
Yao and Wan [92]	30.99	30.69	30.15	29.84
Ye *et al.* [93]	16.74	17.06	14.32	13.67

Chapter 5

The Minimum Description Length Principle for Clustering and Computational Stemmatology

5.1 Introduction

Clustering is one of the fundamental problems in learning, and is important in many fields from artificial intelligence to bioinformatics. A fundamental problem in clustering is the existence of free parameters affecting the outcome. Many clustering algorithms require users to determine either explicitly the number of clusters to output, such as the Gaussian Mixture Model (GMM), K-means, and nonlinear manifold learning [77], or implicitly, such as hierarchical clustering methods [91], graph clustering by graphical cuts [82] and affinity propagation (AP) [29]. Despite allowing those algorithms to be flexible, it imposes difficulties in comparing and interpreting results. Can the number of clusters and other parameters be determined in a principled way? Can a clustering algorithm balance the number of parameters used and the modeling error? These are classic model selection questions addressed in information theory and other communities for a long time [66, 6, 74, 83]. Model selection theories have been applied to cluster multinomial data based on MDL arguments [45], GMM clustering [10] based on the universal prior of integers developed by Rissanen

[64], and K-means clustering [61] based on the Bayesian information criterion (BIC) developed by Schwarz [74].

In real life, one often encounters a very closely related situation where one needs to infer a structural relationship among data points based on an incomplete dataset. Stemmatology is a class of such problems. The goal of stemmatology is to reconstruct a family tree of different variants of a text resulting from imperfect copying, which is a crucial part of textual criticism. In reality, historians often have incomplete data because some variants are not yet discovered and there are missing portions in available variants due to physical damage. Stemmatology is similar to molecular phylogenetics where biologists aim to reconstruct the evolutionary history of species based on genetic or protein sequences. Adoption of phylogenetics methods has lead to encouraging results in automatic stemmatology.

In this chapter, we propose an information-theoretic framework of similarity-based clustering based on the idea of two, or multi, part codes and MDL, and its application to stemmatology. We utilize MDL concepts to the structural inference problem, particularly focusing on stemmatology where in addition to missing data points, the available data points have missing values. We offer new insights on how to handle these issues. Description length is measured information theoretically with the bit as the fundamental unit that the number of clusters can be determined automatically and balanced with the algorithm performance. We argue that similarity-based clustering problems can be turned into problems of combinatorial optimization on graphs and there is an information theoretic rationale of graph-based clustering methods. We develop a general algorithm based on MDL insights that is simple to implement and can be used along with other existing algorithms, and propose a generic MDL encoder with minimal assumptions made about the data, returning a hierarchical clustering of data. We discuss and demonstrate the potential application of the proposed MDL clustering concepts to stemmatology. Our method is applied to realistic datasets and outperforms major existing methods as of June 2010.

5.2 MDL Clustering Code Intuition

In this section, we introduce two intuitive MDL settings for clustering. The first is to partition the data into clusters only. The second is to determine clusters and for each cluster, to determine an exemplar. For both cases, the following notation is used. The number of data points is denoted as N. The data points are denoted as $x_i \in \mathcal{X} \subset R^d$, where $i = 1, \cdots, N$. Let $\underline{x}$ denote $(x_1, \cdots, x_N)$ and $\underline{x} \in \underline{\mathcal{X}} = \mathcal{X}^N$. The number of clusters inferred is denoted as K. For the case in which exemplars need to be identified, the exemplars are denoted as $x_k, k \in \mathcal{K}$,where $\mathcal{K} \subset \{1, \cdots, N\}$ is the index set of exemplars with cardinality K. For the case in which only clusters need to be determined, $k = 1, \cdots, K$ denotes a generic index of clusters.

When only clusters need to be determined, a code which describes a particular choice of clusters consists of three parts:

1. To specify K requires $\log N$ bits;
2. To specify a cluster k requires $\log K$ bits for each data point;
3. To describe X_l given that it is in cluster k requires $\log \frac{1}{P(x_l|k)}$ bits.

The total description length is then

$$\log N + N \log K + \sum_{k=1,\cdots,K} \left(\sum_{l \in \mathcal{L}(k)} \log \frac{1}{P(x_l|k)} \right), \tag{5.1}$$

where $\mathcal{L}(k)$ is the index set of data points in cluster k.

For the case where clusters and their exemplars need to be determined, a code which describes a particular choice of clusters and exemplars consists of four parts:

1. To specify K requires $\log N$ bits;
2. To specify indexes of exemplars requires $\log C_K^N$ bits;
3. To specify an exemplar $x_k, k \in \mathcal{K}$ requires $\log \frac{1}{P(x_k)}$ bits;

4. To specify the exemplars for all other data points requires less then $(N-K)\log K$ bits;

5. To describe x_l given that its exemplar is x_k requires $\log \frac{1}{P(x_l|x_k)}$ bits.

The total description length is then

$$\begin{aligned} \log N &+ \log C_K^N + (N-K)logK \\ &+ \sum_{k\in\mathcal{K}} \left(\log \frac{1}{P(x_k)} + \sum_{l\in\mathcal{L}(k)} \log \frac{1}{P(x_l|x_k)} \right), \end{aligned} \tag{5.2}$$

where $\mathcal{L}(k)$ is the index set of data points with exemplar x_k.

5.3 MDL for Similarity Based Clustering

It is readily seen that MDL provides a basis for clustering from choosing the number of clusters to assign cluster membership to each data points. In addition, MDL-based model selection allows one to easily incorporate prior information or constraints into models by translating them into densities or encoding strategies of model parameters. We consider the MDL setting for similarity-based clustering in this section. We assume that x_i are independent for all i. An other major category of clusterings is density based clustering whose brief overview is given in section 5.10.

For similarity-based clustering, one has to estimate the description length through similarities for selecting clustering models. There are three ways to encode a data point, either to encode it directly, to jointly encode it with some other data, or to encode it provided other encoded data. It is clear that there are different perspectives of how similarity-based clustering can be done under different restrictions of encoding operations, resulting in different two part codes and interpretation. For example, if one allows all joint encoding among any number of data points in the same cluster, it is well known that encoding all data jointly leads to the shortest code length. Hence in this framework, the shortest encoding is the grouping of all data in the same cluster and to jointly encode them, leading to no clustering. Thus one should restrict the encoding parameters so that a clustering interpretation exists. One obvious restriction is to allow joint encoding among at most $\nu < N$ data points. Since most similarity measures used in practice are defined on pairs of data, we shall consider

the case of encoding a data point using at most one other data point, along with model parameters. This leads to following two clustering frameworks, both of which can be turned into optimization on graphs.

1st order similarity-based clustering with weak exemplars: The goal is to minimize the code length of describing data $\underline{x}$ with model parameter vector $\underline{t}$ whose ith element is an integer from $\{1, \cdots, N\}$, the helper data index of the ith data point. The code length is

$$L(\underline{x}|\underline{t}) + L(\underline{t}) = \sum_{x_i : i \neq t_i} L(x_i|x_{s_i}) + \sum_{x_i : i = t_i} L(x_i) + L(\underline{t}). \tag{5.3}$$

If no prior is assumed for $\underline{t}$, the second term is at most $N \log N$. When $t_i = i$, x_i has to be encoded by itself and the code length is $L(x_i)$. When $t_i \neq i$, the code length is $L(x_i|x_{t_i})$. Letting $t_i^q = t_{t_i}^{q-1}$ and $t_i^1 = t_i$, we see that for all data points to be encodable, $t_i^N = t_i^{N+1}$ must hold. This means that if we take data points as N vertexes s on a graph and connect an edge between vertexes $(i, t_i), \forall i \neq t_i$, we have a forest. By adding an additional *base vertex* and connecting it to all vertexes with $i = t_i$, we then have a tree clustering. Thus finding the shortest code length for similarity-based clustering can be turned into a problem of finding an minimum path length arborescence tree (MAT) in a complete directed graph. The weights of directed edges are $L(x_i|x_{t_i})$, and vertexes are data points and the base node. The resulting directed tree has the base vertex as its root, and the number of children of the root is the number of clusters. Data points with a common ancestors up to the children of the root are clustered to be in the same cluster. The resulting tree can be viewed as a hierarchical clustering with exemplars at each branch. If the corresponding similarity measure used is symmetric, it is then the well known problem of finding the minimum spanning tree (MST). Tarjan's algorithm is proved to find the MAT in a directed graph with no negative cycle with complexity $O(N^2)$ [73].

1st order similarity-based clustering with strong exemplars: In this framework, t_i is allowed only if $t_i^2 = t_i$ except for the exemplar data which has $i = t_i$, the other data can only be coded with an exemplar data as helper data. As discussed earlier, this is then a problem of finding a two level arborescence tree in a complete directed graph.

