Creating Fingerprint Databases and a Bayesian Approach to Quantify Dependencies in Evidence

Maria Annette Tackett
Brentwood, TN

B.S., University of Tennessee, 2009
M.S., University of Tennessee, 2010

A Dissertation Presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Statistics

University of Virginia
August, 2018

This work was partially funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreement #70NANB15H176 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, University of California Irvine, and University of Virginia.

Abstract

In 2009, the National Research Council issued *Strengthening Forensic Science in the United States: A Path Forward* about the need for more scientific rigor in forensic science. Since then, there has been an effort to make the methods used to analyze forensic evidence more objective, in part through the use of statistics to interpret the forensic evidence. With Lindley (1977) as a guide, this research focuses on two aspects of statistics in forensic science. The first is the creation of large databases that can be used for the development and implementation of statistical methods. We propose a theoretical framework for *fully-resourced* databases that contain sufficient information to be used for these purposes and demonstrate their use in statistical inference, specifically how the databases can be used to systematically obtain prior information in the Bayesian framework. Recommendations are provided for the type of information that can be included in such databases in the context of fingerprint evidence.

The second aspect is quantifying and interpreting the weight of evidence when multiple candidates are examined as the source of a mark recovered from a crime scene. We propose accounting for the dependencies that exist in the weight of evidence for multiple candidates by imposing a constraint on the set of plausible models and examine the asymptotic properties under the constrained model. This research is used to inform guidelines for the examination of multiple candidates identified by a fingerprint matching system such as the Automated Fingerprint Identification System (AFIS).

Acknowledgments

I would like to express appreciation for my advisor, Professor Spitzner for helping me get thus far in my academic journey. He has encouraged me to challenge myself as a researcher and always be willing to explore new ideas.

Thank you to my committee members - Professor Kafadar, Professor Holt, and Dr. Iyer - for their time and willingness to give guidance and support throughout my dissertation research.

Thank you to the faculty of the Statistics Department for their teaching and encouragement over the years. I learned a lot in each of my classes, and that is in large part to the dedication and enthusiasm each professor brought to the classroom. I would also like to thank Karen Dalton for her unending support and encouragement.

I would like to thank the friends and family who have been with me throughout this journey. To my parents who taught me the value of working hard no matter the task and have supported me through each new undertaking throughout my life. To my friends in A10 Northside, for their continued encouragement and support throughout this journey. They have always known how to make me laugh and smile just when I need it most.

I am most grateful for my husband Matt for his unconditional love and support. He has encouraged me throughout this journey, and I am deeply appreciative for his above and beyond patience and thoughtfulness, especially in the final stages of the Ph.D.

Contents

Chapter 1

Introduction

1.1 Motivation

In 2009, the National Research Council issued *Strengthening Forensic Science in the United States: A Path Forward* [51] about the need for more scientific rigor in forensic science. In this report, the council stated the need for research in the various disciplines in forensic sciences, which could lead to a better understanding of the sources of variability in forensic evidence and the potential biases that can occur in analyses. Given that forensic evidence is often used in the courtroom, establishing more scientific rigor in forensics is especially important, because "it has become apparent, over the past decade, that faulty forensic feature comparison has led to numerous miscarriages of justice" [21, pg. 44].

A decision in the case *Daubert v. Merrell Dow Pharmaceuticals* state that the judge should act as the "gatekeeper" in determining if evidence meets criteria for scientific validity and whether evidence "rests on reliable foundation." [21]. In a 2016 report, the President's Council of Advisors on Science and Technology (PCAST) identify criteria for a judge to consider in making a determination in regards to the scientific validity of forensic evidence: "(1) whether the theory or technique can be (and has been) tested; (2) whether the theory or technique has been subjected to peer review and publication; (3) the known or potential rate

of error of a particular scientific technique; (4) the existence and maintenance of standards controlling the technique's operation; and (5) a scientific technique's degree of acceptance within a relevant scientific community" [21, pg. 41]. To achieve these criteria, forensic methods need to become more objective by reducing the amount of human judgment involved in the process. In addition to allowing for a better estimate of the error rate, objective methods tend to be more accurate, repeatable and reliable than subjective ones [21]. One way to achieve more objectivity is by quantifying the evidence that is used to make decisions in different parts of the forensic examination. For example, PCAST suggested one way to make fingerprint analysis more objective is through the development of areas such image analysis.

To invest in the research and development of more objective forensic science techniques, large databases that contain examples of forensic evidence are required. [21] stated that the creation of such databases was "the most important resource to propel the development of objective methods." [21, pg. 11] Similarly, the Organization of Scientific Area Committees (OSAC)[44], included "the identification of types of databases will be needed to support proposed approaches" in a list of needs to encourage more statistical development in feature comparison methods. The importance of databases is a key motivation for our research. While databases could be developed containing various types of forensic evidence, our focus is on databases for pattern evidence - markings produced when as a result of one object coming into contact with another [4]. We propose a theoretical framework for the creation of of databases that contain enough information to be used for statistical inference. We discuss the development of these databases in chapters 2 and 3.

1.2 Statistical Challenges

Suppose we have X, a numerical summary of pattern evidence produced by a known source and Y, a numerical summary of pattern evidence produced by an unknown source. Throughout this document, one of our main objectives will be to quantify an assessment of whether X and Y were produced by the same source. We state this inquiry in the following

hypotheses:

$$\begin{aligned} H_0 &: X \text{and } Y \text{ were produced by the same source} \\ H_1 &: X \text{and } Y \text{ were produced by the different sources} \end{aligned} \tag{1.1}$$

Given H_0, H_1, and $E = \{X, Y\}$, we use [33] to quantify the strength of all information in support of H_0.

$$\frac{P(H_0|E)}{P(H_1|E)} = \frac{P(H_0)}{P(H_1)} \times \frac{P(E|H_0)}{P(E|H_1)} \tag{1.2}$$

The prior odds, $\frac{P(H_0)}{P(H_1)}$, calculated primarily using non-fingerprint evidence. Therefore, in this document, we are most interested in analyzing the Bayes factor in (6.2).

$$\frac{P(E|H_0)}{P(E|H_1)} = \frac{\int P(E|H_0, \phi)\pi(\phi|H_0)d\phi}{\int P(E|H_1, \phi)\pi(\phi|H_1)d\phi} \tag{1.3}$$

One major challenge in calculating (6.2) is obtaining $\pi(\phi|H_i)$, the prior information for the parameters under hypothesis H_i. We propose developing a sufficiently resourced database that could be used to obtain this prior information using previous cases that are similar to the current one. This database is discussed in more detail in Chapter 2.

Other statistical methods, such as [40] and [41], have used databases that contain a large amount of information about different individuals that can be used to calculate $P(E|H_1)$. For example, the Federal Bureau of Investigation's Next Generation Identification database contains the fingerprints and other information for millions of individuals [23]. The challenge, then, is in the availability of the data needed to calculate $P(E|H_0)$. Databases typically do not contain replicates of evidence (e.g. as multiple fingerprints) produced by a single source (e.g. individual finger). Moreover, it is often difficult to produce prints to represent all of the possible scenarios that could occur in a crime scene. There are some databases created under experimental conditions that contain multiple prints produced by the same source. These can be used to provide some measure of the variability that exists between multiple prints produced by the same source; however, they do not necessarily cover all conditions that may occur in practice.

In order to quantify the variability between multiple prints and marks produced by the same source, Neumann et al. [41] uses a distortion model to simulate thousands of "pseudo-marks" from a single configuration of minutiae. The pseudo-marks produced by the model are a result of linear or rotational translation of the original marks. These translations simulate the distortions that could occur in the minutiae configuration from various motions, such as sliding or turning, that could occur when a mark is produced under uncontrolled conditions. The model is only meant to account for distortions that occur when a finger makes contact with a flat surface, therefore, it is not meant to account for the additional distortion of the minutiae configuration when a mark is left on a curved surface, such as a door knob.

The distortion model is based on the thin-plate splines of Bookstein [8]. In [8], the translation of each point is captured in a bending-energy matrix. Neumann et al. [40] creates sets of these bending-energy matrices for the different types of translations by training the bending-energy matrices on a database of fingerprints from the University of Lausanne which contains 3520 fingerprints consisting of 704 images each of 25 unique fingers. This leads to one of the challenges with using the model in practice. Without availability to such a data set as the one used by Neumann et al. [40], it is very difficult to replicate such a distortion model.

Another challenge is that in Neumann et al. [40] and Neumann et al. [41], the distortion model is only used to simulate prints needed to calculate $P(E|H_0)$ but it is not used to simulate prints for the calculation of $P(E|H_1)$, thus giving unequal treatment to the numerator and denominator of the likelihood ratio in Neumann et al. [41]. Given the large number of prints that can be used to calculate $P(E|H_1)$, however, applying a distortion model to each print could make calculating the probability computationally intensive. We propose that the criteria for databases we present in chapters 2 and 4 could help alleviate these challenges, because they would require databases have the information necessary to calculate $P(E|H_0)$ without the use of simulated data.

A challenge arises in the interpretation of the Bayes factor in (6.2) when more than one candidate is examined as the source of a fingermark. Typically the calculation of the Bayes

factor does not account for other candidates being examined, therefore the evidence from each candidate is interpreted independently [6, 40, 41]. These independent interpretations could cause one to overstate the weight of evidence in support of a match, because they do not account for the additional uncertainty that arises when new candidates are introduced.

One suggestion to minimize this issue is for examiners to use different thresholds for making a conclusion from the evidence depending on the number of candidates that are considered [16]. While this could provide some structure that could be used for interpreting evidence, without criteria on how to change such thresholds, it is still relatively subjective. We propose taking advantage of Bayesian methods by placing a constraint on the model set up that reflects that at most one unique candidate can be the source of a fingermark. By placing this constraint on the models, we can can capture the dependencies in the evidence for multiple candidates when we calculate the weight of evidence. This approach eliminates the need for creating decision thresholds that are based on the number of candidates, since the impact of multiple candidates will be accounted for in the Bayes factor. This is discussed in more detail in Chapter 6.

1.3 Terminology

We intend for the framework we propose to be applicable for multiple types of forensic evidence such as tool marks, shoe prints, and bullet casings; however, the focus on this document is on the analysis of fingerprint evidence. With that in mind, we present terminology that will be used throughout the remainder of the document.

A *fingermark* (*mark*) is the impression produced in a uncontrolled environment such as a crime scene, and a *fingerprint* (*print*) is an impression produced in a controlled environment such as a lab or police station. This terminology was first used in the 1892 text *Fingerprints* by Sir Francis Galton who is credited with developing the first method to quantify fingerprint identification [49]. Since then, it has been the terminology commonly used by British examiners [40]; forensic practitioners in the United States use the terms latent print

and exemplar print to refer to a mark and print, respectively. The *individual* refers to the person who produced the mark, and the *source* refers to the specific finger that produced a mark. Therefore, there can be up to 10 different sources for a given individual. Figure 1-1 shows a fingermark and fingerprint that were determined by the FBI to have been produced by the same source.

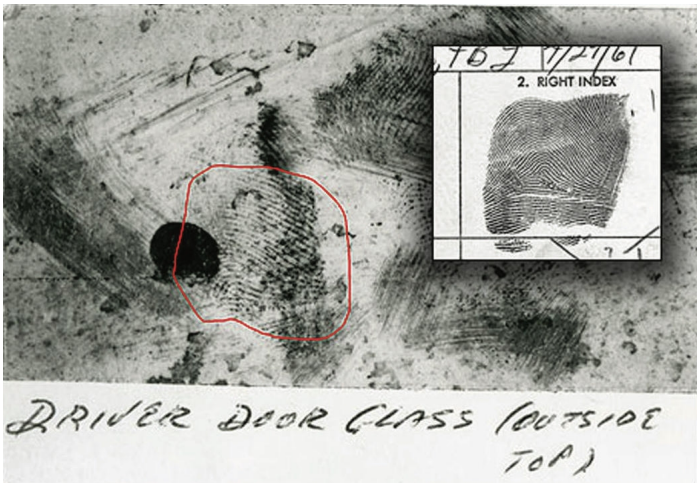

Figure 1-1: A fingermark recovered from a 1969 murder victim's car and the fingerprint of a suspect derived from the FBI's fingerprint database. It was eventually determined that the mark and print were produced by the same individual [25].

Figure 1-2: Types of minutia - ridge ending and bifurcation

A *friction ridge* (*ridge*) is the raised part of the skin on a finger, and a *furrow* is the

depression between two ridges. In most fingerprint images, such as the images in Figure 1-1, the ridges appear dark and the furrows appear light. The ridges create the general pattern that is often used to compare a mark and print. At the end of each ridge are *minutiae*, features that are often analyzed for comparisons during the forensic examination process. (This process is described in more detail in Chapter 4.) There are numerous types of minutiae; however, we focus on two types of minutiae that occur most often - *ridge ending*, the point at which an ridge ends, and a *bifurcation*, the point at which a ridge splits into two different ridges. An example of each type of minutia is shown in Figure 1-2. Fingerprints can be classified based on their overall ridge pattern - *arch*, *loop*, and *whorl* (shown in Figure 1-3). This classification is often used as a first step in comparison. While some methods use patterns to assess evidence, we focus on those based on an examination of minutiae.

Figure 1-3: Arch, loop and whorl are three categories of patterns that can be found in a fingerprint image. There are many variations within each category. For example, the loop shown above is a right slant loop, since the ridges flow to the right side of the impression. A variation on this loop pattern is a left slant loop in which the ridges flow to the left side of the impression. (Image from [28])

1.4 Generating Data From Fingerprint Images

We focus on how data summarized from pattern evidence can be used for statistical analysis, not how the numerical summaries are generated. Therefore, we intend for our framework to be used with any technique that generates a numerical summary from pattern evidence. We present some methods that are used to generate vectors of data from fingerprint images to help the reader have a more complete picture of the variety of data that could be

used in this framework. The data generated from [41] consists of three different numerical summaries from minutiae - *shape*, *direction*, and *type*. We describe each of these below.

A configuration of k minutiae is identified (usually by an examiner), and the center of the configuration is calculated using the arithmetic mean of the Cartesian coordinates of the minutiae. This is shown in Figure 1-4. Then, k triangles are created by connecting each minutiae to the centroid and adjacent minutiae (shown in Figure 1-4). The fingerprint is then summarized in a vector containing information about the shape, direction and type from these k minutiae.

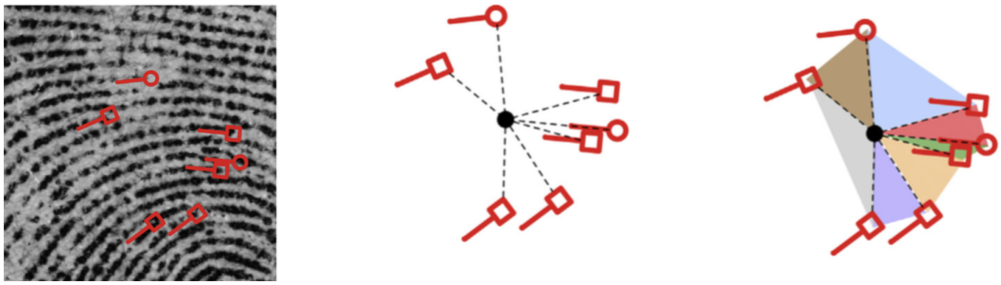

Figure 1-4: Left: Identifying a group of k minutiae in a fingerprint image. Middle: Identifying the centroid in the group of minutiae. Right: Creating k triangles by connecting each minutia to the centroid and to adjacent minutiae. (Image from [41])

Once the minutiae configuration is established, the *shape* is calculated for each of the k triangles. The shape includes two elements - the form factor and the aspect ratio. The *form factor* is the ratio between the area of the triangle and its perimeter, and the *aspect ratio* is the ratio between the diameters of the circumcircle and incircle in the triangle (shown in Figure 1-5).

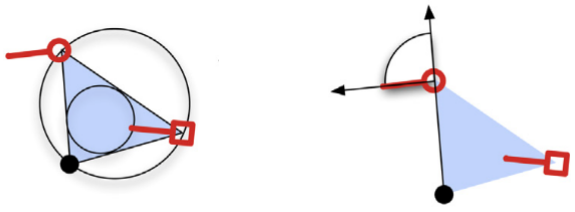

Figure 1-5: Left: Shape - Aspect ratio for a single triangle. Right: Direction - Angle measured counterclockwise from the axis to the direction of the minutia. (Image from [41])

To establish the orientation of the configuration of k triangles, the aspect ratio is used to determine the first triangle. The first triangle is the one that has the minimum aspect ratio; the rest of the triangles are numbered consecutively going counterclockwise. Once the triangles are numbered, the aspect ratio data is removed and the shape data consists only of the form factor. Using empirical evidence, they determine there is some dependency between the form factor of a given triangle and its adjacent triangles; however there is no dependency otherwise. Therefore, the distribution of shape of minutiae configurations from a single source can be modeled using a univariate Gaussian density for the first triangle and bivariate Gaussian densities for the subsequent triangles. Kernel density estimation is used to model the distribution of shape data from multiple sources.

The direction is calculated for each of the k minutiae. It is the angle measured counterclockwise from an axis to the direction of the minutia (shown in Figure 1-5). The axis extends from the centroid to the minutiae location. Using empirical evidence, the authors assume independence between the directions for the minutiae, which simplifies the density functions. Based on the histogram of density estimates for direction, they determine that the distribution is skewed to the right under the null and alternative hypotheses. Therefore, they approximate the distributions of the minutiae directions using non-parametric distributions based on von Mises kernels.

The final piece of data derived from the k minutiae is the type of each minutia using a similar method as [40]. The type is a categorical variable that can be one of the following categories: *ridge ending*, *bifurcation*, *unknown*. The authors assume that the type of each minutia is dependent on the location of the minutiae in the overall ridge flow pattern but not by other minutiae. The densities for the distribution of type depend on the type of minutiae and an examiner's ability to correctly identify it. They establish a table of probabilities based on a survey of 200 latent print examiners in which each examiner was asked to identify the type of a series of minutiae on fingermarks. When analyzing a mark, the probability can be understood as the probability the minutia is type l given an examiner marked it as m. They assume there is no uncertainty in an examiner's determination of minutia type when examining a fingerprint.

In addition to the types of numerical summaries described in [41], another category of summaries are based on data from the Automated Fingerprint Identification System (AFIS). These are "black box" systems that produce scores indicating how closely fingerprint images in the database match an image of a fingermark that was input into the system. The AFIS systems are proprietary, so it is not clear exactly how these scores are calculated; however, they can be used to give an indication about how closely each candidate print matches the mark in question. Some statistical methods, such as those in [19] and [18], and [6] use the AFIS scores to calculate the weight of evidence in support of H_0 from (1.1).

1.5 Outline of the Dissertation

In Chapter 2, we propose a framework for an ideal database that contains enough information to sufficiently be used for statistical analysis. We also discuss the inference that can be conducted in the less ideal databases that are more often found currently. Many numerical summaries of fingerprint evidence are in multivariate vectors, so in Chapter 3, we extend our framework to the multivariate case and show its applicability for a wider variety of data.

Once we have provided a foundation for the databases, we apply them more specifically to the context of fingerprint evidence. In Chapter 4, we discuss the potential sources of variability that could exist in multiple prints produced by the same source. We describe the fingerprint collection process in both controlled and uncontrolled conditions, and use this to provide guidelines for the types of factors to consider for each "instrument" represented in a database.