5.4 Simulations on Syntheses Data

We present simulations using MDL-based MST clustering. The code lengths $L(x_i)$ and $L(x_i|x_j)$ are measured as the sum of the lengths of Rissanen's universal code for integers [64] of data vector elements, or differences of data elements, are quantized to a desired precision. Case 1 consists of four 2-dimensional Gaussian clusters with means forming a square with edge length equal to 30. The covariance matrices are all equal to the identity matrix. In Case 4, two Gaussian clusters with means 45 units apart are simulated. The covariance matrices are I and $5I$. In Case 3, data vectors are drawn uniformly from four rings of outer radius 1.1 and inner radius 0.9 with centers forming a square with edge length of 10. In Case 4, data vectors are drawn from two interlocking rings in 3 dimensional space shown in Figure 3. Both circles have radius 2, centered at the origin and $(2, 0, 0)$, and lie on the $x - y$ plane and the $y - z$ plane respectively. For all cases, each cluster has 100 data points.

The results are shown in Figures 1 to 4. The MDL-MST returns the correct number of clusters for these three cases. Case 3 shows that MDL-MST clustering can resolve clusters of a complicated data structure, while affinity propagation (AP) returns a larger number of clusters. Affinity propagation often returns more clusters than the actual number, even when parameters are set to the suggested values of [29].

5.5 Introduction to Computational Stemmatology

Before printing technology was widespread, text documents had to be copied by hand, mostly with errors. Thus, despite many documents originating from a common original text, they differ from one another. For those variants that survived and were discovered, historians are interested in knowing the relations among them, in particular, the family tree of the copying history. The research of finding such a family tree based on surviving variants is called stemmatology, and a proposed tree is called a stemma. A stemma is ideally a rooted tree where a child node is copied from its ancestor node in the tree. An accurate stemma with geographical, and temporal if available, information of

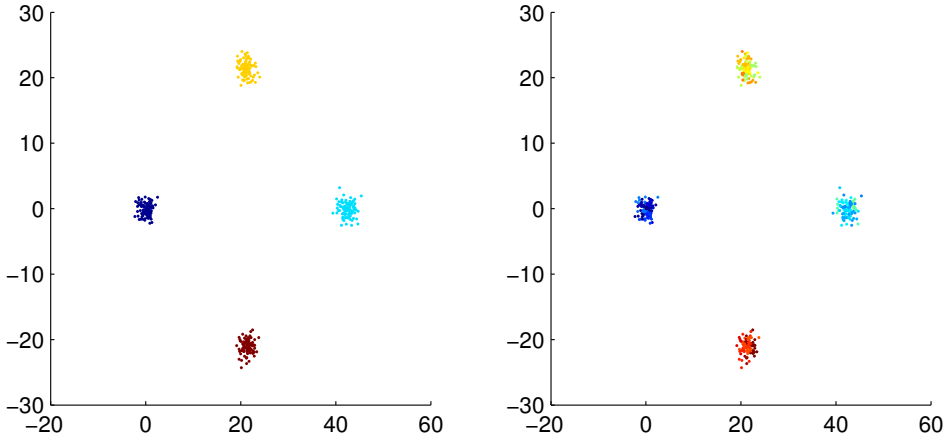

Figure 5.1: Results of Case 1 using MDL-MST returning 4 clusters (left) and AP returning 12 clusters (right).

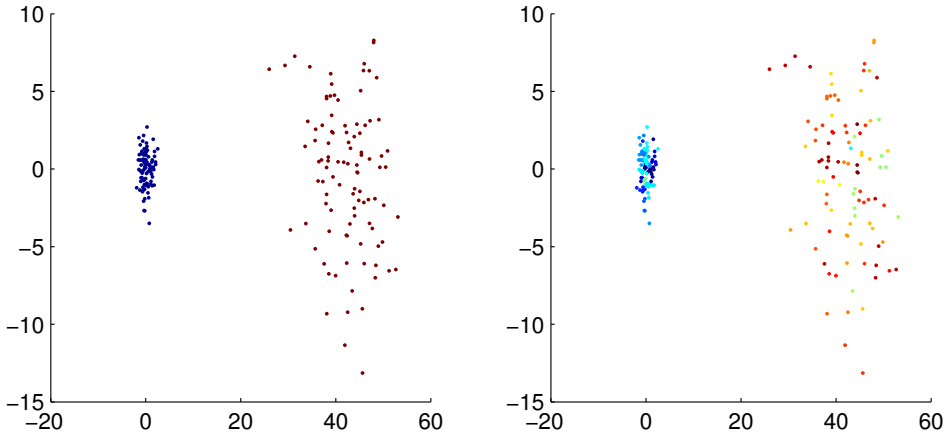

Figure 5.2: Results of Case 2 using MDL-MST returning 2 clusters (left) and AP returning 11 clusters (right).

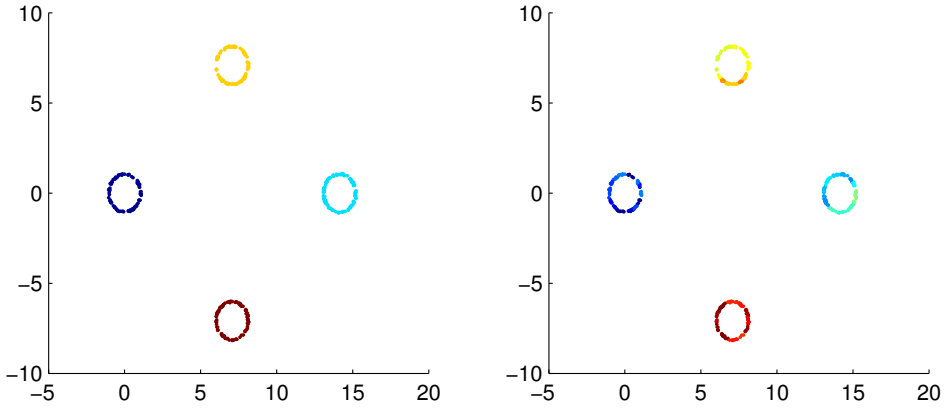

Figure 5.3: Results of Case 3 using MDL-MST returning 4 clusters (left) and AP returning 29 clusters (right).

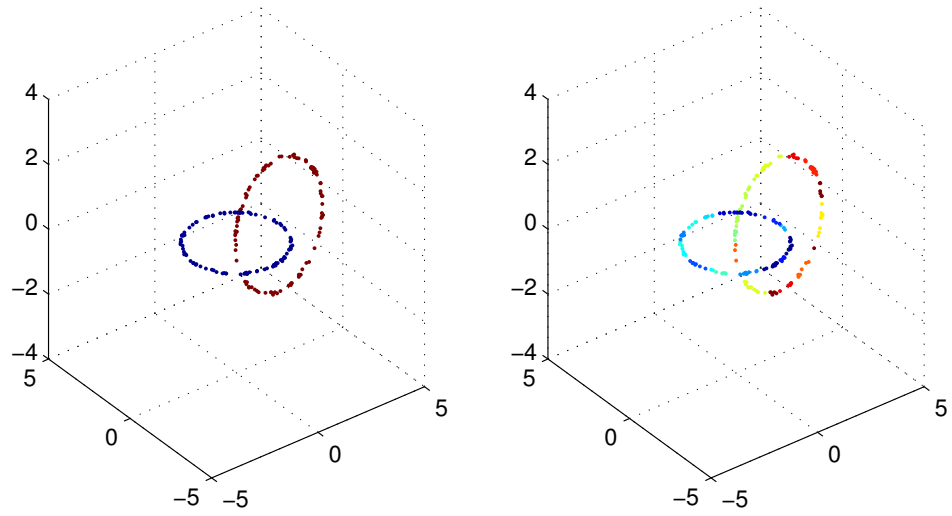

Figure 5.4: Results of Case 4 using MDL-MST returning 2 clusters (left) and AP returning 23 clusters (right)

variants, may provide important historical evidence related to the spread and interaction of variants with local cultures.

There are a number of mechanisms which lead to differences in variants. During the Middle Ages, Latin was no longer an actively spoken or written language. However, many texts were still copied in Latin; the copyists might understand a part of the text. This results in a large amount of unintentional copying error as well as intentional changes. Also for an original text being copied for centuries, the errors accumulate from one copy to another. These have resulted in large differences among surviving variants. Also, to construct a stemma, a number of variants must be considered simultaneously. The number of possible stemmata grows enormously with the number of variants: for example, there are 1.4×10^9 stemmata for 30 variants [28]. Hence, beyond traditional manual approaches, computer aided stemmatology methods are needed.

One can quickly notice that the problem of stemmatology is closely related to phylogenetics. The copying process with error is similar to genetic mutations during the evolution process. Also in both cases, there are missing variants. In biology, there may be no genetic data from extinct species. For both cases, variants whose word orders or genetic sequences are similar to each other are considered to be close in the resulting tree. Many automatic stemmatology methods are inspired by phylogenetic methods, and have been improved since the work of Robinson and O'Hara [68]. These methods have produced encouraging results as they have been applied and evaluated on small datasets where historians have strong confidence in historical relation among variants. For these datasets, there is a consensus stemma based on many forms of evidence.