In Chapter 5, we apply the results from chapters 2 - 4 to a demonstration of how the database could be used in practice. We apply the shape metric from [40] use the information from the database to obtain prior information regarding whether two prints being compared were produced by the same source. This demonstration helps us gain insight in regards to the design of the database along with some potential advantages of our proposed framework.

Once we have fully described our proposed database, we move on the question of quantifying the weight of evidence from (1.2). Using the investigation of Brandon Mayfield and Ouhnane Daoud in the 2004 Madrid terrorist attack [42] as a guide, we discuss the dependencies that exist in the interpretation of the weight of evidence when multiple candidates are examined as the potential source of a fingermark. We extend this analysis from two candidates to the scenario when many candidates are examined and use the results to provide some general guidelines to consider when examining a list of candidates, such as those produced by AFIS.

We conclude the dissertation in Chapter 7 with a summary of the work presented in this document and suggestions for future research directions.

Chapter 2

Foundations for a Fingerprint Database: Univariate Measurements

2.1 Motivation

In a 2016 report, the President's Council of Advisors on Science and Technology (PCAST) recommended steps that could be taken to improve the scientific validity of forensic science. The report was written in response to a 2015 inquiry by President Obama about the progress that had been made to address the criticisms posed in the 2009 report by the National Research Council [51] and the areas for improvement that remained. One of the recommendations made by PCAST was the development of databases that could be used by researchers for methodological development. In regards to the importance of databases, [21] states,

> *The most important resource to propel the development of objective methods would be the creation of huge databases containing known prints, each with many corresponding "simulated" latent prints of varying qualities and completeness, which would be made available to scientifically-trained researchers in academia and industry.* [pg. 103]

Similarly, PCAST included the "development of forensic feature databases, with adequate privacy protections, that can be used in research" [21, pg. 130] in its list of recommendations to create a national research and development strategy for forensic sciences. While this recommendation refers specifically to simulated latent prints (fingermarks), we assume, if unrealistic, a scenario of being able to obtain multiple fingermarks from the same source.

In this chapter, we will establish a theoretical foundation to describe the way databases can be used in forensic research. Though our present focus is on fingerprint analysis, we intend to propose a framework that can be more fully applied to multiple forensic contexts, most especially to other types of pattern evidence. Additionally, our focus is not on the ways fingerprint images are processed to produce data that can be used for statistical analysis, rather we focus on how that data can be used for statistical inference. Therefore, we intend for this framework to work in conjunction with any method that is used to generate data from fingerprint images, such as those described in section 1.4.

We begin the chapter by describing the inference problem and the role databases have in helping us solve that problem in section 2.2. Then in section 2.3, we describe how databases can be used to obtain prior information using a technique inspired by empirical Bayes methods. We offer an alternative approach for obtaining prior information based on variance decomposition in section 2.4, and in section 2.5, we demonstrate our proposed methods using simulated data. The focus of this chapter is in inference for univariate data; we extend the ideas from this chapter to the multivariate case in Chapter 3.

2.2 General Setup

Suppose there is a fingermark recovered from a crime scene and a fingerprint produced by a person of interest. We call Y_1 the numerical summary from the mark that was recovered from the scene using *Instrument* Y and X_1 the numerical summary derived from the print produced using *Instrument* X in a controlled environment such as a lab or police station. The "instrument", then, refers to a combination of personnel, equipment, and investigation

techniques used to recover a fingermark or produce a fingerprint. We will define "instruments" more specifically in Chapter 4. Ultimately, the goal of our analysis mimics that of a latent print examiner - to establish if X_1 and Y_1 were produced by the same source. Let θ_X be the error-free numerical summary from the fingerprint and θ_Y the error-free numerical summary from the fingermark. Using equation (1.2), we quantify the weight of evidence in support of H_0 based on the following hypotheses:

$$\begin{aligned} H_0 &: \theta_X = \theta_Y \\ H_1 &: \theta_X \neq \theta_Y \end{aligned} \tag{2.1}$$

As with any tool used for measuring, *Instrument X* and *Instrument Y* have some measurement error. Additionally, because *Instrument Y* captures marks produced under uncontrolled conditions at a scene, we expect that the measurement error in *Instrument Y* is larger than the error in *Instrument X*. Understanding the differences in the measurement error for each instrument can help us better distinguish between discrepancies due to measurement error and those that indicate that X_1 and Y_1 were produced by different sources. One of our primary objectives, therefore, is to use our inferential framework to quantify the measurement error for each instrument.

Databases can help us in this endeavor in a key way. An ideal database would contain many measurements from previous investigations that were produced by the instrument. These measurements could help us quantify the prior information $\pi(\phi|H_i)$ under hypothesis H_i from (1.3). Though this information could be determined by forensic experts, using databases to collect the information is a more systematic approach. Moreover, only one fingermark and fingerprint are investigated, i.e. $n_X = n_Y = 1$; therefore, any prior information will have a strong impact on the Bayes factor.

Ideally, one could use a database to obtain information about measurement error for the instrument in question, measurement error across a variety of instruments similar to the one in question, and information about the marks that are typically found in cases with

circumstances similar to the one being investigated, in order to find the prior information about the parameters of $P(E|H_0)$ and $P(E|H_1)$, where $E = \{X_1, Y_1\}$. If all of these pieces of information can be derived from the database, we call it *fully-resourced*; otherwise, if certain pieces of information are missing, we refer to it as *sub-resourced*. A sub-resourced database does not have enough information about the variability in measurements error across instruments similar to the ones in the current investigation. Not having this information about variability could cause the weight of evidence for a match to be overstated.

Current databases do not often contain the type of information required for a fully-resourced database in the proposed framework. Some databases, such as those used in the Fingerprint Verification Competitions [13], include repeated impressions produced by the same source; however, all of the impressions were produced under controlled conditions. Moreover, we will define "instrument" to include specific characteristics about the surface on which an impression was made. (This is discussed in detail in Chapter 4.) In many current databases, since the impressions are made under controlled conditions, there is little to no variation in the surfaces represented in the database. Given these limitations, many current databases do not have sufficient information to define the "similar instruments" required for the fully-resourced case.

2.2.1 Description of Databases

Let S_X be the set of instruments in the database that are similar to *Instrument X* and S_Y be the set of instruments in the database similar to *Instrument Y*. In the ideal, fully-resourced database, S_X and S_Y consist of many instruments; however, in the sub-resourced case, S_X and S_Y contain only the instruments in the current investigation, i.e. $S_X = \{X\}$ and $S_Y = \{Y\}$. For each instrument S, we have a collection of measurements $U_{S,ij}$. Let $i = 1, \ldots, g_S$, such that g_S is the total number of sources measured by instrument S, and $j = 1, \ldots, r_{S,i}$ such that $r_{S,i}$ is the total number of replicated measurements of source i measured by instrument S. Finally, let $n_S = \sum_i r_{S,i}$ be the total number of pattern measurements in the database produced by instrument S.

2.3 Using a Database for Statistical Inference

2.3.1 Model Setup

To illustrate our framework, we will use a scenario in which X_1 and Y_1 are univariate Gaussian data derived from fingerprint images. Though we are presenting the framework using univariate Gaussian distributions, we aim to generalize the framework beyond this scope. We have chosen to start with a Gaussian structure for the data, because it is applicable for many types of numerical summaries of fingerprint images. Even if the numerical summaries do not follow a Gaussian distribution, the standard values of these summaries can be used in this framework. At this point, we assume independence between elements in the vector of numerical data; in Chapter 3, we extend the framework to the more general multivariate Gaussian case where the correlation structure can be derived completely from the information in the database.

When X_1 and Y_1 are produced by the same source, $X_1 \sim G(\theta, \sigma_X^2)$ and $Y_1 \sim G(\theta, \sigma_Y^2)$, where θ is the true measurement of X_1 and Y_1, and σ_X^2 and σ_Y^2 are the variability in instruments X and Y, respectively. We can think of σ_X^2 and σ_Y^2 as the measurement error of instruments X and Y. The prior knowledge about θ, σ_X^2 and σ_Y^2 can be described as

$$\theta \sim G(\theta_0, \sigma_0^2) \qquad \sigma_X^2 \sim \text{logN}(\mu_{X,0}, \tau_{X,0}^2) \qquad \sigma_Y^2 \sim \text{logN}(\mu_{Y,0}, \tau_{Y,0}^2) \tag{2.2}$$

where $\mu_{X,0}$ and $\mu_{Y,0}$ are the means of the respective log Normal distributions, and $\tau_{X,0}^2$ and $\tau_{Y,0}^2$ are the variances. "A commonly used prior distribution for variance parameters is inv-$\chi_\kappa^2(\lambda)$, the scaled inverse chi-squared distribution with κ degrees of freedom and a scale parameter λ. This distribution is often chosen, because it is the conjugate prior distribution for the variance parameter in the Gaussian likelihood function. Since the prior distribution of the θ's does not depend on σ_X^2 or σ_Y^2, we do not have a conjugate model and therefore gain no conjugacy benefits from using the scaled inverse chi-square distribution. "Moreover, we found using this distribution was not computationally feasible. Therefore, we use the lognormal

(as the approach in [7]) to make the computations more manageable as we implement the framework.

In contrast, when X_1 and Y_1 are produced by different sources, $X_1 \sim G(\theta_X, \sigma_X^2)$ and $Y_1 \sim G(\theta_Y, \sigma_Y^2)$, where θ_X and θ_Y are the true measurements of X_1 and Y_1, respectively, and σ_X^2 and σ_Y^2 are defined as before. The prior information about θ_X, θ_Y, σ_X^2 and σ_Y^2 can be described similarly as in (2.2)

$$\theta_X \sim G(\theta_0, \sigma_0^2) \qquad \theta_Y \sim G(\theta_0, \sigma_0^2) \qquad \sigma_X^2 \sim \text{logN}(\mu_{X,0}, \tau_{X,0}^2) \qquad \sigma_Y^2 \sim \text{logN}(\mu_{Y,0}, \tau_{Y,0}^2) \quad (2.3)$$

where the mean and variance parameters for the log-Normal distributions are defined as before.

Using this basic construction, we can obtain prior information about the within instrument variability, σ_X^2 and σ_Y^2, and use this to answer the inferential questions in section 2.2. We begin with the ideal fully-resourced database, then show what can be measured using a sub-resourced database.

2.3.2 Inference Using a Fully-Resourced Database

Let $S_X^0 = S_X - \{X\}$ be the set of instruments in the database similar to *Instrument X* but excluding *Instrument X*. $S_Y^0 = S_Y - \{Y\}$ is defined similarly. Let $U_{S,ij}$, be the individual measurements in database S, and $\mathbf{D}_0 = \{U_{S,ij} : S \in S_X^0 \cup S_Y^0\}$, the complete set of measurements in S_X^0 and S_Y^0. Given (2.3) each $\theta_{S,i} \sim G(\theta_0, \sigma_0^2)$, $\sigma_S^2 \sim \text{logN}(\mu_{X,0}, \tau_{X,0}^2)$ for $S \in S_X$, and $\sigma_S^2 \sim \text{logN}(\mu_{Y,0}, \tau_{Y,0}^2)$ for $S \in S_Y$. We assume the $U_{S,ij}$'s, $\theta_{S,i}$ and the σ_S's are independent for all $i = 1, \ldots, g_S$ and $j = 1, \ldots, r_{S,i}$.

Since our goal is to use the database to obtain prior information, we use $\mathbf{D}_0$ to calculate the parameters for the prior distributions. To do so, we take inspiration from empirical Bayes techniques and obtain prior information by maximizing the likelihood equation, written in general and analytically in (2.4) and (2.5), respectively. Standard techniques for multi-parameter maximization can be used to achieve this. It is worth noting, however, a key

difference between our method and empirical Bayes methods. In the usual empirical Bayes techniques, the data is used to fit the parameters of the prior distribution and that same data is then analyzed. In our method, $\mathbf{D}_0$, the data used to fit the parameters of the prior distributions, does not include the data that is then analyzed as part of the current investigation.

$$
\begin{aligned}
&L(\theta_0, \sigma_0^2, \mu_{X,0}, \tau_{X,0}^2, \mu_{Y,0}, \tau_{Y,0}^2 | \mathbf{D}_0) = \\
&\qquad \prod_{S \in S_X^0} \int \left\{ \pi_X(\sigma_S^2) \prod_{i=1}^{g_S} \int \left\{ \pi(\theta_{S,i}) \prod_{j=1}^{r_{S,i}} \pi(U_{S,ij} | \theta_{S,i}, \sigma_S^2) \right\} d\theta_{S,i} \right\} d\sigma_S^2 \\
&\qquad \times \prod_{S \in S_Y^0} \int \left\{ \pi_Y(\sigma_S^2) \prod_{i=1}^{g_S} \int \left\{ \pi(\theta_{S,i}) \prod_{j=1}^{r_{S,i}} \pi(U_{S,ij} | \theta_{S,i}, \sigma_S^2) \right\} d\theta_{S,i} \right\} d\sigma_S^2
\end{aligned}
\tag{2.4}
$$

$$
\begin{aligned}
&L(\theta_0, \sigma_0^2, \mu_{X,0}, \tau_{X,0}^2, \mu_{Y,0}, \tau_{Y,0}^2 | \mathbf{D}_0) = \\
&\qquad \prod_{S \in S_X^0} \int \tau_{X,0}^{-\frac{1}{2}} (2\pi)^{-\sum_{i=1}^{g_S} (\frac{1}{2} + \frac{r_{S,i}}{2})} (\sigma_S^2)^{-\left(1 + \frac{\sum_{i=1}^{g_S} r_{S,i} - g_S}{2}\right)} \prod_{i=1}^{g_S} (\sigma_S^2 + r_{S,i}\sigma_0^2)^{-\frac{1}{2}} \\
&\qquad \times \exp\left\{ -\frac{1}{2} \left[\frac{1}{\tau_{X,0}^2} (\log(\sigma_S^2) - \mu_{X,0})^2 + \sum_{i=1}^{g_S} \left(\frac{\theta_0}{\sigma_0^2} + \frac{\sum_{j=1}^{r_{S,i}} U_{S,ij}^2}{\sigma_S^2} - \frac{(\theta_0 \sigma_S^2 + \sigma_0 \sum_{j=1}^{r_{S,i}} U_{S,ij})^2}{(\sigma_S^2 + r_{S,i}\sigma_0^2)^2} \right) \right] \right\} \\
&\qquad \prod_{S \in S_Y^0} \int \tau_{X,0}^{-\frac{1}{2}} (2\pi)^{-(\frac{1}{2} + \sum_{i=1}^{g_S} \frac{r_{S,i}}{2})} (\sigma_S^2)^{-\left(1 + \frac{\sum_{i=1}^{g_S} r_{S,i} - g_S}{2}\right)} \prod_{i=1}^{g_S} (\sigma_S^2 + r_{S,i}\sigma_0^2)^{-\frac{1}{2}} \\
&\qquad \times \exp\left\{ -\frac{1}{2} \left[\frac{1}{\tau_{Y,0}^2} (\log(\sigma_S^2) - \mu_{Y,0})^2 + \sum_{i=1}^{g_S} \left(\frac{\theta_0^2}{\sigma_0^2} + \frac{\sum_{j=1}^{r_{S,i}} U_{S,ij}^2}{\sigma_S^2} - \frac{(\theta_0 \sigma_S^2 + \sigma_0 \sum_{j=1}^{r_{S,i}} U_{S,ij})^2}{(\sigma_S^2 + r_{S,i}\sigma_0^2)^2} \right) \right] \right\}
\end{aligned}
\tag{2.5}
$$

Once we have completely specified the prior distribution using $\mathbf{D}_0$, we can update the densities of σ_X^2 and σ_Y^2 using $\mathbf{D}_X$ and $\mathbf{D}_Y$, the set of measurements collected using *Instrument X* and *Instrument Y*. The density function $p(\sigma_X^2 | \mathbf{D}_X)$ takes form in (2.6); $p(\sigma_Y^2 | \mathbf{D}_Y)$ can be written in a similar fashion.

$$
\begin{aligned}
p(\sigma_X^2|\mathbf{D}_X) &\propto \prod_{i=1}^{g_X} \int \left\{ \pi(\theta_{X,i}) \prod_{j=1}^{r_{X,i}} \pi(U_{X,ij}|\theta_{X,i},\sigma_X^2) \right\} d\theta_{X,i} \\
&\propto \pi_X(\sigma_X^2) \tau_{X,0}^{-\frac{1}{2}} (2\pi)^{-\sum_i \frac{r_{X,i}+1}{2}} (\sigma_X^2)^{-\sum_i \frac{r_{X,i}-g_X}{2}} \left[\prod_{i=1}^{g_X} \left(\frac{\sigma_X^2 + r_{X,i}\sigma_0^2}{\sigma_0^2 \sigma_X^2} \right)^{-\frac{1}{2}} \right] \\
&\times \exp\left\{ -\frac{1}{2} \left[\frac{(\log(\sigma_X^2) - \mu_{X,0})^2}{\tau_{X,0}^2} + \sum_{i=1}^{g_X} \left(\frac{\theta_0^2}{\sigma_0^2} + \frac{\sum_{j=1}^{r_{X,i}} U_{X,ij}^2}{\sigma_X^2} - \frac{(\theta_0 \sigma_X^2 + \sigma_0^2 \sum_{j=1}^{r_{X,i}} U_{X,ij})^2}{(r_{X,i}\sigma_0^2 + \sigma_X^2)^2} \right) \right] \right\}
\end{aligned}
\tag{2.6}
$$

We can now use the prior information obtained from the database to calculate the Bayes factor. Let M_0 be the model under H_0 (X_1 and Y_1 from the same source), and M_1 the model under H_1. The Bayes factor is

$$
\begin{aligned}
BF_{01}(X_1, Y_1) &= \frac{\pi_{M_0}(X_1, Y_1)}{\pi_{M_1}(X_1, Y_1)} \\
&= \frac{\int\int\int \pi(X_1|\theta,\sigma_X^2)\pi(Y_1|\theta,\sigma_Y^2)\pi(\theta)p(\sigma_X^2|\mathbf{D}_X)p(\sigma_Y^2|\mathbf{D}_Y)d\theta d\sigma_X^2 d\sigma_Y^2}{\int\int\int\int \pi(X_1|\theta_X,\sigma_X^2)\pi(Y_1|\theta_Y,\sigma_Y^2)\pi(\theta_X)\pi(\theta_Y)p(\sigma_X^2|\mathbf{D}_X)p(\sigma_Y^2|\mathbf{D}_Y)d\theta_X d\theta_Y d\sigma_X^2 d\sigma_Y^2} \\
&= \frac{\int\int A\, p(\sigma_X^2|\mathbf{D}_X)p(\sigma_Y^2|\mathbf{D}_Y)d\sigma_X^2 d\sigma_Y^2}{\int\int B\, p(\sigma_X^2|\mathbf{D}_X)p(\sigma_Y^2|\mathbf{D}_Y)d\sigma_X^2 d\sigma_Y^2}
\end{aligned}
\tag{2.7}
$$

such that

$$
A = (\sigma_Y^2\sigma_0^2 + \sigma_X^2\sigma_0^2 + \sigma_X^2\sigma_Y^2)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \left[\frac{X_1^2}{\sigma_X^2} + \frac{Y_1^2}{\sigma_Y^2} + \frac{\theta_0^2}{\sigma_0^2} - \frac{(X_1\sigma_Y^2\sigma_0^2 + Y_1\sigma_X^2\sigma_0^2 + \theta_0\sigma_X^2\sigma_Y^2)^2}{\sigma_0^2\sigma_X^2\sigma_Y^2(\sigma_Y^2\sigma_0^2 + \sigma_X^2\sigma_0^2 + \sigma_X^2\sigma_Y^2)} \right] \right\}
$$

$$
B = \left(\frac{1}{2\pi}\right)(\sigma_0^2+\sigma_X^2)^{-\frac{1}{2}}(\sigma_0^2+\sigma_Y^2)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \left[\frac{X_1^2}{\sigma_X^2} + \frac{Y_1^2}{\sigma_Y^2} + \frac{\theta_0^2}{\sigma_0^2} - \frac{(X_1\sigma_0^2 + \theta_0\sigma_X^2)^2}{\sigma_0^2\sigma_X^2(\sigma_0^2+\sigma_X^2)} - \frac{(Y_1\sigma_0^2 + \theta_0\sigma_Y^2)^2}{\sigma_0^2\sigma_Y^2(\sigma_0^2+\sigma_Y^2)} \right] \right\}
$$

"The scale in Table 2.1 is a guide for interpreting the Bayes factor to assess the original question - what is the weight of evidence from X_1 and Y_1 in support of H_0? This scale was originally presented by Kass and Raftery [32] and is often used as a standard for interpreting Bayes factors. Given the wealth of information in the fully-resourced database, this

assessment can be made having a better understanding of the variability we expect from the instruments similar to those in the current investigation. Currently, such fully-resourced databases are rare, so we move to the sub-resourced case which better reflects the information that is often available in current practice.

$\mathbf{2log(B_{01})}$	**Evidence in support of $\mathbf{H_0}$**
0 to 2	Neutral
2 to 6	Positive
6 to 10	Strong
> 10	Very strong

Table 2.1: Interpretation of Bayes factors based on [32]. The negative of this scale can be interpreted similarly as evidence in support of H_1.

2.3.3 Inference Using a Sub-Resourced Database

In a sub-resourced database, $S_X = \{X\}$ and $S_Y = \{Y\}$, therefore, we do not have $\mathbf{D}_0$ from which we can obtain prior information. Thus, we are unable to get information about the distribution of σ_X^2 and σ_Y^2. The best we can do is calculate estimates of the variability within each instrument in the current investigation, $\hat{\sigma}_X^2$ and $\hat{\sigma}_Y^2$, but there is no sense of the uncertainty associated with these estimates. Using a similar approach as before, we maximize the likelihood function in (2.8) (shown analytically in (2.9)) using multi-parameter maximization techniques to calculate the desired parameters and estimates.

$$\begin{aligned} L(\theta_0, \sigma_0^2, \sigma_X^2, \sigma_Y^2 | \mathbf{D}_X, \mathbf{D}_Y) = & \prod_{i=1}^{g_X} \int \left\{ \pi(\theta_{X,i}) \prod_{j=1}^{r_{X,i}} \pi(U_{X,ij} | \theta_{X,i}, \sigma_X^2) \right\} d\theta_{X,i} \\ & \times \prod_{i=1}^{g_Y} \int \left\{ \pi(\theta_{Y,i}) \prod_{j=1}^{r_{Y,i}} \pi(U_{Y,ij} | \theta_{Y,i}, \sigma_Y^2) \right\} d\theta_{Y,i} \end{aligned} \tag{2.8}$$

$$
\begin{aligned}
&L(\theta_0, \sigma_0^2, \sigma_X^2, \sigma_Y^2 | \mathbf{D}_X, \mathbf{D}_Y) = \\
&\quad (2\pi)^{-\sum_i \frac{r_{X,i}}{2}} (\sigma_X^2)^{-\sum_i \frac{r_{X,i}-g_X}{2}} \left[\prod_{i=1}^{g_X} \left(\frac{\sigma_X^2 + r_{X,i}\sigma_0^2}{\sigma_0^2 \sigma_X^2} \right)^{-\frac{1}{2}} \right] \\
&\quad \times \exp\left\{ -\frac{1}{2} \sum_{i=1}^{g_X} \left(\frac{\theta_0^2}{\sigma_0^2} + \frac{\sum_{j=1}^{r_{X,i}} U_{X,ij}^2}{\sigma_X^2} - \frac{(\theta_0 \sigma_X^2 + \sigma_0^2 \sum_{j=1}^{r_{X,i}} U_{X,ij})^2}{\sigma_0^2 \sigma_X^2 (r_{X,i}\sigma_0^2 + \sigma_X^2)} \right) \right\} \\
&\quad \times (2\pi)^{-\sum_i \frac{r_{Y,i}}{2}} (\sigma_Y^2)^{-\sum_i \frac{r_{Y,i}-g_Y}{2}} \left[\prod_{i=1}^{g_Y} \left(\frac{\sigma_Y^2 + r_{Y,i}\sigma_0^2}{\sigma_0^2 \sigma_Y^2} \right)^{-\frac{1}{2}} \right] \\
&\quad \times \exp\left\{ -\frac{1}{2} \sum_{i=1}^{g_Y} \left(\frac{\theta_0^2}{\sigma_0^2} + \frac{\sum_{j=1}^{r_{Y,i}} U_{Y,ij}^2}{\sigma_Y^2} - \frac{(\theta_0 \sigma_Y^2 + \sigma_0^2 \sum_{j=1}^{r_{Y,i}} U_{Y,ij})^2}{\sigma_0^2 \sigma_Y^2 (r_{Y,i}\sigma_0^2 + \sigma_Y^2)} \right) \right\}
\end{aligned} \tag{2.9}
$$

We can then calculate the Bayes factor in (2.10) and interpret it using Table 2.1.

$$
BF_{01}(X_1, Y_1) = \frac{\int \pi(X_1|\theta, \hat{\sigma}_X^2)\pi(Y_1|\theta, \hat{\sigma}_Y^2)\pi(\theta)d\theta}{\int\int \pi(X_1|\theta_X, \hat{\sigma}_X^2)\pi(Y_1|\theta_Y, \hat{\sigma}_Y^2)\pi(\theta_X)\pi(\theta_Y)d\theta_X d\theta_Y} \tag{2.10}
$$

$$
\begin{aligned}
BF_{01}(X_1, Y_1) = &\left(\frac{(\sigma_0^2 + \hat{\sigma}_X^2)(\sigma_0^2 + \hat{\sigma}_Y^2)}{(\hat{\sigma}_Y^2 \sigma_0^2 + \hat{\sigma}_X^2 \sigma_0^2 + \hat{\sigma}_X^2 \hat{\sigma}_Y^2)} \right)^{\frac{1}{2}} \\
&\times \exp\left\{ -\frac{1}{2} \left[\frac{(\sigma_0^2 X_1 + \hat{\sigma}_X^2 \theta_0)^2}{\sigma_0^2 \hat{\sigma}_X^2 (\sigma_0^2 + \hat{\sigma}_X^2)} + \frac{(\sigma_0^2 Y_1 + \hat{\sigma}_Y^2 \theta_0)^2}{\sigma_0^2 \hat{\sigma}_Y^2 (\sigma_0^2 + \hat{\sigma}_Y^2)} \right.\right. \\
&\left.\left. - \frac{(\sigma_0^2 \hat{\sigma}_Y^2 X_1 + \sigma_0^2 \hat{\sigma}_X^2 Y_1 + \hat{\sigma}_X^2 \hat{\sigma}_Y^2 \theta_0)^2}{\sigma_0^2 \hat{\sigma}_X^2 \hat{\sigma}_Y^2 (\sigma_0^2 \hat{\sigma}_Y^2 + \sigma_0^2 \hat{\sigma}_X^2 + \hat{\sigma}_X^2 \hat{\sigma}_Y^2)} - \frac{\theta_0^2}{\sigma_0^2} \right] \right\}
\end{aligned}
$$

The Bayes factor calculated from (2.10) should be interpreted with some caution. Though we are able to obtain estimates of the measurement error for each instrument, we do not have a full understanding of the variability in that estimate. This may be especially important when trying to evaluate close non-matching prints in which it is more difficult to determine the discrepancies that indicate the prints are from different sources rather than discrepancies due to usual measurement error.

2.4 Variance Decomposition

In the previous section, we described a method of obtaining information from a database based on maximizing the appropriate likelihood function. There are a relatively small number of parameters being fitted in the univariate Gaussian case, thus optimization can be done relatively smoothly. This is not necessarily the case when we move to more complex data structures, such as the multivariate Gaussian case in Chapter 3. Therefore, in this section, we propose an alternative procedure for gathering information about parameters of interest from databases that is based on Analysis of Variance (ANOVA). Once we obtain the parameters, we can calculate the Bayes factors and any available prior information as in the previous section.

Though we are utilizing an ANOVA decomposition to quantify sources of variation, we are not interested in testing as is common in traditional ANOVA analyses. Thus, we are not as concerned that our data will not meet some of the usual assumptions such as equal variance across groups. We begin with the ANOVA decomposition for a sub-resourced database followed by the fully-resourced case.

2.4.1 Sub-Resourced Database

As we discussed in the previous section, a sub-resourced database does not contain measurements from instruments similar to instruments X and Y. Therefore, we only have the data from S_X and S_Y to obtain the desired parameters and estimates of variability. We can calculate θ_0, prior parameter for the source means by simply taking the mean of all measurements in S_X and S_Y; however, calculating the variance parameters will require that we break down the different components of variation within measurements by the same instrument. For a given instrument S, we can decompose the variance in the collection of measurements produced S using Table 2.2. Using the same notation as before, g_S is the number of sources measured by S, $r_{S,i}$ is the number of measurements for the i^{th} source, and n_S is the total number of measurements produced by S.

	DF	SS	MS
Between	$g_S - 1$	$\sum_i r_{S,i}(\bar{U}_{S,i} - \bar{U}_{S,..})^2$	$SSB/(g_S - 1)$
Within	$n_S - g_S$	$\sum_i \sum_j (U_{S,ij} - \bar{U}_{S,i})^2$	$SSW/(n_S - g_S)$
Total	$n_S - 1$	$\sum_i \sum_j (U_{S,ij} - \bar{U}_{S,..})^2$	

Table 2.2: ANOVA decomposition for a single instrument, such that $i = 1, \ldots, g_S$, $j = 1, \ldots, r_{S,i}$, and $n_S = \sum_i r_{S,i}$ is the total number of measurements for instrument S. SS represents the sum of squares and MS is the mean square.

Using this decomposition, we can calculate $\hat{\sigma}_S^2$, the estimated variability in the measurements produced by S, using the classical result that the sample variance can be estimated using the mean square error. Therefore, we estimate the measurement error for instruments X and Y using MSW, the within mean square.

To fit σ_0^2, we need to take into account measurements from both S_X and S_Y. Thus we calculate σ_S^2 using (2.11) which is an estimate of the variation between sources in databases S_X and S_Y..

$$\sigma_0^2 = \frac{SSB_X + SSB_Y}{n_X + n_Y - 2} \tag{2.11}$$

2.4.2 Fully-Resourced Database

Using a fully-resourced database, we can utilize information from $\mathbf{D}_0 = \{U_{S,ij} : S \in S_X^0 \cup S_Y^0\}$ to calculate the desired parameters as before. Similar to the sub-resourced case, we can calculate θ_0 using the sample mean across all measurements in a database. To calculate the parameters associated with the sources of variability, we use ANOVA to decompose the variability observed in a single database S_d as shown in Table 2.3. Now, SSB measures the sum of squares between instruments in a database S_d and SSW measures the sum of squares within a single instrument.

"To calculate σ_0^2 in (2.12), the variation between sources in $\mathbf{D}_0$ is pooled (similar to (2.11)).

	DF	SS	MS
Between	$N_{S_d} - 1$	$\sum_s n_S(\bar{U}_s - \bar{U}_.)^2$	$SSB/(N_{S_d} - 1)$
Within	$\sum_s n_S - N_{S_d}$	$\sum_s \sum_i (U_{s,i} - \bar{U}_{s,.})^2$	$SSW/(\sum_s n_S - N_{S_d})$
Total	$\sum_s n_S - 1$	$\sum_s \sum_i (U_{s,i} - \bar{U}_.)^2$	

Table 2.3: ANOVA decomposition for a database. $s = 1, \ldots, N_{S_d}$ is the number of instruments in the database, and $i = 1, \ldots, n_S$ is the number of measurements for instrument S.

$$\sigma_0^2 = \frac{\sum_{i=1}^{N_{S_d}} SSB_i}{\sum_{i=1}^{N_{S_d}} (n_{S_i} - 1)} \tag{2.12}$$

To complete the prior information that is obtained from a fully-resourced database, we find the parameters ($\mu_{X,0}$, $\tau^2_{X,0}$) and ($\mu_{Y,0}$,$\tau^2_{Y,0}$) that are used to derive prior information about σ^2_X and σ^2_Y, respectively. We write the equations in terms of *Instrument X*; the same structure is used for *Instrument Y*.

Our aim is to derive the prior distribution of within instrument variation for the instruments in S^0_X, so we use the decomposition in Table 2.2 to estimate the within instrument variation. Given the i^{th} instrument in S^0_X, $S^2_i = \frac{SSW_i}{n_{S_i}-1}$ and

$$\mu_{X,0} = \text{mean}\{\log(S_1), \ldots, \log(S_N)\} \qquad \tau^2_{X,0} = \text{variance}\{\log(S_1), \ldots, \log(S_N)\} \tag{2.13}$$

such that N is the number of instruments in S^0_X.

2.5 Numerical Demonstration

One of the main objectives of the following simulation is to understand how the methods perform under different configurations of the proposed database. We are most interested in

how the number of unique sources in the database, the number of replications for each source, and the number of similar instruments (in the fully-resourced case) affect the performance of our framework in fitting the desired parameters to obtain prior information. Other aspects to be explored through these simulations is the general performance of our method and the computational feasibility of the approach if it were to be put into practice. We also want to compare the performance of the optimization-based approach to the one based on variance decomposition. We begin by looking at the sub-resourced case in section 2.5.1 and then move onto the the fully resourced case in section 2.5.2.

2.5.1 Sub-Resourced Database

To test the performance of the sub-resourced database methods described in sections 2.3 and 2.4, we simulate databases S_X and S_Y using the following parameters: $\theta_0 = 10$, $\sigma_0^2 = 1000$, $\sigma_X^2 = 9$, and $\sigma_Y^2 = 100$. Different settings are tested for the number of sources in each database,g_S, and the number of repeated measurements for each source, $r_{S,.}$. There were four different levels for the number of unique sources in each database: 100,500,1000,5000. These levels were chosen based on some of the database sizes used by [40], [41], [12], and others. There were three different levels used for the number of repeated measurements for each source: 5, 20, and 70. Five repeated measurements was chosen based on [12] who did not see much different in results in their analysis of the effects of database size using five repeated measurements versus larger numbers of repeated measurements. Moreover, for metrics (such as an AFIS score) which produce one numerical summary value for an entire fingerprint, it will be difficult to obtain large numbers of repeated measurements for each source. Therefore, having a framework that can work well with few repeated measurements will be more feasible to implement in practice. We also test having 70 repeated measurements for fingerprints in database S_X, since examiners typically work around 70 - 80 minutiae on a fingerprint. For similar reasons, we tested 20 repeated measurements for the fingermarks in S_Y. In these scenarios, the metric used to translate fingerprint images into numerical data would need to have a separate measurement for each minutiae, such as the metrics in [41].

For each case, the simulation results were generated using the following process:

1. Databases S_X and S_Y were simulated using the source and replication specifications for that case. The observations were generate using the form described in section 2.2 and the following parameters: $\theta_0 = 10$, $\sigma_0^2 = 1000$, $\sigma_X^2 = 9$, and $\sigma_Y^2 = 100$.

2. The parameters θ_0 and σ_0^2 were fit using each proposed method. The fitted values are θ_{0*} and σ_{0*}^2 in the results tables. Similarly, σ_X^2 and σ_Y^2 were estimated using each proposed method; the estimated values are $\hat{\sigma}_X^2$ and $\hat{\sigma}_Y^2$ in the results tables. Nelder-Mead, a simplex optimization method, was used to find the fitted and estimated values for the maximization approach. This represents how the historical data from the databases could be used in practice.

The simulation results are shown in tables 2.4, 2.5, and 2.6. In general, both approaches perform very well. Both methods have the worst parameter fits in Case 4, when both the number of unique sources and the number of repeated measurements for each source are small. Additionally, the optimization method has some sensitivity to the starting value, as shown by the uneven level of performance among the seven cases explored. In each case, multiple starting values were used and the parameters corresponding the overall maximized likelihood function are shown in the results. Using this multiple starting values made this approach very slow compared to the variance decomposition method. Other maximization techniques could be used in order to improve the performance of the optimization method; however, the high level of performance and very fast run time of the variance decomposition method make it more desirable to implement in practice.

2.5.2 Fully-Resourced Database

To understand how our proposed methods perform for a fully-resourced database, we test different configurations of the number of unique sources, g_S, and the number of instruments in each database, N_S. The levels 100, 1000, and 5000 are used to assess the effect of the number of unique sources on the performance, and the levels 4, 10, and 20 are used to assess

Case	g_X	g_Y	r_X	r_Y	θ_0	θ_{0*}: Opt.	θ_{0*}: Decomp.
1	5000	5000	5	5	10	10.431	10.435
2	500	500	5	5	10	10.848	10.828
3	500	500	70	20	10	7.667	7.121
4	100	100	5	5	10	13.177	13.163
5	100	100	70	20	10	13.151	13.249
6	1000	500	5	5	10	9.181	9.185
7	1000	500	70	20	10	10.562	10.953

Table 2.4: Simulation results fitting θ_0 using a sub-resourced database. "Opt" results are based on the likelihood maximization method in section 2.3.3; "Decomp" results are based on the variance decomposition method in section 2.4.1.

Case	g_X	g_Y	r_X	r_Y	σ_0^2	σ_{0*}^2: Opt.	σ_{0*}^2: Decomp.
1	5000	5000	5	5	1000	990.623	1001.326
2	500	500	5	5	1000	949.570	961.294
3	500	500	70	20	1000	1029.078	978.985
4	100	100	5	5	1000	1029.248	1040.854
5	100	100	70	20	1000	1002.615	1029.498
6	1000	500	5	5	1000	916.095	1221.777
7	1000	500	70	20	1000	988.770	1125.080

Table 2.5: Simulation results fitting σ_0^2 using a sub-resourced database. "Opt." results are based on the likelihood maximization method in section 2.3.3; "Decomp." results are based on the variance decomposition method in section 2.4.1.

the effect on performance based on number of instruments. We are especially interested in how the number of instruments affects the performance of our methods, as this will provide guidance when we define "instrument" in Chapter 4. For each case, we used 5 repeated measurements, since the results from the sub-resourced case suggested this did not have as much of an effect on the performance as the number of unique sources. Moreover, as shown in Table 2.10, the optimization approach runs very slowly, so having a large number of repeated measurements would make this approach not practical to be used in real analysis.