Despite those successful automatic stemmatology results, several challenges remain. In particular, early test datasets are relatively small and ideal in that there are few missing variants, and most available variants have few missing portions. However as mentioned before, it is known that historical variants have missing portions due to physical damage. This poses additional challenges compared to phylogenetics, where in most cases, full gene or protein sequences are available. On the other hand, it is reasonable for phylogenetics to construct a bifurcation tree with all variants as leaves since there rarely is an occasion where more than two species mutated and evolved from an ancestor at exactly the same time, and the surviving species should indeed be the result of the latest mutations. This is not the case in stemmatology, where several copyists can copy from an identical source and

surviving variants need not be the latest copies. Other issues such as contamination where a variant is copied based on two or more sources are also unique in stemmatology. In the review by Roos and Heikkilä [70], 13 major algorithms are evaluated on three artificially generated datasets with known true stemmata. The datasets are generated by subjects copying texts but not real historical data. Notably one of them, the *Heinrichi* dataset, is a much more realistic dataset where nearly half of the variants are missing, and available variants have large missing portions. Even though the best performing method on the *Heinrichi* dataset obtains good accuracy slightly lower than results from simpler datasets, surprising failures of several promising methods, such as `CompLearn` [17], indicate that more should be done to address the issue of incomplete data [56].

5.6 Computational Challenges and Datasets

Inferring structure among data points in the presence of missing data is tied closely to a set of graphical optimization problems collectively called the Steiner Tree problem. In the Steiner tree problem, a graph $G = (V, E)$ and a subset S of V are given. The goal is to find a tree G' that connects all nodes in S and minimizes the total edge weights in G'. In stemmatology and phylogenetics, S is then the set of variants or genetic sequences of interests, and V is then the set of all possible variants or genetic sequences that are relevant to the problem at hand. The edge weight is a distance or similarity measure we pick. In general, the Steiner tree problem is $\mathcal{NP}$ hard, and it is even $\mathcal{NP}$ to have a close approximation. Under the case with missing portions in available variants, in the worst cases, even the optimal imputation and structural inference among only the available variants is a Steiner Tree problem.

Rather than adopting to our problem a general algorithm for the Steiner tree problem, we develop a novel approach based on the specific properties of the datasets in stemmatology. Our goal is to learn what concepts should yield algorithms that perform well. The current best performing algorithm RHM developed by Roos, *et at.,* is in fact closely related to a Steiner tree algorithm. We give an MDL interpretation of why RHM succeeds.

The *Heinrichi* dataset and the *Parzival* dataset are used to evaluate algorithm performances in the Computer-Assisted Stemmatology Challenge [70]. The *Heinrichi* dataset consists of an original text,

a 17th century late medieval Finnish folktale *Piispa Henrikin Surmavirsi*, written in old Finnish. 17 copyists participated to produce 67 text variants with contamination. The copyists are mostly Finnish but can only understand some ancient words, which resembles the situation in real stemmatology problems. In simulation, large portions of available variants are deleted on purpose, and only 37 variants are available. Thus it is similar to real world stemmatology with the physical damage and variants uncovered. Each variant has around 1200 words with an average of 300 missing words. *Heinrichi* is currently the most realistic data set with a large amount of incomplete data.

On the other hand, the *Parzival* [70] dataset is smaller consisting of 21 variants of the German poem *Parzival* by Matthew Spencer and Heather F. Windram. Only 5 out of the 21 variant are missing to the algorithm and no missing portions except those generated by copying error. This dataset is mostly for validation that any algorithm should produce reasonable performance on it, and it is easy to analyze results on this dataset.

5.7 Notations

A variant of a text is denoted as $x_j = (x_1^j, \cdots, x_n^j)$, where j is the variant index, n is the supposed number of words in a variant, and $x_i^j, i = 1, \cdots, n$ is a word or '?' if the word is missing at a location. The set of all variants of interest is denoted S. The number of variants of interest is denoted N. Here it is assumed that the variants are aligned. For this purpose, multiple aligmment techniques exist similar to the well known Needleman-Wunsh algorithm. While alignment is indeed a relevant issue, it is not the focus of this paper, and aligned data are used in simulations. A stemmatology structure, called the stemma, of a set of variants is a connected graph of nodes V and edges E such that $x_j \in V$ for all j. Note that V may contain auxiliary nodes depending on what algorithms are being used, and the graph need not be a tree as one person may produce a variant by referring to multiple sources, known as contamination in the stemmatology literature. The set of all words appearing in the set of variants is denoted $\mathcal{X}$, and the total number of elements in $\mathcal{X}$ is m.

In order to compare structural differences between two stemmatology graphs, the *average sign similarity* is introduced [70]. For a given undirected graph G, the simple path length between two nodes A and B is defined as the number of edges along the shortest path between A and B defined on

G. The simple path distance between two nodes on the true graph is denoted as $d(A, B)$, and the distance between them on the inferred graph is denoted as $d'(A, B)$. For any three nodes A, B, and C, the sign agreement index is defined as

$$u(B, C|A) = 1 - \frac{1}{2}|\text{sign}(d(A, B) - d(A, C)) \\ -\text{sign}(d'(A, B) - d'(A, C))|, \tag{5.4}$$

where $|\cdot|$ is the absolute value. This index measures if the proposed graph has the same ordering of B and C, given a reference node A, as the true graph. It is equal to 1 if the order is the same in the truth and the proposed graph, 1/2 if one and only one of them is zero, i.e. B and C have the same distance from A, and 0 if the ordering is mismatched. The *average sign similarity* between two stemmatology graphs G and H given a set of variants x_j is defined as

$$D(G, H) = \sum_{x_i \neq x_j \neq x_k} u(x_j, x_k|x_i)/6. \tag{5.5}$$

Note that the division by 6 is to discount equivalents due to permutation. Given a true structure T, the score of an inferred structure G is defined as $D(G, T)/D(T, T)$.

5.8 MDL Concepts for Stemmatology

The stemmatology problem can be approached using MDL and information theoretic ideas as in [70, 51, 17]. We start with assuming no missing data and no missing word in any variant to illustrate the key idea. The MDL concept to be presented has to be generalized for datasets with missing words and missing variants, which will be discussed in the following two subsections. In general, given an encoding function (Z), which may be a general purpose compression algorithm or model, we denote $Z(x_i)$ as the number of bits to encode x_i by itself, and $Z(x_i|x_j)$ as the number of bits to encode x_j given x_i. As mentioned in [51], finding a efficient encoding of all variants is equivalent to finding the minimum spanning tree of a fully connected undirected graph with variants as its nodes and pairwise code length as the length of its edges. Besides using a compression algorithm, one can also use a generic code with the indexed bag of words $\mathcal{X}$. To describe a variant, without any other

information, it takes $n \log m$ bits where $m = |\mathcal{X}|$. Furthermore, it takes

$$\log N + \log n + \log \binom{n}{k} + k \log m \tag{5.6}$$

bits to describe the variant x_i based on the variant x_j if they differ in k words. The first $\log N$ bits are used to describe the index j, $\log n$ bits are used to describe the number of differences, and $\log \binom{n}{k}$ bits are used to describe the locations of differences. In this encoding scheme, all variants have the same code length when coding by itself, and the code length from x_i to x_j is the same as x_j to x_i.

5.8.1 MDL Concepts for Missing Words

When there are words missing in some or even all variants, there are problems in determining the code length between two variants. Two rough approaches are to either encode locations with words missing, denoted as '?', as an additional symbol added to $\mathcal{X}$, or simply consider available words only and using some model or compression algorithm such as `gzip` for encoding. Both approaches have the same major drawback in cases where two variants have large number of locations with words missing. Suppose two variants x_i and x_j both have $n\frac{1}{4}$ words missing but at different locations, while the overlapping part of available words are identical. This high word missing rate is common in stemmatology datasets. On the other hand, direct encoding also exaggerates the difference that encoding the words of a child in locations where words are missing in the parent requires a large number of bits as if those words are directly encoded by themselves. This also suggests a reason why the normalized compression distance (NCD) based CompLearn does not work well in one of the testing datasets, the *Heinrichi* dataset, with high word missing rate [70]. The NCD is defined by

$$e_Z(A, B) = \frac{Z(A, B) - \min\{Z(A), Z(B)\}}{\max\{Z(A), Z(B)\}}, \tag{5.7}$$

where Z measures the number of bits required using a compression algorithm. For the example case under discussion, the NCD between those two variants will roughly be $\frac{1}{3}$. In general, when two variants have a large number of non-overlapping locations with words missing, the NCD would

be overly large. It is because $Z(A, B)$ takes roughly the number of bits to encode overlapping available words plus non-overlapping words, and $Z(A)$ takes roughly the number of bits to code overlapping available words plus words in A but missing in B. Assume without loss of generality that $Z(A) > Z(B)$, $Z(A, B) - Z(B)$ is then roughly the code length for words in A at locations that words are missing in B which is approximately the number of such locations times the entropy of words. After divided by $Z(A)$, the resulting NCD is largely biased by missing parts mismatch, which is mainly a result of physical degradation of text variants not the copying process. The negative effects of missing words and needs of specific approaches to deal with missing words for NCD are also discussed in detail in [56].