For each case, the simulation results were generated using the following process:

1. Databases S_X^0 and S_Y^0 were simulated using the source and replication specifications for that case. The observations were generated using the form described in section 2.2 and the following parameters: $\mu X, 0 = \log(9)$, $\tau_{X,0}^2 = \log(10)$, $\mu_{Y,0} = \log(100)$, $\tau_{X,0}^2 = \log(10)$.

Case	g_X	g_Y	r_X	r_Y	σ_X^2	$\hat{\sigma}_X^2$: Opt	$\hat{\sigma}_X^2$: Decomp.	σ_Y^2	$\hat{\sigma}_Y^2$: Opt	σ_Y^2: Decomp
1	5000	5000	5	5	9	8.927	8.926	100	99.551	99.587
2	500	500	5	5	9	8.826	8.826	100	100.668	100.638
3	500	500	70	20	9	8.941	8.942	100	97.880	97.846
4	100	100	5	5	9	8.606	8.605	100	96.137	96.162
5	100	100	70	20	9	9.106	9.106	100	98.481	98.465
6	1000	500	5	5	9	8.839	8.836	100	100.782	100.813
7	1000	500	70	20	9	8.975	8.976	100	100.519	100.537

Table 2.6: Simulation results estimating σ_X^2 using a sub-resourced database. "Optimization" results are based on the likelihood maximization method in section 2.3.3; "Decomposition" results are based on the variance decomposition method in section 2.4.1.

2. Using the data in $S_{X,0}$ and $S_{Y,0}$, the parameters θ_0 σ_0^2, $\mu_{X,0}$,$\tau_{X,0}^2$, $\mu_{Y,0}$,and $\tau_{Y,0}^2$were fit using each proposed method. The fitted values are indicated with an $*$ in tables 2.7, 2.8, and 5.20. Nelder-Mead, a simplex optimization method, was used to find the fitted and estimated values for the maximization approach.

3. New means for the sources in S_X^0 and S_Y^0 are generated using Gaussian distributions and the fitted values θ_{0*} and σ_{0*}^2. Similarly, new error variances for the instruments in S_X^0 and S_Y^0 are generated using a log Normal distribution and the fitted values $\mu_{X,0*}$,$\tau_{X,0*}^2$, $\mu_{Y,0*}$,and $\tau_{Y,0*}^2$.

4. Using each proposed method, the fitted values were obtained as before; these values are θ_0 and σ_0^2, $\mu_{X,0}$,$\tau_{X,0}^2$, $\mu_{Y,0}$, $\tau_{Y,0}^2$ in tables 2.7, 2.8, and 5.20.

5. Comparing the fitted values, then, helps us assess how well the proposed method is describing the patterns in the data.

The following parameter values were used to generate S_X^0 and S_Y^0, the databases of historical data used to generate the parameters: $\theta_0 = 10$, $\sigma_0^2 = 1000$, $\mu_{X,0} = \log(9)$, $\tau_{X,0}^2 = \log(10)$, $\mu_{Y,0} = \log(100)$, $\tau_{Y,0}^2 = \log(10)$. The results are shown in tables 2.7, 2.8, and 5.20. As in the sub-resourced case, we can assess the performance of each approach using by comparing the $*$ values to the fitted values for each parameter (for example, θ_{0*} to θ_0).

In general, the decomposition method tends to perform better than the optimization method. As in the sub-resourced case, Nelder-Mead was used to fit the parameters, which

			Optimization				Decomposition			
Case	N_S	g_S	θ_{0*}	θ_0	σ^2_{0*}	σ^2_0	θ_{0*}	θ_0	σ^2_{0*}	σ^2_0
1	4	100	8.718	8.399	1060.221	1146.218	9.231	8.351	1078.022	1145.506
2	4	1000	9.745	10.937	1011.091	766.882	9.735	9.950	1038.449	1050.251
3	4	5000	9.991	12.451	999.349	607.666	9.987	9.652	1031.675	1074.921
4	10	100	10.458	11.105	1101.185	1115.684	9.964	10.654	1110.004	1223.417
5	10	1000	10.311	1.077	967.277	1427.911	10.304	10.046	1007.157	1056.734
6	10	5000	9.818	7.766	1004.410	975.221	9.812	9.956	1018.829	1026.635
7	20	100	9.438	9.253	978.615	990.765	9.535	9.168	1022.196	1001.942
8	20	1000	10.159	9.787	1009.077	1108.724	10.135	10.013	1036.780	1043.462
9	20	5000	10.141	2.984	1026.941	1003.889	10.130	9.948	1070.711	1119.043

Table 2.7: Simulation results fitting θ_0 and σ^2_0 using a fully-resourced database. "Optimization" results are based on the likelihood maximization method in section 2.3.3; "Decomposition" results are based on the variance decomposition method in section 2.4.1.

			Optimization				Decomposition			
Case	N_S	g_S	$\mu_{X,0*}$	$\mu_{X,0}$	$\tau^2_{X,0*}$	$\tau^2_{X,0}$	$\mu_{X,0*}$	$\mu_{X,0}$	$\tau^2_{X,0*}$	$\tau^2_{X,0}$
1	4.	100	3.218	3.314	2.330	1.598	3.040	2.041	3.087	4.805
2	4	1000	2.288	4.637	68.738	0.289	1.621	1.203	7.312	4.307
3	4	5000	3.690	4.142	4.947	1.115	3.080	2.877	3.270	1.183
4	10	100	1.809	2.748	1.496	1.223	1.318	0.844	3.144	2.460
5	10	1000	21.900	4.729	97.071	2.628	1.815	2.213	3.967	2.655
6	10	5000	3.488	3.077	35.524	0.178	1.311	1.359	2.542	2.886
7	20	100	2.035	2.206	0.811	0.486	1.585	2.034	2.226	2.288
8	20	1000	3.296	3.820	13.665	5.122	2.204	2.112	3.305	3.051
9	20	5000	-3.525	2.480	130.038	2.828	1.998	1.360	2.984	3.609

Table 2.8: Simulation results fitting $\mu_{X,0}$ and $\tau^2_{X,0}$ using a fully-resourced database. "Optimization" results are based on the likelihood maximization method in section 2.3.3; "Decomposition" results are based on the variance decomposition method in section 2.4.1.

has some sensitivity to the starting value. The variance decomposition method has less variability in its performance, which does make it more ideal for implementing in practice. Case 1 which has both a low number of unique sources and a low number of instruments seems to have the worst performance of the tested cases. This indicates that our methods should not be used with very small databases. As expected, the methods perform better when there are larger amounts of available data; however, very large databases are not required. The cases in which there are 10 instruments perform as well as when there are 20; therefore, when creating a database,the most important aspect to take into consideration is the number of unique sources.

			Optimization				Decomposition			
Case	N_S	g_S	$\mu_{Y,0*}$	$\mu_{Y,0}$	$\tau^2_{Y,0*}$	$\tau^2_{Y,0}$	$\mu_{Y,0*}$	$\mu_{Y,0}$	$\tau^2_{Y,0*}$	$\tau^2_{Y,0}$
1	4	100	4.169	4.248	0.912	0.039	3.870	3.423	1.609	1.034
2	4	1000	4.410	0.240	0.973	1.503	4.086	4.533	1.660	0.800
3	4	5000	4.249	0.708	38.052	2.413	4.134	4.825	3.196	2.737
4	10	100	4.676	4.411	1.934	1.824	4.663	4.326	3.939	7.004
5	10	1000	4.333	2.480	1.740	3.617	4.371	4.022	1.749	4.334
6	10	5000	3.170	2.459	194.920	1.692	3.975	3.559	3.673	2.421
7	20	100	4.433	4.711	2.122	1.262	4.150	4.164	3.246	2.584
8	20	1000	4.399	4.019	1.476	1.284	4.106	3.942	2.292	2.536
9	20	5000	7.925	1.498	86.902	0.894	4.946	4.277	3.307	3.895

Table 2.9: Simulation results fitting $\mu_{Y,0}$ and $\tau^2_{Y,0}$ using a fully-resourced database. "Optimization" results are based on the likelihood maximization method in section 2.3.3; "Decomposition" results are based on the variance decomposition method in section 2.4.1.

Finally, Table 2.10 shows the number of minutes required to fit the parameters for each method. The run time for the optimization approach was substantially impacted by the number of sources in each database; however, there is negligible impact on the variance decomposition approach. Given our assessment that having a larger number of sources is desirable in practice, the optimization approach may not be as feasible in practice as the decomposition. To determine the effect the chosen optimization method had on the performance on our approach, we tested cases for the fully-resourced database using Broyden-Fletcher-Goldfar-Shanno (BFGS), a gradient based optimization technique, and simulated annealing, a Markov Chain Monte Carlo optimization technique. In both cases, there was not much improvement in the performance or run time Thus given the results of this analysis, we propose using the ANOVA decomposition based approach in practice, given its level of accuracy and very small run time.

2.6 Conclusion

In this chapter, we provided a theoretical foundation for the information that can be gained from sub- and fully-resourced databases, and we demonstrated this framework with univariate data. From the numerical analysis, we have shown that an ANOVA decomposition

Case	N_S	g_S	Optimization	Decomposition
1	4	100	1.452	0.000
2	4	1000	34.240	0.002
3	4	5000	156.388	0.012
4	10	100	6.749	0.000
5	10	1000	504.490	0.005
6	10	5000	538.277	0.025
7	20	100	24.532	0.001
8	20	1000	131.365	0.010
9	20	5000	547.354	0.061

Table 2.10: Total time (in minutes) to fit parameters for a fully-resourced database.

performs comparable to the optimization approach with far less run time. In order to implement this framework in a realistic scenario, there are two elements that need to be explored further. The first is extending the framework to the multivariate case. While there are some metrics, such AFIS scores, that translate a fingerprint image into univariate data, the majority of data translation methods (such as those discussed in 1.4), translate the image into a multivariate vector. In the next chapter, we extend our framework to the multivariate case, with a focus on an approach that makes use of the Multivariate Analysis of Variance (MANOVA) decomposition. The second is to clearly define an "instrument". In order to fully understand what contributes to this variability and what constitutes an "instrument", we need to understand the potential factors that could contribute to variations in images of fingerprints or fingermarks produced by the same source. In Chapter 4, we describe the fingerprint investigation process in order to understand these potential factors. To make this theoretical framework feasible in the practical realm, we also provide recommendations for the sources of variability that should be included for a database to be considered fully-resourced.

Chapter 3

Foundations for a Fingerprint Database: Multivariate Measurements

In the last chapter, we presented a framework for obtaining prior information from a database using a univariate measurement. We now extend our framework to the multivariate Gaussian case with a focus on the ANOVA-inspired variance decomposition approach. The model is set up in section 3.1; section 3.2, includes the statistical inference from a sub-resourced database; the fully-resourced database is discussed in section 3.3. In section 3.4.1 is a demonstration of the method using simulated data, and we conclude in section 3.5.

3.1 Model Setup

We refer the reader to section 2.2 for a description of the inferential questions and the general structure of databases. We begin here with the data structure, which is the multivariate analog to the model described in section 2.3.1.

Let $\mathbf{X}_1$ be a vector of d_X measurements produced by *Instrument X*. $\mathbf{X}_1 \sim G(\boldsymbol{\theta}_X, \boldsymbol{\Sigma}_X)$, such that $\boldsymbol{\theta}_X$ is a vector of length d_X and $\boldsymbol{\Sigma}_X$ is a $d_X \times d_X$ covariance matrix . Similarly, $\mathbf{Y}_1$ is a set of d_Y measurements from *Instrument Y*, such that $\mathbf{Y}_1 \sim G(\boldsymbol{\theta}_Y, \boldsymbol{\Sigma}_Y)$. We assume

that $d_X = d_Y$, so we drop the subscript and denote the length of each vector using d. When $\mathbf{X}_1$ and $\mathbf{Y}_1$ are from different sources, the source means are generated from

$$\boldsymbol{\theta}_X \sim G(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_0) \qquad \boldsymbol{\theta}_Y \sim G(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_0) \tag{3.1}$$

When $\mathbf{X}_1$ and $\mathbf{Y}_1$ are form the same source, $\boldsymbol{\theta}_X = \boldsymbol{\theta}_Y = \boldsymbol{\theta}$, such that $\boldsymbol{\theta} \sim G(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_0)$.

The prior information for each covariance matrix, $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Sigma}_Y$, is calculated using the separation strategy of [7]. The covariance matrix $\boldsymbol{\Sigma}$ is decomposed into a diagonal matrix of the standard deviations $\mathbf{S}$ and a matrix of the correlations $\mathbf{R}$ as shown in (3.2).

$$\boldsymbol{\Sigma} = \mathbf{SRS} \tag{3.2}$$

Therefore, the prior information for $\boldsymbol{\Sigma}$ is $p(\mathbf{S}, \mathbf{R}) = p(\mathbf{R}|\mathbf{S})p(\mathbf{S})$. We will define $p(\mathbf{S}, \mathbf{R})$ in way such that $p(\mathbf{R}|\mathbf{S}) = p(\mathbf{R})$, i.e. the standard deviations and correlations can be handled independently. This provides more flexibility than the commonly used inverse Wishart distribution which requires a single scale matrix for both the variances and the correlations. Using [7], we will model the prior information for $\mathbf{S}$ using d independent log-Normal distributions, such that S_i, the i^{th} element on the diagonal of $\mathbf{S}$ can be generated from $\text{logN}(\mu_{S_i,0}, \tau^2_{S,0})$; we will use an inverse Wishart prior for $\mathbf{R}$. Let $\mu_{S_i,0}$ be log of the i^{th} diagonal of $\mathbf{S}$ and $\tau^2_{S,0} = Var(S_1, \ldots, S_d)$, the variance of the log of the diagonal elements of $\mathbf{S}$. We discuss how to calculate $\mu_{S_i,0}, \tau^2_{S,0}$ from the database in section 3.3.

3.2 Inference Using a Sub-Resourced Database

Similar to the univariate case, a sub-resourced database does not contain measurements from instruments similar to instruments X and Y, therefore we only have the data from S_X and S_Y, the set of measurements produced by instruments X and Y. We calculate the parameter $\boldsymbol{\theta}_0$ by find the mean of all measurements in S_X and S_Y. To calculate the parameters associated with variance, we will use Multivariate Analysis of Variance (MANOVA), a

	DF	SSCP
Between	$g_S - 1$	$\sum_i r_{S,i}(\bar{\mathbf{U}}_{S,i.} - \bar{\mathbf{U}}_{S,..})(\bar{\mathbf{U}}_{S,i.} - \bar{\mathbf{U}}_{S,..})^T$
Within	$n_S - g_S$	$\sum_i \sum_j (\mathbf{U}_{S,ij} - \bar{\mathbf{U}}_{S,i.})(\mathbf{U}_{S,ij} - \bar{\mathbf{U}}_{S,i.})^T$
Total	$n_S - 1$	$\sum_i \sum_j (\mathbf{U}_{S,ij} - \bar{\mathbf{U}}_{S,..})(\mathbf{U}_{S,ij} - \bar{\mathbf{U}}_{S,..})^T$

Table 3.1: MANOVA decomposition for a single instrument, such that $i = 1, \ldots, g_S$, $j = 1, \ldots, r_{S,i}$, and $n_S = \sum_i r_{S,i}$ is the total number of measurements for instrument S. The sum of squares is now denoted as SSCP to indicate it is a matrix of sum of squares and cross products.

variance decomposition approach similar to the one described in Section 2.4.1.

Let $\boldsymbol{U}_{S,ij}$ be the j^{th} measurement on the i^{th} mark produced by instrument S. Given there are n_S measurements in the database produced by *Instrument S*, we decompose the sources of variability among the n_S measurements as shown in Table 3.1.

The variability within each instrument $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Sigma}_Y$ can be estimated using the analog to the univariate within mean squares as shown below. As in the univariate case, we can only estimate the covariance matrices, $\hat{\boldsymbol{\Sigma}}_X$ and $\hat{\boldsymbol{\Sigma}}_Y$; we have no additional instruments from which we can obtain information about the variability in measurement error from such instruments. Lastly, we obtain $\boldsymbol{\Sigma}_0$, shown below, by pooling the SSCPB for S_X and S_Y.

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}}_X &= \frac{SSCPW_X}{n_X - g_X} \\
\hat{\boldsymbol{\Sigma}}_Y &= \frac{SSCPW_Y}{n_Y - g_Y} \\
\boldsymbol{\Sigma}_0 &= \frac{SSCPB_X + SSCPB_Y}{n_X + n_Y - 2}
\end{aligned}$$

Using our calculated values for $\boldsymbol{\theta}_0$, $\boldsymbol{\Sigma}_0$, $\hat{\boldsymbol{\Sigma}}_X$ and $\hat{\boldsymbol{\Sigma}}_Y$ and $\mathbf{D}_X$ and $\mathbf{D}_Y$, the set of measurements in S_X and S_Y, we can quantify the weight of evidence in support of H_0 using a Bayes factor, as before.

	DF	SSCP
Between	$N_{S_d} - 1$	$\sum_s n_S(\bar{\mathbf{U}}_{s.} - \bar{\mathbf{U}}_{..})(\bar{\mathbf{U}}_{s.} - \bar{\mathbf{U}}_{..})^T$
Within	$\sum_s n_s - N_{S_d}$	$\sum_s \sum_i (\mathbf{U}_{si} - \bar{\mathbf{U}}_{s.})(\mathbf{U}_{si} - \bar{\mathbf{U}}_{s.})^T$
Total	$\sum_s n_s - 1$	$\sum_i \sum_j (\mathbf{U}_{si} - \bar{\mathbf{U}}_{..})(\mathbf{U}_{si} - \bar{\mathbf{U}}_{..})^T$

Table 3.2: ANOVA decomposition for a database. $s = 1, \ldots, N_{S_d}$ is the number of instruments in the database, and $i = 1, \ldots, n_S$ is the number of measurements for instrument S.

$$
\begin{aligned}
BF_{01}(\mathbf{X}_1, \mathbf{Y}_1, \mathbf{D}_X, \mathbf{D}_Y) &= \frac{\int \pi(\mathbf{X}_1|\boldsymbol{\theta}, \hat{\boldsymbol{\Sigma}}_X)\pi(\mathbf{Y}_1|\boldsymbol{\theta}, \hat{\boldsymbol{\Sigma}}_Y)\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int\int \pi(\mathbf{X}_1|\boldsymbol{\theta}_X, \hat{\boldsymbol{\Sigma}}_X)\pi(\mathbf{Y}_1|\boldsymbol{\theta}_Y, \hat{\boldsymbol{\Sigma}}_Y)\pi(\boldsymbol{\theta}_X)\pi(\boldsymbol{\theta}_Y)d\boldsymbol{\theta}_X d\boldsymbol{\theta}_Y} \\
&= \left(\frac{|\boldsymbol{\Sigma}_0||\boldsymbol{\Sigma}_X^{-1} + \boldsymbol{\Sigma}_0^{-1}||\boldsymbol{\Sigma}_Y^{-1} + \boldsymbol{\Sigma}_0^{-1}|}{|\boldsymbol{\Sigma}_X^{-1} + \boldsymbol{\Sigma}_Y^{-1} + \boldsymbol{\Sigma}_0^{-1}|} \right)^{\frac{1}{2}} \\
&\quad \times \exp\left\{ -\frac{1}{2}\Big[\mathbf{B}_X^T \mathbf{A}_X \mathbf{B}_X + \mathbf{B}_Y^T \mathbf{A}_Y \mathbf{B}_Y - \mathbf{B}_0^T \mathbf{A}_0 \mathbf{B}_0 - \theta_0^T \boldsymbol{\Sigma}_0^{-1} \theta_0 \Big] \right\}
\end{aligned}
\tag{3.3}
$$

such that

$$
\begin{aligned}
\mathbf{A}_X &= \boldsymbol{\Sigma}_X^{-1} + \boldsymbol{\Sigma}_0^{-1} & \mathbf{B}_X &= (\boldsymbol{\Sigma}_X^{-1} + \boldsymbol{\Sigma}_0^{-1})^{-1}(\boldsymbol{\Sigma}_X^{-1} X_1 + \boldsymbol{\Sigma}_0^{-1}\theta_0) \\
\mathbf{A}_Y &= \boldsymbol{\Sigma}_Y^{-1} + \boldsymbol{\Sigma}_0^{-1} & \mathbf{B}_Y &= (\boldsymbol{\Sigma}_Y^{-1} + \boldsymbol{\Sigma}_0^{-1})^{-1}(\boldsymbol{\Sigma}_Y^{-1} Y_1 + \boldsymbol{\Sigma}_0^{-1}\theta_0) \\
\mathbf{A}_0 &= \boldsymbol{\Sigma}_X^{-1} + \boldsymbol{\Sigma}_Y^{-1} + \boldsymbol{\Sigma}_0^{-1} & \mathbf{B}_0 &= (\boldsymbol{\Sigma}_X^{-1} + \boldsymbol{\Sigma}_Y^{-1} + \boldsymbol{\Sigma}_0^{-1})^{-1}(\boldsymbol{\Sigma}_X^{-1} X_1 + \boldsymbol{\Sigma}_Y^{-1} Y_1 + \boldsymbol{\Sigma}_0^{-1}\theta_0)
\end{aligned}
$$

3.3 Inference Using a Fully-Resourced Database

We now present the ideal case, in which we have sufficient information in the database to obtain full prior information about $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Sigma}_Y$. We decompose the variation for a database S_d in Table 3.2.