Under the MDL concept, the key is to encode data efficiently. Thus for locations with words missing, one should either not encode the words or encode them in a way that leads to efficient encoding of other variants. In fact, if missing words in x_i are available in x_j, one can have x_i filled in under a low bit cost. To do so, one first copies x_i as x_i and encodes only the difference between x_i and x_j whenever the words are available in x_j. Clearly due to this encoding strategy, the actual number of bits to encode x_i given x_j further depends on the parent of x_j. For example shown in 5.5, encoding x_j given A or B requires the same number of bits. However, choosing B as the parent of x_j leads to a shorter total description length using the described encoding method. Likewise, if one fixes the parent variant, what to fill in depends on its children. Thus searching the shortest code length becomes a problem of simultaneously finding the optimal tree structure as well as optimal words to fill in locations with words missing for all variants that are not leaves of the resulting tree. This problem is related to the well-known Steiner tree problem, which is know to be $\mathcal{NP}$-hard.

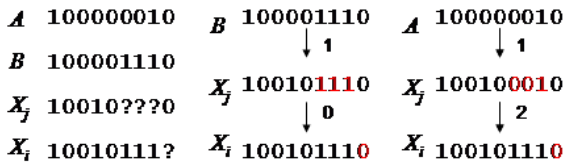

Figure 5.5: The total code length depends on what are filled in locations with words missing, which in turn depends on the tree structure.

As the difficulty of the Steiner tree problem lies in estimating the miss parts of the variants, one can instead attempt to directly estimate the code length between two variants based only on the

available parts. We can view that there exists a channel from one variant to the other with channel capacity denoted as $C_{i|j}$. The code length from x_j to x_i is then $n(1 - C_{i|j})$ when n is large. Thus in analogy, given a compression algorithm Z, we can estimate the $Z(x_i|x_j)$ as $\frac{n}{n_{i,j}} Z(\hat{x}_i|\hat{x}_j)$, where $\hat{\cdot}$ denotes the overlapping available part, and $n_{i,j}$ is the length of the overlapping available part. This assumes that the statistical property of the differences of available words are asymptotically the same as the channel. This is a reasonable model for typos but not for insertion and deletion. Since we are considering the parts available in both variants and the variants are aligned, the effects of insertion or deflection may be suppressed. We can also use (5.6) by substituting n with $n_{i,j}$ and k with $\hat{k}$ which is the number of differences in the overlapping part.

5.8.2 MDL Concepts for Missing Variants

In real stemmatology or phylogenies problems, it is known that not all variants are available in the dataset. As mentioned earlier, there is a strong emphasis on learning the correct structural relation among variants in stemmatology problems. Hence it is reasonable to ask, if it would lead to a more efficient code when one adds auxiliary variants in the inferred tree. It may be a surprise that the answer is indeed yes, and also adding auxiliary variants help infer the stemma better. For any three nodes A, B, C where A is the parent, if there exists a $D \in \mathcal{X}^n$ that

$$Z(D|A) + Z(B|D) + Z(C|D) < Z(B|A) + Z(C|A), \tag{5.8}$$

then adding D as an auxiliary node leads to a more efficient encoding. Likewise if D is an auxiliary node, (5.8) can be used to decide if D should be removed. For example, using the encoding given by (5.6) for A, B, and C, it is easy to show that when there is a large portion of locations where B and C are the same but different from A, adding an auxiliary node D helps encoding. Such a node D is constructed to coincide with B on those locations otherwise the same as A leads to a more efficient encoding. Practically speaking, (5.8) can only be evaluated once the encoder Z and nodes to be considered are known, but the main concept here is that conditions satisfying (5.8) exist, and (5.8) can be used to add or delete auxiliary nodes in a principled way.

Thus, when there are no missing words in any variant, finding the most efficient encoding of variants of interest given Z while allowing adding auxiliary nodes is a case of Steiner tree problems. In a Steiner tree problem, a graph $G = (V, E)$ and a strict subset of nodes $S \subset V$ are given. The goal is to find a subgraph G' of G with minimum total edge length such that G' is connected and contains all nodes in S. S can be viewed as the variants of interests. V can be viewed as the set of all the length n sequence of words from $\mathcal{X}$. The edges E are the code lengths defined by Z. A Steiner tree problem is in general $\mathcal{NP}$-complete. There exist a number of heuristic search and approximation algorithms (see [34] for a classical review). The RHM algorithm described in [69] utilizes a similar concept while forcing the stemma to be a bifurcating tree.

5.9 MDL for Computational Stemmatology Simulation Results and Conclusions

For a given encoding method Z and a set of variants of interest S, run the following algorithm:

for $i \neq j$ **do**
  $L = \{l : x_l^i, x_l^j \neq ?\}, n_{i,j} = |L|$
  $\hat{x}_i \leftarrow$ concatenation of $x_l^i, l \in L$
  $\hat{x}_j \leftarrow$ concatenation of $x_l^j, l \in L$
  $E_{i,j} = \frac{n}{n_{i,j}} Z(\hat{x}_i, \hat{x}_j)$
end for
return $G = \mathrm{MST}(S, E)$

This is closely related to the well-known Chou-Liu algorithm [16]. When Z is chosen to the generic MDL encoding whose code length is described as in (5.6), one can note that $\log \begin{pmatrix} n_{i,j} \\ k_{i,j} \end{pmatrix}$ is approximately $n_{i,j} H(\frac{k_{i,j}}{n_{i,j}})$ when $n_{i,j}$ is large, remembering $k_{i,j}$ is the number of differences between $\hat{x}_i$ and $\hat{x}_j$. Also if $k_{i,j}$ is small relative to $n_{i,j}$, $H(\frac{k_{i,j}}{n_{i,j}})$ can be approximate by a straight line $c\frac{k_{i,j}}{n_{i,j}}$ where c is a constant. This reduces the numerical problem for computing large combinatorial terms. Thus

Table 5.1: Performance of 14 algorithms on the Heinrichi dataset and the Parzival dataset

Method	Heinrichi (%)	Parzival (%)
MDL	79.0	78.5
RHM	76.0	79.9
Parsimony	74.4	77.8
Parsimony BS	73.6	85.4
Neighbor Joining	64.4	81.5
Neighbot Joining BS	62.9	87.1
Least squares	64.2	81.5
Least squares BS	62.6	79.8
n-Gram clustering	64.4	79.3
NeighborNet	59.1	77.8
SplitDecomp.	53.1	74.5
ParsimonySplits	56.8	83.7
CompLearn	52.7	81.5
Hierarchical clustering	51.4	72.6

(5.6) becomes

$$\log N + \log n_{i,j} + k_{i,j}(c + \log m). \tag{5.9}$$

By keeping the leading term associated with $k_{i,j}$ of (5.9) and plugging into the computation of edge length E we have

$$E_{i,j} = \frac{k_{i,j}}{n_{i,j}} n(c + \log m). \tag{5.10}$$

Note that the minimum spanning tree is invariant to uniform scaling in the input edge lengths, hence only the ratio between $k_{i,j}$ and $n_{i,j}$ is needed in this case. This results in a simple normalized Hamming distance approximating the code length between pairs of variants.

Using the result in (5.9), we construct a stemmatology structure for both the *Heinrichi* and *Parzival* dataset used in [70]. The resulting *average sign similarity* described in 5.5 for the *Heinrichi* dataset is 79.0%, which is better then all 13 algorithms reviewed in [70]; only three of them achieve above 65%. The resulting tree is shown in Figure 3. For the Parzival dataset, the result is 78% which is around the middle of the 13 algorithms with 6 of them above 80%. The results of the Parzival dataset is shown in Figure 5.9, with the ground truth presented in Figure 5.8. Table 5.9 summerizes our performance along with the 13 algorithms reviewed in [70].

5.10 A Note on Density Based Clustering and MDL

In this section, we briefly review of MDL principle and the density based clustering. We start from the stochastic complexity as the foundation of MDL, and then describe how to use MDL principle in density based clustering. Stochastic complexity as described in [7] and [66] starts from classes of models, or probability distributions

$$\mathcal{M}_\gamma = \{P(x|\theta,\gamma) : \theta \in \Theta_\gamma\}, \tag{5.11}$$

and the union of model classes $\mathcal{M} = \bigcup_{\gamma\in\Gamma} \mathcal{M}_\gamma$. If γ is known, a naïve guess for the shortest description length is $\log P(\underline{x}|\hat{\theta},\gamma)^{-1}$, where $\hat{\theta}(x)$ is the maximum likelihood (ML) estimate of θ given $\underline{x}$. While the shortest length itself is computable, codes based on $\log P(\underline{x}|\hat{\theta},\gamma)^{-1}$ are not decodable, since information about the ML estimate is not yet coded. One reasonable objective is to have a code which minimizes the worst case redundancy over the optimal code if the ML estimate was given. The length of such a minimum redundancy achieving code is defined as the *stochastic complexity* of $\underline{x}$ relative to the model class $\mathcal{M}_\gamma$.