The parameter $\boldsymbol{\theta}_0$ can be calculated finding the mean of all observations in the database

S_d, and $\mathbf{\Sigma}_0$ can be found using the estimates of variability between sources in each database.

$$\mathbf{\Sigma}_0 = \frac{\sum_{i=1}^{N_{S_d}} SSCPB_i}{\sum_{i}^{N_{S_d}} (n_{S_i} - 1)}$$

To find the parameters for the prior distributions of the covariance matrices, we make use of the separation strategy in (3.2). We separate the covariance matrix for within instrument variability in database S_K^0 into $\mathbf{\Sigma}_{S_K^0} = \mathbf{S}_{S_K^0}\mathbf{R}_{S_K^0}\mathbf{S}_{S_K^0}$. The estimate of the covariance matrix can be obtained by pooling together the estimates of the within source variability for each instrument in the database.

$$\hat{\mathbf{\Sigma}}_{S_K^0} = \sum_{i=1}^{N_{S_K^0}} \frac{SSCPW_i}{g_{S_i} - 1} \tag{3.4}$$

Therefore, we calculate $\mu_{S_K^0,i}$ and $\tau_{S_{K^0},0}$ using the following:

$$\mu_{S_K^0,i} = \log(\hat{\sigma}_i) \qquad \tau^2_{S_K^0} = Var\{\log(\hat{\sigma}_i), \ldots, \log(\hat{\sigma}_i)\} \tag{3.5}$$

such that $\hat{\sigma}_i$ is the standard deviation obtained from the i^{th} diagonal of the estimated covariance matrix $\hat{\mathbf{\Sigma}}_{S_K^0}$. To model the prior information for the correlation matrix $\mathbf{R}_K$, we use the inv-Wishart$(\nu_{S_K^0}, \mathbf{\Lambda}^{-1}_{S_K^0})$ with the degrees of freedom, $\nu_{S_K^0}$, and scale matrix $\mathbf{\Lambda}^{-1}_{S_K^0}$ as shown in (3.6).

$$\nu_{S_{K^0}} = N_{S_K^0} - 1 \qquad \mathbf{\Lambda}^{-1}_{S_K^0} = \hat{\boldsymbol{S}}^{-1}_{S_K^0}\hat{\mathbf{\Sigma}}_{S_K^0}\hat{\boldsymbol{S}}^{-1}_{S_K^0} \tag{3.6}$$

Note that the degrees of freedom $\nu_{S_{K^0}}$ must be the at least as large as d, the dimension of the data. Therefore, it in order to conduct inference on high-dimensional data, it may be necessary to define the set of "similar instruments" in a broad way such that many instruments can be included.

$$p(\boldsymbol{\Sigma}_K|\mathbf{D}_K) \propto p(\boldsymbol{S}_K)p(\mathbf{R}_K)\prod_{i=1}^{g_K}\int\left\{\pi(\boldsymbol{\theta}_{K,i})\prod_{j=1}^{r_{K,i}}\pi(\boldsymbol{U}_{K,ij}|\boldsymbol{\theta}_{K,i},\boldsymbol{\Sigma}_K)\right\}d\boldsymbol{\theta}_{K,i} \tag{3.7}$$

Lastly, we can calculate the Bayes factor in (3.8).

$$BF_{01}(\mathbf{X}_1,\mathbf{Y}_1,\mathbf{D}_X,\mathbf{D}_Y) = \frac{\int\int\int\pi(\mathbf{X}_1|\boldsymbol{\theta},\boldsymbol{\Sigma}_X)\pi(\mathbf{Y}_1|\boldsymbol{\theta},\boldsymbol{\Sigma}_Y)\pi(\boldsymbol{\theta})p(\boldsymbol{\Sigma}_X|\mathbf{D}_X)p(\boldsymbol{\Sigma}_Y|\mathbf{D}_Y)d\boldsymbol{\theta}d\boldsymbol{\Sigma}_X d\boldsymbol{\Sigma}_Y}{\int\int\int\int\pi(\mathbf{X}_1|\boldsymbol{\theta}_X,\boldsymbol{\Sigma}_X)\pi(\mathbf{Y}_1|\boldsymbol{\theta}_Y,\boldsymbol{\Sigma}_Y)\pi(\boldsymbol{\theta}_X)\pi(\boldsymbol{\theta}_Y)p(\boldsymbol{\Sigma}_X|\mathbf{D}_X)p(\boldsymbol{\Sigma}_Y|\mathbf{D}_Y)d\boldsymbol{\theta}_X d\boldsymbol{\theta}_Y d\boldsymbol{\Sigma}_X d\boldsymbol{\Sigma}_Y} \tag{3.8}$$

Given the challenges of integrating out $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Sigma}_Y$ from the numerator and denominator, we use [14] to calculate the Bayes factor in (3.8). We describe the process in more detail in the next section.

3.4 Numerical Demonstration

3.4.1 Sub-Resourced Database

We begin by exploring the performance under the sub-resourced database under different specifications of the data in S_X and S_Y. In each example, we use five repeated measurements based on the numerical results from the univariate analysis. Here we are most interested in exploring the performance based on the specifications related to the multivariate structure of the data. In each of the results tables, we have the following data specifications: d is the length of the vector of numerical data, g is the number of sources in the database, and r is the correlation between elements in the numerical vector.

Table 3.3 shows the performance of our approach using the sub-resourced database to fit the parameters $\boldsymbol{\theta}_0$ and $\boldsymbol{\Sigma}_0$ and estimate $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Sigma}_Y$. Under each database configuration, the table shows $d(actual, fitted)$ the distance between the actual and fitted values. The last

column of the table shows the deviance of the log-likelihood function calculated $2(\log L_{actual} - \log L_{fitted})$. From the results, we see that regardless of the dimension of the data or the correlation structure, the framework performs better when there are more unique sources in the database. This result is not surprising given the results from the univariate case in Chapter 2.

				d(actual,fitted)				Deviance
Case	d	g	r	$\boldsymbol{\theta}_0$	$\boldsymbol{\Sigma}_0$	$\boldsymbol{\Sigma}_X$	$\boldsymbol{\Sigma}_Y$	$\log L$
1	5	100	0.2	12.226	1013.120	9.420	101.297	-47.805
2	5	100	0.5	8.870	1101.285	9.215	71.862	-44.996
3	5	100	0.8	13.740	1118.346	20.866	260.886	-56.557
4	5	1000	0.2	2.358	567.104	3.979	26.928	-52.581
5	5	1000	0.5	1.109	530.954	2.203	39.640	-38.375
6	5	1000	0.8	0.838	713.652	1.594	19.050	-31.919
7	5	10000	0.2	0.637	177.205	1.126	10.748	-40.559
8	5	10000	0.5	1.422	178.802	1.128	10.700	-55.422
9	5	10000	0.8	0.942	352.502	2.952	7.006	-39.081
10	10	100	0.2	18.046	5080.907	40.660	410.284	-182.116
11	10	100	0.5	21.096	4556.417	32.070	474.462	-181.906
12	10	100	0.8	30.302	6578.181	93.076	667.223	-192.205
13	10	1000	0.2	6.155	2096.153	7.999	123.934	-173.619
14	10	1000	0.5	12.594	1893.803	18.970	161.531	-199.556
15	10	1000	0.8	7.624	1195.027	25.787	198.277	-159.033
16	10	10000	0.2	2.072	858.992	3.737	32.362	-147.218
17	10	10000	0.5	2.416	689.080	5.039	52.913	-151.675
18	10	10000	0.8	1.509	419.314	2.276	63.811	-161.800
19	20	100	0.2	29.481	23867.231	148.958	1682.057	-685.193
20	20	100	0.5	31.727	20074.769	142.960	1794.017	-659.323
21	20	100	0.8	25.910	17739.539	96.308	6214.074	-687.663
22	20	1000	0.2	11.390	8956.786	43.629	563.018	-651.184
23	20	1000	0.5	9.479	7344.097	47.411	610.242	-616.865
24	20	1000	0.8	8.069	5481.657	72.474	1030.002	-661.339
25	20	10000	0.2	2.595	2216.803	16.656	149.442	-587.042
26	20	10000	0.5	2.036	2017.998	26.277	220.360	-562.199
27	20	10000	0.8	2.241	7982.408	35.021	197.442	-627.206

Table 3.3: Results from the sub-resourced database. d is the length of the vector of numerical data, g is the number of sources in the database, and r is the correlation between elements in the numerical vector. $d(actual, fitted)$ is the distance between the actual and fitted values, and the deviance of the likelihood is calculated $2(\log L_{actual} - \log L_{fitted})$.

3.4.2 Fully-Resourced Database

We now look at the performance of our method in a fully resourced database. The results are show in tables 5.16, 3.5, 3.6, and 3.7. For each of the cases, we keep the number of sources constant ($g = 1000$) so we can explore different cases of the number of instruments in databases S_X^0 and S_Y^0 shown under the column N in the tables. In general, the parameter values are fit reasonably well. In table 5.16, we see some large values for the distance of the $\mathbf{\Sigma}_0$ in some scenarios. These large values are due to instances of large error for one or a small few of the fitted values. In general, there does not seem to be a clear advantage of using a database with a lot of similar instruments N verses a database with a few, so there will be some flexibility when we recommend a database structure in Chapter 4.

Since the value of ν is required to be at least as large as the dimension of the data, we used the maximum of $(d, N-1)$ to calculate these values. Given this calculation of the parameter, it is not surprising that the distance between the actual and predicted values of ν in tables 3.5 and 3.6 are 0 given the actual value of ν used to generate the correlation structure for each instrument.

3.4.3 Sensitivity Analysis

Now that we have assessed the performance of our approach for data that follows a multivariate Gaussian distribution, we now consider its robustness by examining its performance for data that do not meet the underlying assumptions. To do so, we conduct the analysis using a fully-resourced database that contains data simulated from multivariate t distributions. In general, the multivariate t distribution has more variability and heavier tails than the multivariate Gaussian distribution, especially for small degrees of freedom. This analysis will provide an indication of the performance of our approach when the data follow a potentially heavier-tailed distribution like multivariate t.

The results of the sensitivity analysis are shown in Table 3.8. In each case, the database is created using simulated data that follow different specifications of the number of instruments,

number of sources, and degrees of freedom used to specify the multivariate t distribution. In every case, the covariance matrix for the multivariate t distribution is specified such that the correlation between each of the d components is 0.5 the standard deviation is two. Additionally, we assume three replicates for each source in the database. The data are simulated using the using the `rmvt` function in the `mvtnorm` R package. Once we have simulated the data, we use the variance decomposition approach from Section 3.3 to fit the parameters that specify the prior distributions of the model parameters.

We use the fitted parameters to simulate data under the model specifications described in Section 3.1. We then compare the distribution of the original data, call it C_1, and the distribution of the data simulated using the fitted parameters, call it C_2. We use a non-parametric rank-order test by Munzel and Brunner [37] to determine whether there is a statistically significant difference between the two distributions. If we determine there is no statistically significant difference between the two distributions, then we can conclude that our approach is robust to some violations of the multivariate Gaussian assumption under the specifications of the relevant test scenario.

Comparing Multivariate Distributions

Since we are using [37] to compare the distributions C_1 and C_2, we describe the test in terms of two treatment levels. We want to test the hypotheses $H_0 : C_1 = C_2$ vs. $H_1 : C_1 \neq C_2$.

Let $\mathbf{Y}^* = \{\mathbf{Y}_{11}, \ldots, \mathbf{Y}_{1n_1}, \mathbf{Y}_{21}, \ldots \mathbf{Y}_{2n_2}\}$ be the set of all observations in the original and newly simulated data sets. Each $\mathbf{Y}_{ij} = \{Y_{ij,1}, \ldots, Y_{ij,d}\}$, such that $i = 1, 2$ specifies the distribution, $j = 1, \ldots, n_i$ is the observation number, and d is the dimension of the data. Additionally, let Z be a $(n_1 + n_2)$-length vector containing values $\{1, 2\}$ that indicate which distribution an observation is from.

For each of the d components of $\mathbf{Y}^*$, we rank the observations and store the ranks in $\mathbf{R}$, a $d \times (n_1 + n_2)$ matrix. The ranks for the elements in a single $\mathbf{Y}_{ij}$ are stored in $\mathbf{R}_{ij} = (R_{ij}^{(1)}, \ldots, R_{ij}^{(d)})$, a vector of length d. Using $\mathbf{R}$, we define the following SSCP matrices

$$
\begin{aligned}
H &= \sum_{i=1}^{2} (\bar{\mathbf{R}}_{i.} - \tilde{\mathbf{R}}_{..})(\bar{\mathbf{R}}_{i.} - \tilde{\mathbf{R}}_{..})^T \\
G &= \sum_{i=1}^{2} \left(1 - \frac{n_i}{n_1 + n_2}\right) \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\mathbf{R}_{ij} - \bar{\mathbf{R}}_{i.})(\mathbf{R}_{ij} - \bar{\mathbf{R}}_{i.})^T
\end{aligned}
\tag{3.9}
$$

such that $\bar{\mathbf{R}}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{R}_{ij}$ and $\tilde{\mathbf{R}}_{..} = \frac{1}{n_1+n_2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} \mathbf{R}_{ij}$

We calculate the test statistic $T_A = \frac{\text{tr}(H)}{\text{tr}(G)}$ such that T_A follows the F distribution with estimated degrees of freedom $\hat{f}_1$ and $\hat{f}_2$ shown in (3.10). The p-value is calculated as $P(F_{\hat{f}_1, \hat{f}_2} \geq T_A)$.

$$
\hat{f}_1 = \frac{\text{tr}(G)^2}{\text{tr}(G^2)} \qquad \hat{f}_2 = \frac{4}{\sum_{i=1}^{2} \frac{1}{n_i - 1}} \hat{f}_1 \tag{3.10}
$$

Results

Cases 1-36 in Table 3.8 show the model comparisons under different specifications of the database containing data simulated from a multivariate t distribution, and Cases 37 - 42 show the performance when the data was simulated from a multivariate Gaussian distribution. In each case, we compare the distributions C_1 and C_2 by randomly selecting the same 90 observations from each distribution. This value was chosen to remove the effect of sample size in calculating and interpreting the p-value (Case 1 consists of 90 observations).

The 25^{th} percentile (Q_1), median (Q_2), 75^{th} (Q_3), and middle 95% of values p-values are shown in Table 3.8. To obtain the summary statistics in the table, we conduct 100 iterations of simulating data from C_2, randomly selecting 90 observations from C_1 and C_2, and calculating the test statistic and p-values in [37].

The results from this analysis provide some insights about the performance of our approach. Overall, if the number of observations of the database is small, the distribution specified by the fitted parameters does not adequately describe the structure of the data, even when the data follow

a multivariate Gaussian distribution. We see this in the cases in which there are 3 instruments and 10 sources (cases 1, 7, 13, 19, 25, 31, 37) and when there are 10 instruments and 10 sources (cases 4, 10, 16, 22, 28, 34, 40). Moreover, for each specification of the multivariate t distribution, the fit was generally the worst when the parameters were fit using a database with ten instruments and ten sources in each instrument.

The distributions specified by the fitted parameters describes the data well when the database is sufficiently large, even when there are large departures from the multivariate Gaussian assumption, such as when $df = 1$. However, when there are large departures from the Gaussian distribution, the results from the model should still be interpreted with caution. Finally, from the table we see that the number of sources measured by each instrument has a larger effect on the model performance than the number of instruments. Thus, when creating a database, it is preferable to combine instruments that have very few sources in a thoughtful way, since it is preferable to have a large number of sources measured by each instrument than a large number of instruments. We will discuss database design more in Chapter 4.