This can be formulated as a problem of selecting a distribution $Q(\underline{x})$ which is computable given $\underline{x}$ and the model class $\mathcal{M}_\gamma$, such that $Q(\underline{x})$ minimizes the worse case code length redundancy over shortest description length, which is

$$\max_{\underline{x}} \log \frac{P(\underline{x}|\hat{\theta}(\underline{x}),\gamma)}{Q(\underline{x})}. \tag{5.12}$$

The optimal distribution found in [75] is

$$Q^* = \frac{P(\underline{x}|\hat{\theta}(\underline{x}),\gamma)}{\sum_{\underline{x}\in\underline{\mathcal{X}}} P(\underline{x}|\hat{\theta}(\underline{x}),\gamma)}, \tag{5.13}$$

which is called the normalized maximum likelihood (NML) distribution. The code length using $Q^*(\underline{x})$ is

$$\log \frac{1}{P(\underline{x}|\hat{\theta}(\underline{x}),\gamma)} + \log \sum_{\underline{x}\in\underline{\mathcal{X}}} P(\underline{x}|\hat{\theta}(\underline{x}),\gamma). \tag{5.14}$$

The second part is due to the unknown parameters and is defined as the *parametric complexity* [7]. Note that it can be shown that two-part codes achieve the same code length asymptotically [66]. In

two part codes, the code length is

$$\log \frac{1}{P(\underline{x}|\hat{\theta}(\underline{x}), \gamma)} + L\left(\hat{\theta}(\underline{x})\right), \tag{5.15}$$

where L is the code length of its input variable. Under the conditions of [66], the Fisher information matrix $I(\theta)$ of $P(\underline{x}|\hat{\theta})$ given $\underline{x}$ exists and the parametric complexity asymptotically achieves

$$\frac{d_\theta}{2} \log \frac{n}{2\pi} + \log \int |I(\theta)|^{\frac{1}{2}} d\theta + o(1) \tag{5.16}$$

as n gets large, where d_θ is the dimension of θ.

Comments Computation of the stochastic complexity of some model class $\mathcal{M}_\gamma$ can be approached from either the NML coding (5.14) or two part codes by checking the conditions of Rissanen [66]. The NML coding imposes less constraints, while two part codes are more intuitive and results derived by Rissanen [66] largely reduce the computation cost. One should note that stochastic complexity is a property of a model class. Only discrete data can possibly be encoded in finite length. If $\underline{x}$ is continuous, one can discretized $\underline{x}$ to some precision δ to approximate the coding length. The summation in the denominator of (5.13) is changed to an integral. If the integral exists, the stochastic complexity is then (5.14) or (5.15) with an additional precision cost $\log \delta$.

The goal of density based clustering is to take $\underline{x}$ as input and output a cluster index vector $\underline{c}_K$ whose ith element is an integer in $\{1, \cdots, K\}$ as the cluster index for x_i. We can rewrite θ in two parts $\theta = (\underline{c}_K, \underline{\theta}_{K,\eta})$, where $\theta_{k,\eta}$ denotes the parameters of the density of the kth cluster.

Theorem 1: For density based clustering, if conditions in Section II of [66] are satisfied, the stochastic complexity of $\underline{x}$ to a model class $\gamma = (K, \eta)$ is upper bounded by

$$\begin{aligned} & \sum_{i=1,\cdots,N} \log \frac{1}{P(x_i|\hat{c}_i, \hat{\theta}_{c_i,\eta}(\underline{x}))} \\ + \quad & \log \sum_{j=0,\cdots,K-1} (-1)^j \binom{K}{K-j} (K-j)^N \\ + \quad & \frac{d_\theta}{2} \log \frac{n}{2\pi} + \log \int |I(\theta)|^{\frac{1}{2}} d\theta + o(1) + L(K, \eta), \end{aligned}$$

where $\hat{c}_i$ and $\hat{\theta}_{c_i,\eta}$ are ML estimates given $\underline{x}$, and $L(K,\eta)$ is the code length of the model class index, which is $\log \frac{1}{P_{K,\eta}(K,\eta)}$ if a prior is assumed, or $\log N|\Gamma|$ otherwise.

Proof: For two part codes, the code length is

$$\log \frac{1}{P(\underline{x}|\underline{\hat{c}}_K, \underline{\hat{\theta}}_{K,\eta}(\underline{x}))} + L(\underline{\hat{c}}_K, \hat{\theta}_{K,\eta}(\underline{x})) + L(K,\eta), \tag{5.17}$$

From the independence assumption, the first part is then

$$\begin{aligned} & \sum_{i=1,\cdots,N} \log \frac{1}{P(x_i|\underline{\hat{c}}_K, \underline{\hat{\theta}}_{K,\eta}(\underline{x}))} \\ = & \sum_{i=1,\cdots,N} \log \frac{1}{P(x_i|\hat{c}_i, \hat{\theta}_{c_i,\eta}(\underline{x}))}, \end{aligned} \tag{5.18}$$

where the last equality is to evaluate the code length of each x_i using its cluster density. The second part can be upper bounded by the length of encoding $\underline{c}_K$ and $\underline{\theta}_{K,\eta}$ independently. The code length of $\underline{c}_K$ is loosely upper bound by $N \log K$. A tighter bound can be obtained by counting the number of sequences the elements can take, and must include all integer values from 1 to K, which is

$$\sum_{j=0,\cdots,K-1} (-1)^j \binom{K}{K-j} (K-j)^N. \tag{5.19}$$

The worst case code length is the log of (5.19) if each sequence is assumed to be equally likely. Given an assumed family of densities, the code length of $\underline{\theta}_{K,\eta}$ can be computed using the Fisher information matrix as shown in [66], which is equal to the parametric complexity expressed in (5.16). $L(K,\eta)$ is trivial. This complete the proof of theorem 1.

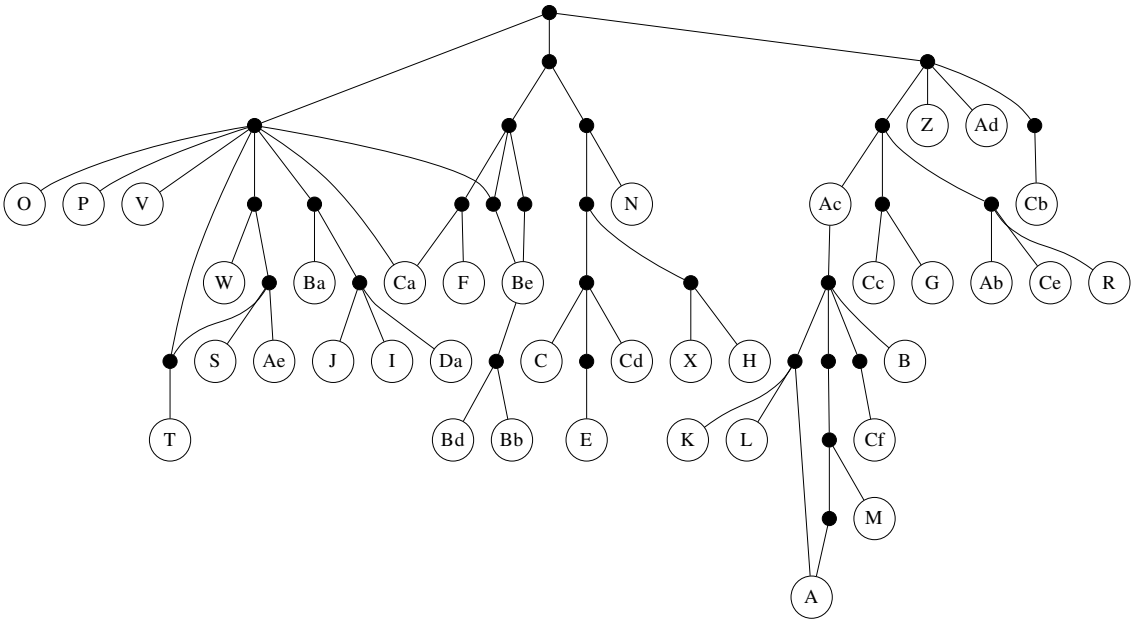

Figure 5.6: The true stemma of the *Heinrichi* dataset. Filled dots represent missing variants. Nodes with two parents are due to contamination.

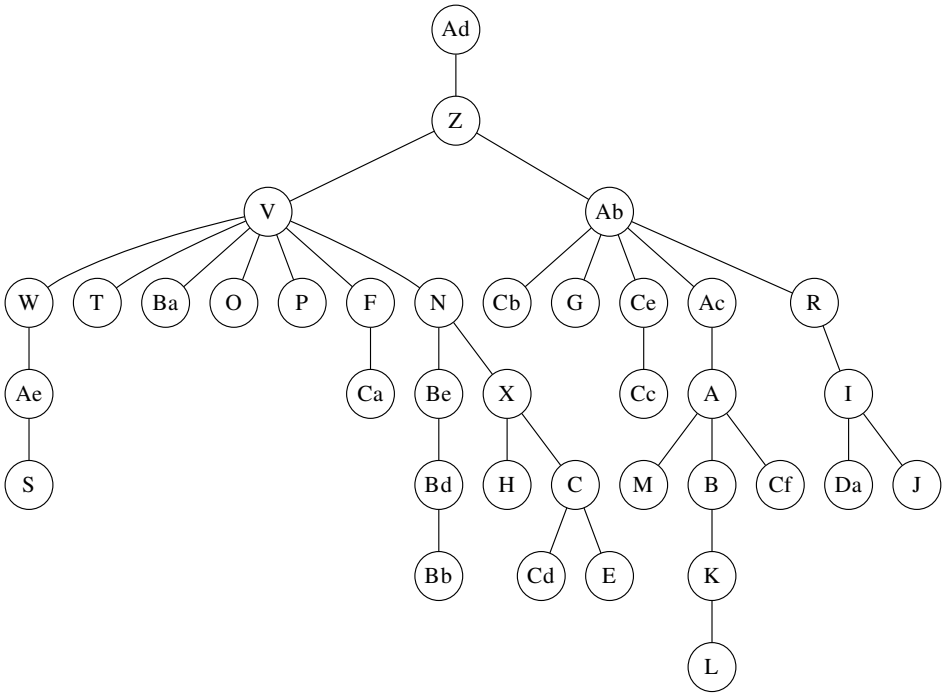

Figure 5.7: The resulting stemma generated from the minimum spanning tree based on the generic MDL code. Note its similarity with the true stemma can be further noticed by focusing on the neighborhood relation among available nodes (labeled with alphabets) which may be connected through unavailable nodes. For example, in the true graph node B is in fact closely connected to A, Cf, M, L, and K while equally far away from the group of Be, Bd, Bb and the group of C, Cd, E, as the inferred stemma suggests.

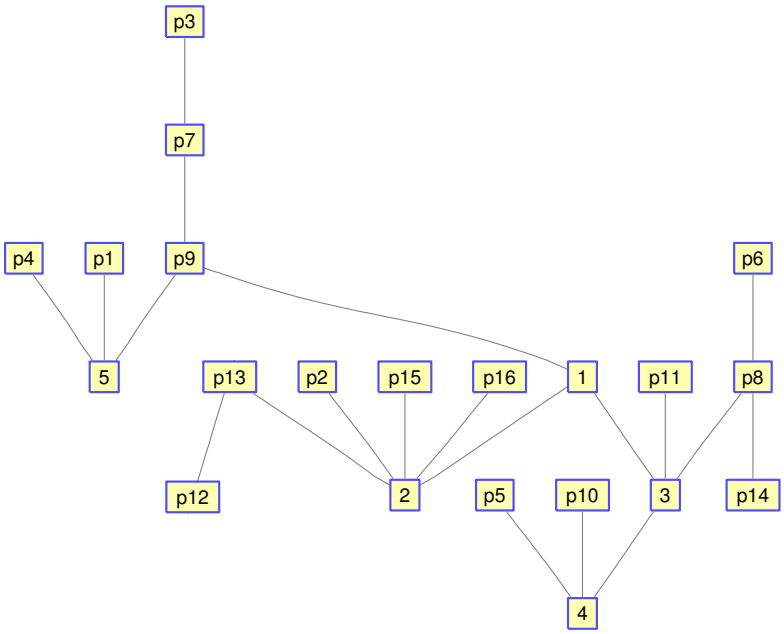

Figure 5.8: The true stemma of the *Parzival* dataset. The nodes labeled with pure numbers are missing variants that are not available to the algorithm

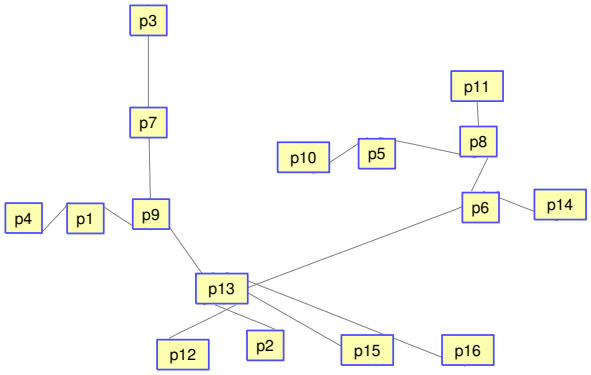

Figure 5.9: The inferred stemma of the *Parzival* dataset is the minimum spanning tree of a graph with edge weight being the normalized Hamming distance. Note that in the true stemma, there are five variants unavailable to the algorithm. This results in several errors in the sign similarity measure. For example variant 8 is directly connect to variant 5 and variant 6 in the inferred stemma, while there are actually two missing variants between variant 8 and 5. On the other hand, in the true stemma if we view two variants connected though unavailable variants as directly connected, the inferred structure is actually close to the true structure.

References

[1] A. Adler. Vulnerabilities in biometric encryption systems. In *Proceedings of the 5th International Conference on Audio and Video Based Biometric Person Authentication*. Hilton Rye Town, NY, USA, 2005.

[2] Foteini Agrafioti and Dimitrios Hatzinakos. ECG based recognition using second order statistics. In *Proceedings of the Communication Networks and Services Research Conference 2008, Nova Scotia, Canada*, 2008.

[3] R. F. Ahlswede and I Csiszàr. Common randomness in information theory and cryptography. I. secret sharing. *IEEE Transaction on Information Theory*, 39(4):1121–1132, 1993.

[4] R. F. Ahlswede and I Csiszàr. Common randomness in information theory and cryptography. II. CR capacity. *IEEE Transaction on Information Theory*, 44(1):225–240, 1998.

[5] R. F. Ahlswede and J. Korner. Source coding with side information and a converse of degraded broadcasting channels. *IEEE Trans. Info. Theory*, 21:629–637, 1975.

[6] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Autom. Control*, 19:716–722, 1974.

[7] A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Info. Theory*, 44:2473–2760, 1998.

[8] Charles H. Bennett, Ming Li, and Bin Ma. Chain letters and evolutionary histories. *Scientific American*, 288(6):76–81, 2003.

[9] Lena Biel, Ola Pettersson, Lennart Philipson, and Peter Wide. ECG analysis: a new approach in human identification. In *Proceedings of the IEEE Instrumentation and Measurement Technology Conference*, volume 16, 1999.

[10] Charles A. Bouman, Michael Shapiro, Gregory W. Cook, C. Brian Atkins, Hui Cheng, Jennifer G. Dy, and Sean Borman. Cluster: An unsupervised algorithm for modeling gaussian mixtures. April. Available from http://www.ece.purdue.edu/~bouman.

[11] Adrian Da Ca Chan, Mohyledin M. Hamdy, Armin Badre, and Vesal Badee. Wavelet distance measure for person identification using electrocardiograms. *IEEE Transactions on Instrumentation and Measurement*, 57(2):248–253, 2008.

[12] Mei Chen. *LDV Signals for Biometrics*. Master Thesis, Washington University in Saint Louis,, 2007.

[13] Mei Chen, Joseph A. O'Sullivan, Alan D. Kaplan, Po-Hsiang Lai, Eric J. Sirevaag, and John W. Rohrbaugh. Biometrics with physical exercise using laser doppler vibrometry measurements of the carotid pulse. In *The First IEEE International Conference on Biometrics, Identity and Security, Tampa, Florida, USA*, 2009.

[14] Mei Chen, Joseph A. O'Sullivan, Naveen Singla, Eric J. Sirevaag, and John W. Rohrbaugh. Laser doppler vibrometry measures of physiological function: evaluation of biometric capabilities. *IEEE Transaction on Information Forensics and Security*, (5), 2010.

[15] Chuang-Chien Chiu, Chou-Min Chuang, and Chih-Yu Hsu. A novel personal identity verification approach using a discrete wavelet transform of the ecg signal. In *The 2nd International Conference on Multimedia and Ubiquitous Engineering, Busan, Korea*, 2008.

[16] CK Chou and CN Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

[17] Rudi Cilibrasi and Paul M. B. Vitányi. Clustering by compression. *IEEE Trans. Info. Theory*, 51:1523–1524, 2005.

[18] David Pereira Coutinho, Ana L. N. Fred, and Mario A. T. Figueiredo. One-lead ecg-based personal identification using ziv-merhav cross parsing. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3858–3861, aug. 2010.