3.5 Conclusion

In this chapter, we extended the theoretical framework for a database to the context of multivariate Gaussian measurements. Similar to the univariate scenario, we demonstrated how such a database could be used for statistical inference and obtaining information that can be used to quantify the weight of evidence in support of H_0 from (1.2). Now that we have a foundation for the databases, we apply the framework in the context of fingerprint evidence. In the next chapter, we explore the possible sources of variability between multiple prints and marks produced by the same source with the goal of providing recommendations for what constitutes a fully resourced database. We also provide guidelines in regards to defining "instrument" in the database.

				d(actual,estimate)	
Case	d	N	r	θ_0	$\mathbf{\Sigma}_0$
1	3	3	0.2	2.518	30742.364
2	3	3	0.5	7.980	1257.725
3	3	3	0.8	7.406	2984.550
4	3	10	0.2	25.345	25977.259
5	3	10	0.5	8.739	2655.473
6	3	10	0.8	1.276	422.001
7	3	15	0.2	0.934	291.097
8	3	15	0.5	3.581	680.675
9	3	15	0.8	1.772	622.490
10	8	3	0.2	7.665	3417.851
11	8	3	0.5	19.573	10108.923
12	8	3	0.8	21.586	61966.657
13	8	10	0.2	28.519	29978.134
14	8	10	0.5	2.228	8673.962
15	8	10	0.8	17.078	51569.327
16	8	15	0.2	2.670	2424.598
17	8	15	0.5	5.680	8407.143
18	8	15	0.8	76.674	1592057.886
19	15	3	0.2	43.176	49286.872
20	15	3	0.5	128.719	401749.520
21	15	3	0.8	46.644	73367.167
22	15	10	0.2	117.603	1338501.253
23	15	10	0.5	20.124	578507.443
24	15	10	0.8	10.838	560874.050
25	15	15	0.2	20.007	149864.841
26	15	15	0.5	16.507	268414.414
27	15	15	0.8	2.988	51575.923

Table 3.4: $d(actual, fitted)$: Distance between the actual and fitted values of $\boldsymbol{\theta}_0$ and $\mathbf{\Sigma}_0$ using a fully-resourced database.

				d(actual,estimated)			
Case	d	N	r	$\boldsymbol{\mu}_{X,0}$	$\tau^2_{X,0}$	$\nu_{X,0}$	$\boldsymbol{\Lambda}_{X,0}$
1	3	3	0.2	2.839	0.402	0.000	0.283
2	3	3	0.5	0.694	1.336	0.000	0.318
3	3	3	0.8	0.605	0.399	0.000	0.632
4	3	10	0.2	0.449	0.539	0.000	0.143
5	3	10	0.5	0.404	0.418	0.000	0.395
6	3	10	0.8	0.600	0.609	0.000	0.672
7	3	15	0.2	0.665	0.299	0.000	0.085
8	3	15	0.5	0.418	0.457	0.000	0.468
9	3	15	0.8	1.099	0.416	0.000	1.001
10	8	3	0.2	1.897	3.412	0.000	2.135
11	8	3	0.5	4.306	11.567	0.000	7.749
12	8	3	0.8	0.981	1.870	0.000	8.175
13	8	10	0.2	1.153	2.543	0.000	3.148
14	8	10	0.5	1.922	2.611	0.000	3.171
15	8	10	0.8	1.523	1.989	0.000	7.460
16	8	15	0.2	2.517	2.072	0.000	2.405
17	8	15	0.5	1.437	1.681	0.000	5.328
18	8	15	0.8	1.437	1.441	0.000	5.718
19	15	3	0.2	3.759	4.072	0.000	5.848
20	15	3	0.5	4.951	8.757	0.000	15.192
21	15	3	0.8	6.393	8.265	0.000	27.573
22	15	10	0.2	2.249	4.762	0.000	7.015
23	15	10	0.5	3.079	3.169	0.000	17.812
24	15	10	0.8	4.535	4.533	0.000	23.458
25	15	15	0.2	3.499	3.813	0.000	6.546
26	15	15	0.5	2.158	2.056	0.000	14.080
27	15	15	0.8	3.282	3.770	0.000	26.426

Table 3.5: $d(actual, fitted)$: Distance between the actual and fitted values for the prior parameters for Σ_X.

				d(actual,estimate)			
Case	d	N	r	$\boldsymbol{\mu}_{Y,0}$	$\tau^2_{Y,0}$	$\nu_{Y,0}$	$\boldsymbol{\Lambda}_{Y,0}$
1	3	3	0.2	5.990	1.936	0.000	1.076
2	3	3	0.5	2.136	0.843	0.000	0.834
3	3	3	0.8	3.082	4.964	0.000	1.510
4	3	10	0.2	3.578	1.184	0.000	0.370
5	3	10	0.5	3.607	1.586	0.000	0.789
6	3	10	0.8	2.889	1.087	0.000	1.170
7	3	15	0.2	3.123	4.080	0.000	0.354
8	3	15	0.5	2.898	0.875	0.000	0.545
9	3	15	0.8	2.785	0.893	0.000	0.954
10	8	3	0.2	9.409	4.722	0.000	14.991
11	8	3	0.5	8.494	10.068	0.000	10.156
12	8	3	0.8	7.071	3.120	0.000	6.911
13	8	10	0.2	8.120	5.448	0.000	5.614
14	8	10	0.5	9.118	2.916	0.000	9.907
15	8	10	0.8	7.069	4.932	0.000	10.911
16	8	15	0.2	9.433	2.509	0.000	4.750
17	8	15	0.5	7.816	1.707	0.000	6.662
18	8	15	0.8	7.086	3.414	0.000	10.411
19	15	3	0.2	7.846	15.247	0.000	69.067
20	15	3	0.5	13.440	24.765	0.000	94.938
21	15	3	0.8	9.812	19.598	0.000	112.371
22	15	10	0.2	9.335	34.487	0.000	153.512
23	15	10	0.5	11.319	24.755	0.000	127.036
24	15	10	0.8	14.485	38.530	0.000	117.333
25	15	15	0.2	16.261	21.004	0.000	66.304
26	15	15	0.5	13.001	17.219	0.000	60.357
27	15	15	0.8	7.542	13.481	0.000	84.451

Table 3.6: $d(actual, fitted)$: Distance between the actual and fitted values for the prior parameters for Σ_Y.

				Deviance	
Case	d	N	r	$\mathbf{\Sigma}_X$	$\mathbf{\Sigma}_Y$
1	3	3	0.2	13.701	22.721
2	3	3	0.5	-2.409	5.234
3	3	3	0.8	16.540	14.720
4	3	10	0.2	0.309	2.047
5	3	10	0.5	0.125	-1.927
6	3	10	0.8	17.040	6.582
7	3	15	0.2	18.762	36.184
8	3	15	0.5	0.577	1.507
9	3	15	0.8	-2.463	-0.428
10	8	3	0.2	49.007	30.583
11	8	3	0.5	60.029	144.468
12	8	3	0.8	3.797	34.512
13	8	10	0.2	342.969	399.808
14	8	10	0.5	77.051	80.123
15	8	10	0.8	-32.335	-1.594
16	8	15	0.2	8.732	18.350
17	8	15	0.5	-6.662	7.392
18	8	15	0.8	24.993	46.269
19	15	3	0.2	33.163	186.498
20	15	3	0.5	317.985	684.693
21	15	3	0.8	246.651	579.534
22	15	10	0.2	993.380	1475.957
23	15	10	0.5	1279.637	1658.334
24	15	10	0.8	695.483	1328.097
25	15	15	0.2	548.787	724.326
26	15	15	0.5	863.752	971.863
27	15	15	0.8	274.616	543.093

Table 3.7: Deviance in the likelihood calculated as $2(\log L_{actual} - \log L_{fitted})$.

				p-values			
Case	N	g	df	Q_1	Q_2	Q_3	Middle 95%
1	3	10	1	0.001	0.024	0.134	(0.000, 0.607)
2	3	100	1	0.070	0.259	0.484	(0.001, 0.958)
3	3	1000	1	0.080	0.370	0.612	(0.003, 0.935)
4	10	10	1	0.000	0.001	0.029	(0.000, 0.603)
5	10	100	1	0.085	0.251	0.581	(0.005, 0.944)
6	10	1000	1	0.080	0.219	0.528	(0.009, 0.942)
7	3	10	10	0.000	0.035	0.249	(0.000, 0.703)
8	3	100	10	0.198	0.344	0.653	(0.008, 0.927)
9	3	1000	10	0.206	0.435	0.739	(0.018, 0.979)
10	10	10	10	0.000	0.000	0.016	(0.000, 0.221)
11	10	100	10	0.088	0.300	0.558	(0.005, 0.886)
12	10	1000	10	0.225	0.478	0.744	(0.019, 0.925)
13	3	10	100	0.002	0.022	0.235	(0.000, 0.924)
14	3	100	100	0.127	0.322	0.567	(0.005, 0.931)
15	3	1000	100	0.152	0.375	0.646	(0.043, 0.946)
16	10	10	100	0.000	0.000	0.000	(0.000, 0.286)
17	10	100	100	0.065	0.266	0.555	(0.001, 0.904)
18	10	1000	100	0.230	0.437	0.689	(0.025, 0.922)
19	3	10	1000	0.001	0.015	0.136	(0.000, 0.657)
20	3	100	1000	0.136	0.351	0.650	(0.006, 0.959)
21	3	1000	1000	0.209	0.425	0.634	(0.031, 0.972)
22	10	10	1000	0.000	0.002	0.032	(0.000, 0.667)
23	10	100	1000	0.098	0.309	0.573	(0.003, 0.88)
24	10	1000	1000	0.285	0.506	0.662	(0.012, 0.925)
25	3	10	10000	0.004	0.052	0.245	(0.000, 0.674)
26	3	100	10000	0.151	0.367	0.657	(0.005, 0.973)
27	3	1000	10000	0.244	0.492	0.798	(0.060, 0.972)
28	10	10	10000	0.000	0.000	0.012	(0.000, 0.744)
29	10	100	10000	0.109	0.326	0.585	(0.000, 0.875)
30	10	1000	10000	0.212	0.394	0.740	(0.031, 0.911)
31	3	10	$1e5$	0.019	0.111	0.375	(0.000, 0.841)
32	3	100	$1e5$	0.117	0.297	0.599	(0.006, 0.903)
33	3	1000	$1e5$	0.295	0.496	0.677	(0.043, 0.958)
34	10	10	$1e5$	0.000	0.004	0.048	(0.000, 0.395)
35	10	100	$1e5$	0.098	0.306	0.619	(0.002, 0.977)
36	10	1000	$1e5$	0.288	0.474	0.714	(0.025, 0.958)
37	3	10	Gaussian	0.001	0.057	0.222	(0.000, 0.687)
38	3	100	Gaussian	0.080	0.283	0.646	(0.002, 0.933)
39	3	1000	Gaussian	0.196	0.443	0.700	(0.009, 0.908)
40	10	10	Gaussian	0.000	0.001	0.022	(0.000, 0.504)
41	10	100	Gaussian	0.047	0.180	0.411	(0.001, 0.824)
42	10	1000	Gaussian	0.127	0.469	0.758	(0.007, 0.936)

Table 3.8: Results of sensitivity analysis assessing the robustness of the proposed framework under different scenarios. N: number of instruments, g: number of sources measured by each instrument, df: degrees of freedom for the multivariate t distribution. For each case, the dimension $d = 3$, and there are $r = 3$ replicates for each source. The median standard error for these results is 0.027.

Chapter 4

Sources of Variability in Fingerprints

4.1 Introduction

In chapters 2 and 3, we introduced a theoretical foundation for fully-resourced databases and demonstrated their advantages over the more commonly available sub-resourced databases. In this chapter, we apply the ideas of such a database in the context of fingerprints and provide recommendations on factors that can be used to define the "instrument" introduced in Chapter 2 based on the sources of variability between multiple marks produced by the same source.

The factors that contribute to variability in fingerprints (or fingermarks) produced by the same source can be divided into two categories: controllable and uncontrollable. The controllable factors are those that could be replicated and thus could be used to define an instrument in the database. Because uncontrollable factors cannot be easily replicated, they cannot be accounted for in a database; however, they are included in measurement error. In section 4.2, we describe fingerprint examination process to provide context around the process in which a fingerprint database could be used. In section 4.3, we discuss the potential sources of variability that could exist between marks produced by the same source through the lens of a crime scene investigation. In section 4.4, the potential sources of variability between prints produced by the same source are explored, and in section 4.5 are recommendations of the sources of variability that should be accounted for in a fully-resourced database.

4.2 ACE-V

In an forensic examination, the main objective for a latent print examiner is to determine whether a mark from a crime scene and a print produced in a controlled environment are from the same source. In order to make such an assessment, the examiner must understand the cause of any discrepancies found between the mark and the print. According to *The Fingerprint Sourcebook*[29], a manual produced by the U.S. Department of Justice in 2011,

> *There is no such thing as a perfect or exact match between two independent prints or recordings from the same source. Each print is unique; yet an examiner can often determine whether unique prints originated from the same unique source.* [pg. 9-8]

Forensic examiners rely on knowledge and expertise, set protocol, and experience to determine whether the discrepancies indicate the mark and print are from different sources, or if they are from the same source and the discrepancies exist due to the conditions under which the mark and print were produced. Before we describe potential sources of discrepancies considered by examiners, we outline *ACE-V*, the process examiners use to analyze and compare a print and mark to determine if they were produced by the same source. ACE-V stands for *Analysis*, *Comparison*, *Evaluation*, and *Verification*; we describe each of these steps in detail below. In general, there is not one set of formal guidelines that every department adheres to regarding the ACE-V procedures and documentation. The Scientific Working Group on Friction Ridge Analysis (SWGFAST, now known as the Organization of Scientific Area Committees [OSAC])) has produced *Standard for Documentation of ACE-V*; however, this standard is not widely enforced [52]. Our description is based on notes from [29] and the description of the process given by two latent print examiners from the Virginia Department of Forensic Sciences. Therefore, we describe the process in general terms, keeping in mind that it can vary by department.

ACE-V: Analysis The goal of the analysis phase is to determine if each mark recovered from the crime scene is of high enough quality to be used for comparison. If there are multiple overlapping marks, a separate assessment is made for each mark. To make the assessment, the examiner determines if a discernible pattern exists in the mark and a possible explanation about why the pattern exists. They make note of three levels of detail:

- *Level I*: Overall flow and pattern of the mark
- *Level II*: Details about individual ridges
- *Level III*: Additional features that may not show in the print. These features may include interesting shapes, widths (which vary due to pressure), pores, creases, etc.

One feature that is especially useful for comparison is the *continuous ridge*, a ridge that contains no minutiae. Another feature that may be seen on a mark is a *dot*, a ridge in which its width is equal to its length. These are often not as useful as a continuous ridge, because it may be difficult to determine what is actually a dot versus residue or some other distortion.

As the examiner inspects the mark, they document relevant features using a system indicating the quality level of each feature. The examiner then determines whether the mark is "suitable" for examination based on the quantity and quality of information that can be garnered from the mark. Examiners must use Level I or Level II details to determine suitability, since Level III detail is the most unreliable. Because the goal of this phase is to merely determine whether a mark is suitable for comparison, the examiner may not identify all of its features. They often document just enough features to meet the criteria for making a determination regarding suitability. If a mark is determined to be suitable, the examiner proceeds with the next phase - comparison. Otherwise, the process ends.

Comparison During this phase, a side-by-side comparison is made between the mark and the fingerprints on a ten-print card from a person of interest. The examiner looks both for similarities and dissimilarities between the mark and each print. They start with the overall ridge flow pattern and *target groups*, unique clusters of minutiae identified in the mark, to determine which fingerprint to more thoroughly examine first.

Since a side-by-side comparison is used, the fingerprint is notated based on the mark; there is no independent analysis of the print. At this point, the examiner also identifies any features in the mark not documented during the analysis. A study by [52] found that there is a high rate of change in feature markups between the analysis and comparison phases. This result is expected, since many examiners identify only enough detail to determine whether a mark is suitable for comparison in the analysis phase, anticipating that there may be more features to document during comparison.

ACE-V: Evaluation Once the examiner compares a mark and print, a conclusion is made regarding whether or not the mark and print are from the same source. The conclusion is reached considering a guidelines and standards set by The Organization of Scientific Area Committees (OSAC, formerly SWGFAST). There are three possible conclusions: identification (formerly individualization), exclusion, and inconclusive.

An exclusion must be made based on at least two target groups and a focal point (like an anchor point). These criteria are needed because there may be an area in the mark that is not shown in the print and vice versa. For example, the mark may be of the side of a finger, but the prints on the ten-print card typically do not include the side of the finger. Moreover, because the marks are produced in uncontrolled conditions, the mark will likely contain only part of the finger, and it is sometimes unclear which part of the finger it is. Often, the data supporting an exclusion decision include spatial information in relation to the anchor point.

A decision of inconclusive is made when either the mark or print do not provide enough information to make a conclusion. The examiner will specify if the inconclusive decision is based on the mark or based on the print. If the decision is based on the print, a request can be made to re-do the ten-print card of the person in question.

ACE-V: Verification Not all forensic labs regularly conduct the verification. During this step, the verifying examiner may know the original decision (identification, exclusion, inconclusive); however, they will independently conduct the analysis and comparison phases to reach their own conclusion. A *blind verification*, in which the verifying examiner does not know the original conclusion, is preferable to reduce bias in the assessment. One risk of not using blind verification is the analysis by the verifying examiner can be incomplete and inaccurate [48], since the verifying examiner knows the previous conclusion. After the 2004 Madrid train bombing investigation by the FBI (discussed in more in Chapter 6), one recommendation to reduce the risk of erroneous identifications was to keep the original examiner's documentation "sealed or withheld from the [verifier]." [48].

4.3 Variability in Multiple Fingermarks Produced by the Same Source

Fingermarks are obtained from a crime scene by forensic crime scene investigators or police officers trained in collecting evidence. Because the fingermarks were produced under uncontrolled and often times volatile situations, there are many potential causes for distortion or other unusual patterns observed on the fingermark. Latent print examiners must consider these causes when comparing the fingermarks to a print from an identified individual. According to [53], "a criticism of the latent print community is that the examiners can too easily explain a 'difference' as an 'acceptable distortion' in order to make an identification." Chapter 9 of [29] gives a list of reasons for which distortion (variability) could be observed in marks and prints produced by the same source. Using this list as a guide, we identify five basic factors from which variability in fingermarks produced may derive: *investigator, surface, equipment, source* and *scene*. In order to design a database that sufficiently resourced, it is important to understand each of these factors in detail and how they are related not only to the variability seen in marks produced by the same source but also the distribution of that variability, i.e. how the amount of variability may change as the factors change.

4.3.1 Investigator

We start with the investigator who collects fingermark evidence from a crime scene, because as a forensic practitioner, (s)he is the most important "instrument" in this portion of the process [15]. The investigator makes decisions regarding the type of evidence to collect and the equipment and methods used to collect it. Our primary focus is understanding the causes of distortion in a fingermark, so an investigator's decision regarding which pieces of evidence to collect is beyond the scope of this research. However, the appropriateness of the equipment and process the investigator uses to obtain fingermark evidence form an item could significantly impact the appearance and quality of the fingermark.

Various factors such as experience, training, level of knowledge, mood at the time of the investigation, and confirmation bias can affect the decisions made by an investigator [17, 15]. The

method to collect evidence is largely based on the surface on which a mark is found; therefore, the investigator must rely heavily on their knowledge and previous experience to make accurate and timely decisions in each step of the evidence collection process. Though there is not a lot of research specific to the performance crime scene investigators, there is research assessing the performance of forensic practitioners more broadly. Because crime scene investigators use similar information to make decisions in their work as other forensic practitioners, the conclusions about the broader forensic community are still helpful in our understanding of how the investigator potentially affects fingermark evidence. Thus, we examine the variability in fingermarks that stems from the differences in the investigators' ability to accurately make these decisions. There are two factors that can be used to describe this ability: an investigator's experience and their ability to make the correct decisions. Edmond et al. [17] refers to this as being an *expert* versus having *expertise*.

Edmond et al. [17] describes an *expert* as someone who has had thousands of hours in the field due to years of experience, compared to a novice who has very little time in the field. Because the expert has extensive prior experience, they tend to draw on the knowledge they've gained from previous investigations to make decisions. Moreover, they have developed strategies and an instinct about how to proceed in an investigation, which leads them to the correct decision. In contrast, the novice relies more heavily on formal textbook training rather than instinct, because their instinct is often inaccurate.

Though these distinctions between experts and novices exist in the decision-making process, there is conflicting evidence regarding the differences in the performance between these two groups. Some studies suggest that practitioners with more experience perform better than novices on measures of performance [55, 31]; however, others find little relationship between the amount of experience and measure of performance [45, 47]. Experience alone should not be used to establish scientific validity [21]; therefore, in addition to experience, we should also account for the *expertise*, i.e. the level of knowledge and skill, of the investigator.