[19] David Pereira Coutinho, Ana L. N. Fred, and Mario A. T. Figueiredo. Personal identification and authentication based on one-lead ecg using ziv-merhav cross parsing. In *Proceedings of the 10th International Workshop on Pattern Recognition in Information Systems (PRIS)'10, Funchal, Portugal*, pages 15–24, 2010.

[20] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, Inc., New York, NY., 1991.

[21] Imre Csiszár. Linear codes for sources and source networks: Error exponents, universal coding. *IEEE Transactions on Information Theory*, 28(4):585–592, 1982.

[22] Imre Csiszár and János Körner. *Information Theory*. Cambridge University Press, 2011.

[23] Luc Devroye, Laszlo Györfi, and Gabor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[24] A. W. Eckford, F. R. Kschischang, and S. Pasupathy. Analysis of ldpc codes in channels with memory. Proceedings of the 21st Queen's Biennial Symposium on Communication, Ontario, Canada, 2002.

[25] M. K. Reiter F. Monrose and S. Wetzel. Password hardening based on keystroke dynamics. In *Proceedings of the 6th ACM conference on Computer and Communications Security*. 2001.

[26] Shih-Chin Fang and Hsiao-Lung Chan. Human identification by quantifying similarity and dissimilarity in electrocardiogram phase space. *Pattern Recogn.*, 42:1824–1831, 2009.

[27] S. Zahra Fatemian and Dimitrios Hatzinakos. A new ECG feature extractor for biometric recognition. In *The 16th International Conference on Digital Signal Processing*, 2008.

[28] Joseph Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Inc., 2004.

[29] Brendan J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[30] Nir Friedman. The Bayesian structural EM algorithm. In *Proceedings of the 15th international conference on Machine learning*, 1998.

[31] Keinosuke Fukunaga and Raymond R. Hayes. Effects of sample size in classifier design. *IEEE Transactions on Pattern Analysis and Machine Learning*, 11(8):873–885, 1991.

[32] J. Garcia-Frias and W. Zhong. Approaching shannon performance by iterative decoding of linear codes with low-density generator matrix. *IEEE Communication Letters*, 7:266–268, 2003.

[33] Peter D. Grünwald. *The Minimum Description Length Principle*. MIT press, 2007.

[34] Frank K. Hwang, Dana S. Richards, and Pawel Winter. *The Steiner Tree Problem*. North Holland, 1992.

[35] Tanya Ignatenko and Frans M J Willems. Biometric systems: Privacy and secrecy aspects. *IEEE Transactions on Information Forensics and Security*, 4(1).

[36] Tanya Ignatenko and Frans M J Willems. Fundamental limits for biometric identification with a database containing protected templates. In *Proceeding of the international Symposium on Information Theory and its Applications, Taiwan*. 2010.

[37] J.M. Irvine, B.K. Wiederhold, L.W. Gavshon, S.A. Israel, S.B. McGehee, R. Meyer, and M.D. Wiederhold. Heart rate variability: a new biometric for human identification. In *International Conference on Artificial Intelligence (IC-AI'2001), Las Vegas, Nevada, USA*, 2001.

[38] John M. Irvine, Steven A. Israel, W. Todd Scruggs, and William J. Worek. eigenPulse: Robust human identification from cardiovascular function. *Pattern Recognition*, 41:3427–3435, 2008.

[39] Anil K Jain, Karthik Nandakumar, and Abhishek Nagar. Biometric template security. *EURASIP Journal on Advances in Signal Processing*, 2008:1–17, 2008.

[40] Anil K. Jain, Arun Ross, and Umut Uludag. Biometric template security: challenges and solutions. In *in Proceedings of the European Signal Processing Conference (EUSIPCO 05), Antalya, Turkey*. 2005.

[41] A. Juels and M. Wattenberg. A fuzzy commitment scheme. In *Proceedings of the 6th ACM Conference on Computer and Communications Security*, pages 28–36. 1999.

[42] A. Juels and M. Wattenberg. A fuzzy vault scheme. In *Proceedings of the 2002 International Symposium on Information Theory*. 2002.

[43] Alan D. Kaplan, Joseph A. O'Sullivan, Erik J. Sirevaag, and John W. Rohrbaugh. Laser doppler vibrometry measurements of the carotid pulse: biometrics using hidden markov models. In *Proceedings of SPIE, Orlando, Florida, USA*. 2009.

[44] Ibrahim Khalil and Fahim Sufi. Legendre polynomials based biometric authentication using QRS complex of ECG. In *Proceedings of the 4th International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, 2008.

[45] Petri Kontkanen, Petri Myllymäki, Wray Buntine, Jorma Rissanen, and Henry Tirri. An MDL framework for data clustering. In Peter D. Grunwald, In Jae Myung, and Mark A. Pitt, editors, *Advances in Minimum Description Length*, pages 323–353. MIT press, Cambridge, Massachusetts, 2005.

[46] Masaki Kyoso and Akihiko Uchiyama. Development of an ECG identification system. In *Proceedings of the 23rd Annual EMBS International Conference, Istnbul, Turkey*, 2001.

[47] Lifeng Lai, Siu-Wai Ho, and Vincent Poor. Privacy-security trade-offs in biometric security systems part i: Single use case. *IEEE Transactions on Information Forensics and Security*, 6(1).

[48] Po-Hsiang Lai and Joseph A. O'Sullivan. Pattern recognition system design with linear encoding for discrete patterns. In *Proceeding of IEEE International Symposium on Info. Theory (ISIT), Nice, France*. 2007.

[49] Po-Hsiang Lai and Joseph A. O'Sullivan. Toward optimal trade-off between identification and secrecy-key binding using linear codes. In *Proceeding of the 2011 IEEE International Symposium on Information Theory (ISIT), Saint Petersburg, Russia*. 2011.

[50] Po-Hsiang Lai, Joseph A. O'Sullivan, Mei Chen, Eric J. Sirevaag, Alan D. Kaplan, and John W. Rohrbaugh. A robust feature selection method for noncontact biometrics based on laser doppler vibrometry. In *Biometric Symposium (BSYM), Tampa, Florida, USA*. 2008.

[51] Po-Hsiang Lai, Joseph A. O'Sullivan, and Robert Pless. Minimum description length and clustering with exemplars. Proceedings of 2009 IEEE International Symposium on Information Theory, Seoul, Korea, 2009.

[52] Ming Li and S. Narayanan. Robust ecg biometrics by fusing temporal and cepstral information. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 1326–1329, aug. 2010.

[53] Yingbin Liang, H. Vincen Poor, and Shlomo Shamai. Information theoretic security. *Fundations and Trends in Communications and Information Theory*, 5(4-5):355–580, 2008.

[54] André Lourenço, Hugo Silva, and Ana Fred. Unveiling the biometric potential of finger-based ecg signals. *Computational Intelligence and Neuroscience*, 2011(720971), 2011.

[55] Emin Martinian, Sergey Yekhanin, and Jonathan Yedidia. Secure biometrics via syndromes. In *Proceedings of the 2005 Allerton Conference*. 2005.

[56] Toni Merivuori and Teemu Roos. Some observations on the applicability of normalized compression distance to stemmatology. In *Proceedings of 2nd Workshop on Information Theoretic Methods in Science and Engineering*, 2009.

[57] Gary Garcia Molina, Fons Bruekers, Cristian Presura, Marijn Damstra, and Michiel van der Veen. Morphological synthesis of ECG signals for person authentication. In *European Signal Processing Conference, Poznan, Poland*, 2007.

[58] C. Nicola, F. Alajaji, and T. Linder. Decoding ldpc codes over binary channels with additive markov noise. Proceedings of the 2005 Canadian Workshop on Information Theory, Montreal, Canada, 2005.

[59] Ikenna Odinaka, Po-Hsiang Lai, Alan D. Kaplan, Joseph A. O'Sullivan, Eric J. Sirevaag, Sean D. Kristjansson, and John W. Rohrbaugh. Ecg biometrics: A robust short-time frequency analysis. In *2010 IEEE International Workshop on Information Forensics and Security (WIFS)*.

[60] Joseph A. O'Sullivan and Po-Hsiang Lai. Pattern recognition system design based on ldpc matrices. In *Proceedings of the 2005 IEEE International Symposium on Information Theory*, pages 33–36. Adelaide, Australia, 2005.

[61] D. Pelleg and A. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceesings of the 17th international conference of machine learning, San Francisco, USA*, pages 727–734, 2000.

[62] Christain Rathgeb and Andreas Uhl. A survey on biometric cryptosystems and cancelable biometrics. *EURASIP journal on information security*, 3, 2011.

[63] Saruans J. Raudys and Anil K. Jains. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Learning*, 13(3):252–264, 1991.

[64] J. Rissanen. A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11:417–431, 1983.

[65] J. Rissanen. Universal coding, information, prediction and estimation. *IEEE Trans. Info. Theory*, 30:629–636, 1984.

[66] J. Rissanen. Fisher information and stochastic complexity. *IEEE Trans. Info. Theory*, 42:40–47, 1996.

[67] J. Rissanen. *Information and Complexity in Statistical Modeling*. Springer, 2007.

[68] Peter Robinson and Robert J. O'Hara. Report on the textual criticism challenge 1991. *Bryn Mawr Classical Review*, 3(4):331–337, 1992.

[69] Teemu Roos, Toumas Heikkilä, and Petri Myllymäki. A compression-based method for stemmatic analysis. In *Proceedings of 17th European Conference on Artificial Intelligence*, pages 805–806, 2009.

[70] Teemu Roos and Tuomas Heikkilä. Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets. *Literary and Linguistic Computing*, 24(4):417–433, 2009.

[71] Natalia A. Schmid and Joseph A. O'Sullivan. Thresholding method for reduction of dimensionality. *IEEE Trans. Information Theory*, 47(7):2903–2920, 2001.

[72] Natalia A. Schmid and Joseph A. O'Sullivan. Performance prediction methodology for biometric systems using a large deviations approach. *IEEE Transactions on Signal Processing*, 52(10):3036–3045, 2004.

[73] Alexander Schrijver. *Combinatorial Optimization*. Springer, Berlin, 2003.

[74] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 78(2):461–464, 1978.

[75] Y. M. Shtarkov. Universal sequential coding of single messages. *Probl. Inform. Transm.*, 76(3):3–17, 1987.

[76] Bernard W. Silverman. *Density Estimation*. Chapman and Hall, New York, NY, USA, 1986.

[77] Richard Souvenir and Robert Pless. Manifold clustering. In *Proceedings of the 10th International Conference on Computer Vision, Beijing, China*, 2005.

[78] Fahim Sufi and Ibrahim Khalil. An automated patient authentication system for remote telecardiology. In *Proceedings of the 4th International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, 2008.

[79] Fahim Sufi, Ibrahim Khalil, and Ibrahim Habib. Polynomial distance measurement for ecg based biometric authentication. *Security and Communication Networks online*, 2008.

[80] P. Syverson. A taxonomy of replay attacks. In *Proceedings of the Computer Security Foundations Workshop*. Franconia, NH, USA, 1994.

[81] Ertem Tuncel. Capacity/storage tradeoff in high-dimensional identification systems. *IEEE Transactions on Information Theory*, 55(5):2097–2016, 2009.

[82] Stijn Marinus van Dongen. *Graph clustering by flow simulation*. Doctoral Thesis, University of Ultrecht, 2000.

[83] C.S. Wallace and D.L. Dowe. Minimum message length and Kolmogorov complexity. *The Computer Journal*, 42:1294–1299, 1999.

[84] Yongbo Wan and Jianchu Yao. A neural network to identify human subjects with electrocardiogram signals. In *Proceedings of the World Congress on Engineering and Computer Science*, 2008.

[85] Yongjin Wang, Foteini Agrafioti, Dimitrios Hatzinakos, and Konstantinos N. Plataniotis. Analysis of human electrocardiogramfor biometric recognition. *EURASIP Journal on Advances in Signal Processing*, 2008(148658), 2008.

[86] Brandon Westover and Joseph A. O'Sullivan. Achievable rates for pattern recognition. *IEEE Transactions of Information Theory*, 54(1):299–320, 2008.

[87] M. B. Westover. *Image representation and pattern recognition in brains and machines*. Doctoral Thesis, Washington University in Saint Louis,, 2006.

[88] M. B. Westover and J. A. O'Sullivan. Achievable rates for pattern recognition. arXiv cs 0509 0509022, 2006.

[89] Frans M J Willems and Tanya Ignatenko. Identification and secret-key binding in binary-symmetric template-protected biometric systems. In *Proceeding of the IEEE international Symposium on Information Theory, Kanagawa, Japan*. 2010.

[90] Gerd Wubbeler, Manuel Stavridis, Dieter Kreiseler, Ralf-Dieter Bousseljot, and Clemens Elster. Verification of humans using the electrocardiogram. *Pattern Recognition Letters*, 28:1172–1175, 2007.

[91] Rui Xu and Donald Wunsch, II. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16:645–678, 2005.

[92] Jianchu Yao and Yongbo Wan. A wavelet method for biometric identification using wearable ECG sensors. In *Proceedings of the 5th International Workshop on Wearable and Implantable Body Sensor Networks*, 2008.

[93] Can Ye, M.T. Coimbra, and B.V.K.V. Kumar. Investigation of human identification using two-lead electrocardiogram (ecg) signals. In *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, pages 1–8, 2010.

[94] Raymond W. Yeung. *A first course in information theory*. Kluwer Academic and Plenum Publishers, 2002.

[95] W. Zhong, H. Lou, and J. Garcia-Frias. Ldgm codes for joint source-channel coding of correlated sources. 2003.

Vita

Po-Hsiang Lai

Date of Birth June 21, 1982

Place of Birth Taipei, Taiwan

Degrees B.S.A.S. Cum Laude, Electrical Engineering, May 2006
D.Sc. Electrical and Systems Engineering, May 2012

Professional Societies The IEEE Information Theory Society
American Association for the Advancement of Science

Publications

P.-H Lai and J. A. O'Sullivan, "Toward Optimal Trade-off between Identification and Secrecy-Key Binding Using Linear Codes," in *the Proceedings of the 2011 IEEE International Symposium on Information Theory* (ISIT), Saint Petersburg, Russia.

P.-H Lai, T. Roos and J. A. O'Sullivan, "MDL Hierarchical Clustering for Stemmatology" in *Proceedings of the 2010 IEEE International Symposium on Information Theory* (ISIT), Austin, Texas, USA, 2010.

P.-H Lai and J. A. O'Sullivan, "MDL Hierarchical Clustering with Incomplete Data." *Information Theory and Applications Workshop* (ITA), San Diego, USA, 2010.

I. Odinaka, P.-H Lai, A. D. Kaplan, J. A. O'Sullivan, E. J. Sirevaag, S. D. Kristjansson, A. K. Sheffield and J. W. Rohrbaugh, "ECG Biometrics: a Robust Short-Time Frequency Analysis," *in the Proceedings of the 2nd Workshop on Information Forensics and Security* (WIFS), Seattle, Washington, USA, 2010.

I. Odinaka, P.-H Lai, A. D. Kaplan and J. A. O'Sullivan, "On Estimating Biometric Capacity: and Example Based on LDV Biometrics," in *the Proceedings of the 2010 Allerton Conference on Communication, Control and Computing*, Allerton, Illinois, USA.

A. D. Kaplan, J. A. O'Sullivan, E. J. Sirevaag, S. D.Kristjansson, P.-H Lai and J. W. Rohrbaugh, "Hidden State Dynamics in Laser Doppler Vibrometery Measurements of the Carotid Pulse Under Resting Conditions," in *the Proceedings of the 32th Annual International Conference*

of the IEEE Engineering in Medicine and Biology Society, Buenos Aires, Argentina, 2010.

M. Chen, J. A. O'Sullivan, N. Singla, E.J. Sirevaag, S. D. Kristjansson, P.-H Lai, A. D. Kaplan, J. W. Rohrbaugh, "Laser Doppler Vibrometry Measures of Physiological Function: Evaluation of Biometric Capabilities," in *IEEE Transactions on Information Forensics and Security*, vol. 5, number 3, pp. 449 - 460, 2010.

P.-H Lai, J. A. O'Sullivan, and R. Pless," Minimum Description Length and Clustering with Exemplars," in the Proceedings of *the 2009 IEEE International Symposium on Information Theory* (ISIT), Seoul, Korea, 2009.

M. Chen, J. A. O'Sullivan, A. D. Kaplan, P.-H Lai, E. J. Sirevaag, and J. W. Rohrbaugh, "Biometrics with Physical Exercise Using Laser Doppler Vibrometry Measurements of the Carotid Pulse." The *First IEEE International Conference on Biometrics, Identity and Security*, Tampa, Florida, USA, 2009.

P.-H Lai, J. A. O'Sullivan, M. Chen, E. J. Sirevaag, A. D. Kaplan, and J. W. Rohrbaugh, "A Robust Feature Selection Method for Noncontact Biometrics Based on Laser Doppler Vibrometry," Best Paper Award, *The 6th Biometrics Symposium*, Tampa, Florida, USA, 2008.

P.-H Lai and J. A. O'Sullivan, "Pattern Recognition System Design with Linear Encoding for Discrete Patterns," in *the Proceedings of the 2007 IEEE International Symposium on Information Theory* (ISIT), Nice, France, 2007.

J. A. O'Sullivan and P.-H. Lai, "Pattern Recognition System Design Based on LDPC Matrices," in *the Proceedings of the 2005 IEEE International Symposium on Information Theory* (ISIT), Adelaide, Australia, pp. 33-36, 2005.

May 2012