Some of the conflict between experience and expertise may be due to the quality of feedback an investigator has received throughout their career [17]. Therefore, we should consider the quality of training and feedback investigators receive when determining expertise [15]. There are training sources available for investigators, yet there is no national standard in regards to training, making it unclear as to the the best practices for measuring the quality of training. Holder et al. [29]

recommends latent print examiners complete competency testing when they first arrive at an agency to ensure each examiner has the minimum requirements to be successful at the job and proficiency tests throughout their tenure to measure the quality of their performance. A similar structure of testing cold be used for investigators to differentiate the levels of expertise between investigators. There are some online proficiency tests available, such as [38]; however, there is no standardized test used by investigation units around the nation.

4.3.2 Surface & Scene

Aside from the investigator, the surface on which a mark is found is the most important factor contributing to the variability in marks produced by the same source. The condition of the surface on which the mark was made, including the texture, surface area, shape and curvature, condensation, contaminants and other factors, affect the quality of the recovered mark [29]. The different properties of a surface can be categorized as the following: type, texture, curvature, substance, and color. Each of these factors can not only affect the appearance of the mark but they determine the type of equipment and processes the investigator uses to retrieve the mark.

Surface There are three basic types of surfaces: porous, non-porous, and semi-porous. Depending on the type of surface, different techniques are used to process and photograph the mark. *Porous* surfaces, such as paper, cardboard and wood, are absorbent; the mark tends to absorb into the surface making it somewhat durable. Because of the absorption, chemicals (rather than powders) are often used to process marks on these surfaces [50]. *Non-porous* surfaces, such as glass, metal, and lacquered wood, are not absorbent; therefore, marks produced on these surfaces are more fragile. Because the marks on these surfaces are more susceptible to damage, powders are commonly used to process and retrieve the marks [29]. Lastly, *semi-porous* surfaces both absorb and repel residue from fingermarks. These surfaces include objects such as glossy cardboard or magazine covers. When processing marks on these surfaces, the investigator determines if the mark has absorbed into the surface and chooses the appropriate recovery process accordingly.

The texture of a surface affects how well a mark can make contact with it. A finger typically does not make full contact with a highly textured surface, therefore a mark recovered from one of these surfaces will typically not have as much detail than one recovered from a smooth surface.

Additionally, the mark will have many irregularities in its appearance, since the texture of the surface will show through the image of the mark [29]. In addition to texture, the color determines the equipment that should be used to create as much contrast in the image as possible between the mark and the surface. Finally, additional substances on the surface, such as blood or grease, determine how the mark can be recovered. For example, if the mark is impressed into the substance, it cannot be easily lifted and a photograph of the mark is the only documentation of the mark the examiner has to use for comparison.

Scene One of the factors a latent print examiner should consider when analyzing a mark is the condition of the scene starting from the point in time the mark was created to the time the mark is recovered. This includes environmental factors, such as the temperature and humidity [29]. If the surface is under conditions such as extreme temperatures or high humidity, the mark will be potentially degraded by the time it is recovered [43]. In addition to the environmental factors, the examiners should consider whether there are overlapping marks either from the same source or different sources. Depending on the level of overlap, the mark may not be suitable for comparison.

4.3.3 Equipment

Photography & Lighting Police first used photographs in investigations to help them identify the faces of repeat offenders [29]. Since then, the use of photographs have been a regular part of crime scene investigation, because they are objective recordings that give an accurate and detailed depiction of a crime scene, often capturing details that witnesses don't remember [11]. A photograph should always be taken of the surface on which a mark is found, even if the mark can be recovered from the surface or the surface is on an object that can be taken into a forensic examination lab [29]. Sometimes a mark cannot be recovered from a surface, such as when a mark is embedded in a substance such as grease or the surface is fragile such as a wall with peeling paint. In those instances, a photograph of the mark is the primary source of evidence examined in the lab. Photographs are most effective at capturing marks on flat surfaces, because there are complications with depth perception in photographs of marks on textured or irregularly shaped surfaces [29].

Photography can be done digitally or using film. In film photography, the grain size and film format affect the level of detail that can be seen in the photograph. In general, large film with

smaller grain size produce higher resolution images that show finer levels of detail [29]. Currently, the overwhelming majority of photographs are taken using digital photography. In digital photography, the resolution determines the level of detail that can be seen in the photograph. Photographs with higher resolution show finer levels of detail which may be useful in the examination process. SWGFAST set guidelines that a photograph of a mark must have a resolution of at least 1000 pixels per inch (ppi) for it to be used as evidence. In addition to resolution guidelines, digital photos of marks must be stored in uncompressed file formats such as .TIFF or .RAW, so image detail is not lost in the compression process. If a mark is on a flat surface that can be transported to a lab, the mark may also be documented by scanning the object directly. The resolution standards for scanned images are the same as for photographed images [29].

Though an investigator does not need to be an expert photographer, it is important that the proper equipment is used based on the type of evidence that is being photographed and the type of process used to expose the mark [29]. One major consideration when photographing a mark is the type of lighting used. The lighting can come from a variety of sources, such as a photographic lab lamp, electronic flash, forensic light source, and many others. Additionally, various lighting techniques can be used based on the surface on which the mark is found. For example direct reflection lighting should be used on flat surfaces, because it creates a high contrast in which a mark processed with black, gray or silver powder will appear dark on a light background. In contrast, *oblique lighting*, lighting placed at a low angle, should be used to photograph marks that are in surfaces such as grease, blood or wax, since the type of lighting exposes detail in the mark by creating shadows [29]. These are just two of the many lighting techniques an investigator can use, which are primarily on the surface being photographed. Because it is the investigator's responsibility to choose the appropriate lighting technique, we conclude that much of the variability in multiple marks due to the photography and lighting can be understood by differences in the surface and investigator expertise.

Processing Powders & Chemicals Often a mark cannot be easily seen on a surface, so processing using powders or chemicals is done to make it more visible. If a surface is non-porous, a powder can be used to make a mark more visible. Once the surface is photographed, the investigator applies the powder to the mark using a brush, and any excess powder is removed by lightly tapping the surface [46]. The type of powder used is determined by the color and texture of the surface. Gray powders are used on dark surfaces, black powders are used on light surfaces, and there are some

bi-chromatic powders that can be used on both [27]. Fluorescent powder can be used for surfaces that are colors for which it is difficult to see a mark in a photograph; the fluorescent powder makes the print more clear under an ultraviolet light [27]. In addition to color, if the mark is on a sensitive surface, a special powder such as a magnet-sensitive powder can be used to expose the mark. This type of powder is less sticky than the traditional powders; therefore, marking it more likely to only stick to the mark [27].

Chemicals are often used to process marks on porous surfaces, because the oils from the mark tend to absorb into the surface [27]. There are many types of chemicals that can be used to process a mark; however, we will focus on some of the most commonly used chemicals. *Ninhydrin* (Ruhemann's Purple) is commonly used on surfaces such as wood and paper [50]. Within an hour of it being sprayed onto a surface, the mark will appear in a blue or purple color [27] making it visible for photographing and lifting. Ninhydrin is not good to use on wet surfaces; however, *Small Particle Reagent* (SPR) can be used to process surfaces that are wet or greasy. *Cyanocrylate* (Super Glue) can be used to process marks on non-porous surfaces, such as glue or metal, which may not respond well to powder [50]. Once it is sprayed, the mark appears as a white adhesive forms on the surface [27].

[50, p. 139] provides a table about the best chemicals to use based on the surface. Ultimately, it is up to the investigator to know the best material to use depending on the surface, thus much of the variability due to the type of powder or chemical used to process the mark may be understood by the surface and the expertise of the investigator.

Lifters Once the mark has been exposed through the processing material, the investigator *lifts* the mark to be taken into the forensic examination lab. Similar to the photography and processing steps, the method used to lift the mark is determined by the surface on which the mark is located. According to [29], there are four main types of lifting material that are used by investigators: transparent tape, hinge lifters, rubber-gelatin and lifting sheets.

Transparent tape can be frosted or clear. The investigator rolls out tape and captures the mark by pressing the tape on the mark. [50] provides some best practice guidelines for using this method so that the mark is not smudged or contain extra lines from the tape. Once the mark is captured, the tape is placed on a *backer*, a card that should be a contrasting color from the powder. There are special tapes, such as the *stretchable polyethylene tape*, that are used to lift marks from texture

surfaces. These types of tape are thicker than the usual lifting tape, so they are able to effectively capture the mark from the contours of the surface [29].

Rubber-gelatin lifters are used to obtain marks from fragile or irregularly shaped surfaces, such as a wall with peeling paint or a doorknob. These lifters are more pliable and less sticky that the conventional transparent tape, therefore, they are less likely to damage the fragile surface when lifting the mark. Once the mark is lifted, a plastic sheet is placed over the rubber [29]. The final type of lifter is a lifting sheet, also known as a *flexible lifter* [50]. This type of lifting material is used when the mark is being obtained from a deceased person. [50] provides best practices on using these lifters so the investigator accurately captures the mark without distorting it.

4.3.4 Source

One of the main factors contributing to the variability in multiple marks produced by the same source is the source itself. That variability can be divided into two aspects: the contact between the source and surface and the condition of the source.

Contact Between Source and Surface In a crime scene, there is little control in the way the source comes into contact with a surface. Differences in the pressure and movement of the source cause discrepancies between repeated marks. The National Academies Press [51] stated that every impression left by the same finger would be different because of "inevitable variations in pressure, which change the degree of contact between each part of the ridge structure and the impression medium." Maceo [35] studied the variation in the impressions in an experiment that included repeated impressions made by the two index fingers from one individual. Based on this experiment, the harder the source pressed onto the surface (more pressure), the more area of the source that made contact with the surface. When the pressure was very high, the edge of the source made contact with the surface and was thus included in the mark. The edge of the source did not make contact when little pressure was applied. The National Academies Press [51] also observed differences in the details within the image for different levels of pressure. As the pressure increased, the ridges appeared wider and the furrows appeared more narrow. Increased pressure also reduced the depth of the furrows, making finer details, such as dots, appear larger.

Along with pressure, a finger can move multiple ways as it creates an impression: horizontally,

vertically, and in rotation. Fagert and Morris [22] conducted an experiment studying the effects of movement using 30 fingers from 27 individuals. The individuals were recorded performing different types of movements as they created inked impressions on a glass surface. They measured the degree of movement for minutiae in the impressions after each type of motion. One general finding from their work was that regardless of the direction of the motion, the further minutiae were from the core, the more they moved as the finger moved.

When the finger moved horizontally as it created the impression, the minutiae appeared shifted in the direction of the motion [35, 22]. Under the experimental conditions of [22], the lower part of the finger made more contact with the surface than the upper part. Therefore, the minutiae on the lower part of the finger showed more horizontal movement than those on the upper part. Maceo [35] also found that the increased pressure affected the amount of displacement of each minutiae. In general, for translation movement, the harder the pressure, the harder it is to move the "stick region", i.e. the core. Therefore, under high pressure, more extreme movement is required to move the sticking region which results in more distortion of the impression. Besides translating the minutiae, [35] found that furrows expand on the leading side (e.g. to the left of the core when moving left) and compress on the trailing side. They also found some smearing from the initial contact between the source and the surface.

Vertical movement causes similar translation in the minutiae as horizontal movement. Up and down movement of the finger produces up and down displacement of the minutiae, respectively. There were some incidences of horizontal displacement seen in the arched areas of the mark [22]. Maceo [35] also made conclusions regarding vertical movement that were similar to the findings about horizontal movement. When the finger was pushed up, the furrows on the top half of the finger (between the core and tip) expanded and the furrows at the bottom compressed. The opposite occurred when the finger was pushed down. Additionally, there was some smearing from the finger's original contact with the surface.

Rotation of a finger causes minutiae displacement that's congruent to where the minutiae lie on a circle [22]. In general, the minutiae around the core stay in a similar spot while the minutiae around the ridges show a curved displacement [35]. Additionally, there is some relationship between the rotation direction and the pattern type of the finger [35]. If the finger has a right slant loop, the displacement is more flexible in regards to ridge flow when there is counter-clockwise rotation than

clockwise rotation. Fingers with a whorl pattern show similar displacement when rotating clockwise and counter-clockwise.

Condition of the Source In addition to movement, the condition of the source can affect the quality of a mark and contribute to variability between marks from the same source. These factors could be related to the circumstances in which the mark was produced, such as residue on the finger from dirt, sweat or grease on the source. The factors could also be due to more permanent conditions of the skin. Friction ridges maintain the same pattern from the third month of fetal development barring disease or mutilation [27]. Even if the pattern remains the same, the ridges could flatten from aging, occupation, health or disease [29].

4.4 Variability in Multiple Fingerprints from the Same Source

Fingerprints are created under controlled conditions, such as a police station, and the FBI provides detailed protocol regarding how these prints should be collected [24]. Each agency follows these or similar guidelines when capturing fingerprints. Because the prints are captured under controlled conditions, if a print is not of preferred quality when it is originally created, it can be retaken. If a source reproduces a fingerprint, the original fingerprint card should be retained for documentation purposes, but it is not used for comparison [50]. Therefore, the variability between multiple prints by the same source is largely reduced [40]; however, there are some factors that could contribute to the minimal variability seen between prints by the same source: the method in which in the print is captured and the source.

Method There are two primary methods used to capture fingerprints: ink and digital live scan. When capturing a print using ink, a black ink made specifically for capturing prints should be used. Other types of ink (e.g. printer ink) are too thin and may smear easily [29]. The ink is applied to the source either by using an ink roller or through contact between the source and an ink pad made of ceramic or resin. The source is then pressed onto a ten-print fingerprint card, such as the Henry Ten-Print Card, that is made of card stock to create the impression [29]. The right hand makes up fingers 1-5 on the card and the left hand fingers 6-10. On each hand, the fingers are individually

captured in the following order: thumb, index, middle, ring, little. At the bottom of the card are a *plain impression* from each hand,a print made of the four fingers pressed down simultaneously. [27, 29, 50]. The purpose of the plain impression is to ensure each finger is individually captured in the correct order. The inked ten-print card is scanned into the Automated Fingerprint Identification System (AFIS) or Next Generation Identification (NGI) database, so it can be used for comparison.

Fingerprints can also be captured using a high-resolution digital live scan instead of ink. The main advantage of using the live scan over ink is that the prints are automatically transmitted to AFIS and NGI [27]. This reduces the potential of damage to the print from handling. The sequence of prints captured digitally is the same as described for prints captured using ink.

Whether the print is captured through ink or digitally, there is a certain technique that should be used to ensure a high quality print is captured. Similar to collecting evidence at a crime scene, the individual's hands should always be photographed before any prints are captured [50]. The purpose of the photograph is to document any debris or defects on the source that may appear in the impression. When rolling the print, the operator should have full control of the hand and the subject should be relaxed. This means that the operator controls what part of the finger is captured in the impression and the pressure used [27, 50]. The print should be rolled using one smooth and continuous motion from one side of the nail to the other side. The most common reason a print is rejected by NGI is because the impression does not fully extend across both sides of the fingernail and from the top of the finger to the first joint [27]. Therefore, to ensure that every potentially relevant part of the print is captured, the resulting print should extend from the tip to just below the first joint of the finger [27].

Source Individuals should wash their hands using soap and water before the print is captured. Alcohol can also be used if there is residue that is hard to remove [50]. There is protocol the agency should follow to reduce the effect of the source being too wet or too dry, so these are less of a factor in variability in fingerprints. The main variability that can come from a source is any modification to the source, such as visible scars and injuries. If the individual has a temporary injury, such as a wound that will eventually heal, it is preferable to wait until the injury has healed to create the impressions [27]. However, if the print must be captured at the present time, the injury should be documented in the appropriate block of the fingerprint card [27].

4.5 Creating a Fully-Resourced Database

Given the different sources of variability described in sections 4.3 and 4.4, we provide recommendations on the factors to consider when creating a fully-resourced database. To create such a database, we are only able to account for variability due to the controllable factors - those that could be replicated in an experimental design; variability due to uncontrollable factors are included in measurement error. In regards to prints, the main source of controllable variability is whether or not the print is captured using ink or digital live scan. As stated earlier, because of the controlled conditions under which the print is captured, there is very little variability between prints from the same source. Therefore, to keep the database manageable, it is not necessary to account for different conditions under which fingerprints are produced. By eliminating fingerprints from the database, we treat *Instrument X* that captures fingerprints in controlled conditions (such as a digital live scan) as having no measurement error. Thus when comparing whether a fingermark and fingerprint were produced by the same source, the weight of evidence for in (3.8) would take the form of (4.1).

$$BF_{01}(\mathbf{X}_1, \mathbf{Y}_1, \mathbf{D}_X, \mathbf{D}_Y) = \frac{\int \pi(\mathbf{Y}_1|\boldsymbol{X}_1, \boldsymbol{\Sigma}_Y) p(\boldsymbol{\Sigma}_Y|\mathbf{D}_Y) d\boldsymbol{\Sigma}_Y}{\int\int \pi(\mathbf{Y}_1|\boldsymbol{\theta}_Y, \boldsymbol{\Sigma}_Y)\pi(\boldsymbol{\theta}_Y|\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_0) p(\boldsymbol{\Sigma}_Y|\mathbf{D}_Y) d\boldsymbol{\theta}_Y d\boldsymbol{\Sigma}_Y} \tag{4.1}$$

The recommendations for creating a database, therefore, are focused on accounting for variability between multiple fingermarks produced by the same source. In section 4.3, we examined five factors that contribute to differences between fingermarks: investigator, surface, equipment, source, and scene. The source and scene are factors that can not be easily replicated, since they are unique to the specific circumstances under which a mark was originally made. Thus, these are uncontrollable factors that would be included in the measurement error. The equipment, investigator, and surface are all controllable factors that can be used to develop a fully-resourced database. As discussed in section 4.3.3, the investigator uses properties of the surface to determine which equipment to use to recover the mark. Given this, in the spirit of making a database manageable, we recommend considering two factors in the creation of a fully-resourced database; investigator expertise and surface.

- *Investigator Expertise*: The investigator makes decisions regarding the procedure and equipment to use, so these factors are confounded with investigator. The relationship between

experience and expertise is not well defined, so we use both in defining the investigator aspect of an instrument. The database should include three levels of investigator: *novice, proficient* and *expert*. There is currently no national requirement related to proficiency testing for investigators, so we use the International Association of Identification (IAI) certification program [5] as a guide for defining each level. An investigator who has less than one year of experience is classified as a novice, since there is a one year minimum requirement to receive any certification. An investigator who has at least one year of experience and passes proficiency test that is of the same level as the test for the IAI's Certified Crime Scene Investigator certification is categorized as proficient. Lastly, an investigator with at least six years of experience and scores at least a 75% on a proficiency exam equivalent to the IAI's Certified Senior Crime Scene Analyst certification exam is categorized as an expert. Otherwise the examiner with at least six years of experience is categorized as proficient. Since we are merely trying to categorize the investigators for the sake of defining an instrument rather than assess them for certification, the requirements we have established are not as rigorous as those required for the IAI certification; however, these tests are a good guide since they are a standard recognized by investigators in the field. We don't include investigators who fail the proficiency exams, since these examiners are likely to have errors in their coursework that would make the data unreliable to use for our purposes.

- *Surface*: The surface on which a mark is found is the main source of variability in the appearance of multiple marks. Guides for investigators such as the "Latent Print Overview" from the Division of Forensic Sciences in the Georgia Bureau of Investigation [1] focus mainly on the texture and porousness of the surface when explaining how to recover fingerprints from a crime scene. Therefore, we use these elements to help define the surface aspect of instrument in the database. The database should include marks from the same source produced on surfaces with different textures (categorized as *low* and *high*) porousness (categorized as *non-porous*, *semi-porous*, and *porous*).

The database structure shown in Figure 4-1 is based on a blocked experimental design. The levels of investigator are used to define the "blocks", and the levels of texture and porousness are used to define the "treatment" groups within each block. (Note, the block design is merely used as a way to organize the data; it is not intended to define any randomization scheme or assignment of treatments. Therefore, this design has no effect on the analysis described in previous chapters.) The

groups shown in Figure 4-1 are used to determine which instruments are "similar" to the instruments used in the current investigation. For example, Group 1 consists of all instruments that include a novice investigator collecting prints from a non-porous surface with low texture. This group could include marks recovered by an investigator with six months of experience from glass windows, metal surfaces such as door knobs, and ceramic surfaces such as pottery or dishes. Using these categories as a guide, we recommend that a lab define the instruments to be included in their database, based on the surfaces that are relevant to the lab's range of casework.

Investigator = Novice

Texture	Porousness		
	Non-Porous	Semi-Porous	Porous
Low	Group 1	Group 2	Group 3
High	Group 4	Group 5	Group 6

Investigator = Proficient

Texture	Porousness		
	Non-Porous	Semi-Porous	Porous
Low	Group 7	Group 8	Group 9
High	Group 10	Group 11	Group 12

Investigator = Expert

Texture	Porousness		
	Non-Porous	Semi-Porous	Porous
Low	Group 13	Group 14	Group 15
High	Group 16	Group 17	Group 18

Figure 4-1: Proposed database structure based on a blocked experimental design.

The fingermark images in the database will be collected from previous casework. In order for a fingermark to be included in the database, we would require information about the surface from which the fingerprint was recovered along with information indicating the level of expertise of the investigator. As mentioned before, there are currently no standardized proficiency exams from which we could measure the expertise of examiners with over one year of experience. To get around this limitation, the initial database could be created using the casework of those investigators who currently have some level of IAI certification. This is not ideal, since the group of investigators who seek and obtain certification are not representative of all investigators; however, this issue would be reduced going forward as the database is updated. Annual proficiency exams could be

calibrated against the exams given for the IAI certifications, and investigators associated with the new casework entered into the database would have their expertise classified based the results of the proficiency exams.

After initially creating the database, one may wish to combine the data from multiple instruments and classify it all under one instrument. Therefore, when defining the instruments in the database, multiple instruments could be combined and classified as a single instrument. This could be done if one or more instruments in the database have very low sample size, for example. Since one of the main uses of the database is to understand measurement variability, we recommend that two (or more) instruments could be combined and redefined as a single instrument if they have measurement variability that do not differ in a statistically significant way. For example, suppose there are two instruments - *Instrument 1: Expert Investigator, Flat and Highly-Textured Wood Surface* and *Instrument 2: Proficient Investigator, Flat and Highly-Textured Wood Surface*, in other words, the level of expertise for the investigator are the only differentiating properties between these two instruments. Additionally, suppose the measurement error for both instruments show no statistically significant difference. We can combine these two instruments to define a single instrument as *Combined Instrument: Proficient/Expert Investigator, Flat and Highly-Textured Wood Surface*.

One potential issue with combining instrument is the practical significance of the newly defined instrument. When combining instruments, one should be mindful that the combined instrument is defined in such a way that the analysis from the database provides useful and practical information. Therefore, even if two instruments have very similar measurement error, it may not be advisable to combine them into a single instrument if newly defined instrument can't be used to make decisions in a useful way. For example, if *Instrument 1: Expert Investigator, Flat and Highly-Textured Wood Surface* and *Instrument 2: Novice Investigator, Curved and Smooth Brass Surface* have similar measurement errors, we advise against combining these instruments, since the newly combined instrument would not be defined in such a way that insights about such an instrument could be used in a meaningful way.

The fingerprints in the database should come from a collection of cases that are representative of the casework in the lab. For example, if gun homicides make up $X\%$ of the investigations from which a lab processes fingerprints, then approximately $X\%$ of the cases represented in the database should be gun homicides. This is due to the fact that there are different amounts of variability across

crimes and thus the information in the database should reflect actual casework. Moreover, since the statistical methods in chapters 2 and 3 are based on ANOVA decomposition, we recommend that the number of unique sources and the number of replicates of each source be as balanced as possible across instruments. Since the prints in the database will initially come from previous casework, this may not be feasible in practice. Based on the analysis in Chapter 3, our methods can still be used even if this balance does not hold.

Though the database would be publicly available, users would go through a registration process and be approved by an administrator before gaining access to the database. Users would then login to access the database and could download data from the database onto their machine. This process is modeled after the one used for the NIST Ballistics Toolmark Research Database [2] which requires registration and login for use.

As more latent print examiners use statistical methods in their work, they could also access the database to obtain prior information. Similar to researchers, examiners would be required to register and login order to access the data. Additionally, examiners would need to include documentation regarding the data they retrieved from the database for the calculation of prior information. Some items to document would be the group number describing the set of similar instruments used in the calculations, a brief note regarding the reason that group was chosen, the fitted parameter values from the framework, and the Bayes factor that was considered in their decision-making. We discuss some guidelines for reporting this Bayes factor in Chapter 6.

4.6 Conclusion

Using the crime scene investigation process, we identified the controllable and uncontrollable factors that contribute to the variability in multiple fingermarks produced by the same source. These factors informed recommendations for creating a fully-resourced database that meets the criteria for our framework and could be implemented in practice. Now that we have described both the theoretical and practical aspects of our framework, we put these together in the next chapter to show how it can be used from start to finish when comparing a fingermark and fingerprint to determine if they were produced by the same source.

Chapter 5

Application: Fingerprint Database

5.1 Introduction

In this chapter, we apply the ideas from chapters 2 - 4 to a scenario in which a print and a mark are compared to determine if they were produced by the same source. One of the main objectives of this chapter is to gain practical insights about database design and how it can be used based on a demonstration using simulated fingerprints. Moreover, we will apply the metrics from section 1.4 to translate the images in the database to multivariate vectors of data. Lastly, we will use fingerprints from the database to show a complete demonstration of how information from the database is used in calculating the Bayes factor to quantify the weight of evidence that two prints are from the same source. At this point, we have shown our results in a more general form, so the results from this chapter will provide the opportunity to be more specific about how our framework applies.

5.2 Database Creation

Currently in practice, there is no such database like the one we've described that is publicly available. Additionally, there are few collections of fingerprints that include multiple impressions produced by the same source. The databases that most closely meet these criteria are the Fingerprint Verification Competition (FVC) databases [13]. These include multiple fingerprints produced by

each source under different conditions such as varying pressure, moisture on the source, rotation, etc. These variations are related to the source; therefore, we account for the variability represented in these databases in measurement error. Using the specifications in Section 4.5, an instrument is defined based on the investigator and surface, which are not accounted for in the FVC competition databases.

Additionally, our proposed database includes information from previous casework that is representative of the situations that exist in practice. The FVC competition databases are not necessarily concerned with covering the crime scene scenarios that happen in practice, thus it is unclear if these databases achieve any such representation. Since the main objective of the database is to use it to understand variability, it is important that the database include information that accurately represents the variability in practice.

These discrepancies between what was included in each database and how we define our proposed database tended to exist for the publicly available databases that include multiple prints produced by the same source. Therefore, we created a database using simulated fingerprints for this demonstration.

5.2.1 Fingerprint Simulation

Using the insights from chapters 3 and 4, we created a database that contains four instruments, each with 500 sources. As described in Section 4.5, the database only includes fingermarks recovered from a crime scene, since we can assume there is negligible variation between multiple prints produced by the same source under controlled conditions. We used the Anguli Fingerprint Simulator [3] to generate the fingerprints and their subsequent fingermarks for the database. As shown in Figure 5-1, the user can change the settings for noise, translation, number of scratches, and rotation to simulate fingerprints under different conditions. To simulate these marks in databases S_Y and S_Y^0, high levels of the noise and translation were used with wide ranges to account for the large measurement error for instruments in S_Y and S_Y^0. For some instruments in S_Y^0, higher levels for the number of scratches setting were used to simulate prints that were produced on textured surfaces. The metric used for this demonstration accounts for different orientations of a fingerprint (described fully in section 5.3), therefore, to reduce the complexity of the feature detection process,

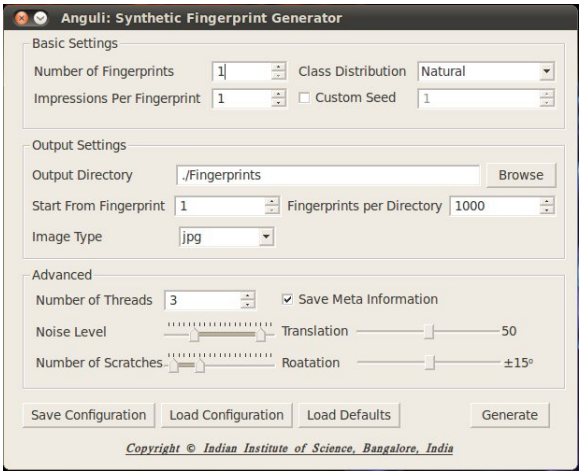

Figure 5-1: Anguli fingerprint generator window. (Image from [3]).

Figure 5-2: Fingerprint impressions from the same source simulated using Anguli. Left: Fingerprint simulated using low levels of noise, translation and scratches. R: Fingermark simulated using higher levels of noise, translation and scratches.

the rotation was held constant for all instruments. Figure 5-2 shows examples of simulated prints produced under various settings in Anguli.

In addition to the simulated fingermarks for S_Y^0 and S_Y, fingerprints were simulated that mimic the fingerprints examiner would use to compare the suspect of interest to the fingermark from the scene. Since these fingerprints represent those that are produced under controlled conditions, very low levels of noise, scratch and rotation were used in Anguli.

5.3 Translating Images to Numerical Summaries

Fingerprint images are stored in the database, therefore they must be translated into numerical summaries to be used with the framework presented in chapters 2 and 3. This process contains two parts: (1) feature extraction and (2) using a metric to translate features into numerical summaries.

In practice, a fingerprint examiner identifies the important features in a fingerprint image during the Analysis and Comparison steps of ACE-V. Features deemed important could include minutiae, general ridge flow pattern, region identification (core, delta, etc), or any other features that are needed to calculate the numerical summaries. The metric used in this demonstration only requires minutiae, so we focus specially on minutiae detection. To find the minutiae, we used MINDTCT, a fingerprint feature extraction program produced by the National Institute of Standards and Technology (NIST). Each fingerprint image is input into the program, and file is returned that contains the minutiae location (described using X and Y coordinates), minutiae angle, type, and quality . Within MINDTCT is a set of criteria to minimize the number of false minutiae identified, so we will assume that the points identified by the program are genuine minutiae. A detailed description of MINDTCT is in Appendix A.

To translate the fingerprint images to numerical data, we will use the shape and type from Neumann et al. [41] and a measurement of the direction produced by MINDTCT that is similar to the direction metric calculated in Neumann et al. [41]. Before obtaining data from the database, we calculate the metrics from the fingerprint of the candidate of interest, since this will be used to inform what information is drawn from the database. In practice, the examiner documents the minutiae on the fingermark of interest and selects k minutiae to be used to quantify the weight of evidence. For our demonstration, we select k minutiae from the fingermark to use. Once the minutiae from the fingermark are chosen, we calculate the numerical data from the k minutiae using the following process:

1. The configuration of k minutiae is selected from the fingermark out of the minutiae that have a quality score of at least 20. This threshold helps provide further assurance that the minutiae used in the calculations are useful features of the fingerprints and not other extraneous marks. The center of the configuration is calculated using the arithmetic mean of the X and Y coordinates of the k minutiae. Once the centroid is established, k triangles are

created by connecting each minutia to the centroid and two adjacent minutiae.

2. Shape is calculated for each of the k triangles. The shape includes two elements - the form factor and the aspect ratio. The form factor is the ratio between the area of the triangle and its perimeter, and the aspect ratio is the ratio between the diameters of the triangle's circumcircle and incircle.

3. The aspect ratio is used to establish the orientation of the configuration. The first triangle is the one with the minimum aspect ratio; triangles $2, \ldots, k$ are numbered consecutively going counterclockwise. The aspect ratio data is removed.

4. The direction of the minutiae is calculated using the angular data provided by MINDTCT as shown by angle (A) in Figure 5-3. All angles are calculated relative to the same orientation, i.e. the horizontal axis pointing to the right is 0°. For a ridge ending, the angle is calculated using a line that extends outwards from the ridge ending. For a bifurcation, the angle is calculated using a line that extends into the "valley" of the bifurcation, i.e. away from the bifurcated ridges. This differs from the direction metric in Neumann et al. [41] (shown in Figure 1-5), in which each angle is calculated using an axis that's drawn based on a minutia's location relative to the centroid of the configuration.

5. The final vector of data used to describe the fingermark is

$$Y = [y_{1,S}, \ldots, y_{k,S}, y_{1,T}, \ldots, y_{k,T}, y_{1,D}, \ldots, y_{k,D}]$$

such that $y_{i,S}$ is the form factor data, $y_{i,T}$ is the type data, and $y_{i,D}$ is the direction data for the i^{th} minutia.

We choose the k minutiae from the fingerprint that will be used as a reference to pull the relevant information from the database. The minutiae chosen from the fingerprint are the set of k minutiae that most closely align with the minutiae from the fingermark. In order to determine the minutiae that most closely align with the fingermark, we use the following process:

5. Identify all possible minutiae configurations of size k using the minutiae with a quality score of at least 20. As before, using this threshold provides further assurance that the minutiae chosen are useful features of the fingerprint rather than extraneous details. There will be

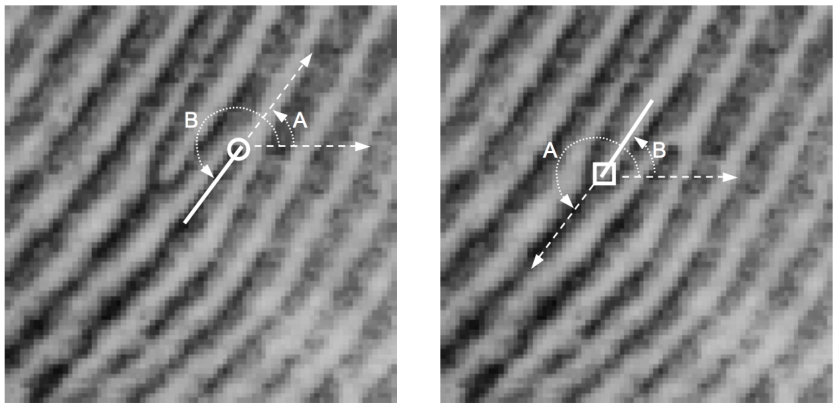

Figure 5-3: Calculation of direction angle used in this demonstration from the MINDTCT output. Left: Angle calculation for a ridge ending. Right: Angle calculation for a bifurcation. In each picture, (A) is the angle output by MINDTCT that is used in this demonstration. (B) is the angle calculation used in IAFIS. Image credit: [54].

$\binom{n}{k}$ combinations, where n is the number of minutiae identified by MINDTCT that meet the threshold, so this also helps reduce the number of configurations being considered in a way that makes the calculations computationally feasible without losing important information from the fingerprints.

6. For each configuration, the numerical summary of data is calculated using the process described in steps 1 - 4 above. The resulting data for each minutiae configuration is

$$X_i = [x_{i1,S}, \ldots, x_{ik,S}, x_{i1,T}, \ldots, x_{ik,T}, x_{i1,D}, \ldots, x_{ik,D}]$$

where X_i is the numerical summary of the i^{th} minutiae configuration.

7. The distance between the summary from each configuration X_i and Y is calculated using

$$d(X_i, Y) = d(X_{i.,S}, Y_{.,S}) + d(X_{i.,T}, Y_{.,T}) + d(X_{i.,D}, Y_{.,D})$$

such that

$$d(X_{i.,S}, Y_{.,S}) = \sqrt{(X_{i1,S} - Y_{1,S})^2 + \cdots + (X_{ik,S} - Y_{k,S})^2}$$
$$d(X_{i.,T}, Y_{.,T}) = \sum_{j=1}^{k} 1[X_{ij,T} \neq Y_{j,T}]$$
$$d(X_{i.,D}, Y_{.,D}) = \sqrt{(X_{i1,D} - Y_{1,D})^2 + \cdots + (X_{ik,D} - Y_{k,D})^2}$$

8. The minutiae configuration that minimizes the distance to Y is selected. We call this configuration X.

The numerical summary X is then used to identify the k-minutiae configurations to use for each mark in the database. For each mark in the database, we use the same process that was used to select X. This time, X is the reference print we use to identify the configurations. Additionally, since this process is being done for thousands of marks in the database, we want to ensure that what we are doing is computationally feasible. The number of possible minutiae configurations can become quite large, so for computational efficiency, we randomly select 10,000 minutiae configurations to be considered if the number of configurations is too large. Since we are still using a very large subset of the minutiae configurations, the gains in computational efficiency are more substantial than the potential information lost from the minutiae configurations that are not selected.

As it stands, our proposed model is built for data that follow a multivariate Gaussian distribution. Thus, we will incorporate the form factor and direction data in our framework to calculate the weight of evidence. Our current framework is not built to handle categorical data, therefore, we will not use type in the calculation of the weight of evidence. The information from type was used to select the best minutiae configuration, so this data was taken into account in the analysis. In Chapter 7, we discuss future work to extend the framework to categorical data.

The raw direction data is circular defined on the interval $(0°, 360°]$. A common approach for modeling circular data is the von-Mises distribution, which does not fit into our framework. Thus, instead of using the raw directions from each of the fingermarks in the database, we will consider the direction relative to the reference print. The following formula is used to quantify the direction

for the i^{th} minutia.

$$Y_{i,D*} = \begin{cases} -[(X_{i,D} - Y_{i,D}) \bmod 360] & Y_{i,D} \leq X_{i,D} \\ (X_{i,D} - Y_{i,D}) \bmod 360 & Y_{i,D} > X_{i,D} \end{cases} \tag{5.1}$$

Using the data in this form, we not only capture how much the minutiae angle from a fingermark deviates from the minutiae angle of the reference print, but we also capture in which direction the angle deviates. Since this is circular data, we should be mindful of the endpoint, since $0° = 360°$. Since our data is defined on the interval $(0°, 360°]$, any minutia with a direction pointing horizontally and to the right is recorded as $360°$. The distributions of D^* by minutia position (component) when $k = 5$ is shown in Figure 5-5.

5.3.1 Neumann et al. (2015)

Since the metric used in this demonstration is largely based on [41], we will describe how they use shape, direction, and type data in their calculation of the weight of evidence. We also discuss some differences between the way they apply the metric compared to our application of it in order to provide some context for our results.

Given a fingermark Y recovered from a crime scene, a person of interest Mr. X, and X, a fingerprint from Mr. X, Neumann et al. [41] calculates a likelihood ratio based on the following hypotheses:

$H_0 : X$ and Y were both produced by Mr. X

$H_1 : X$ was produced by Mr. X, and Y was produced by another individual in the relevant population

They use k minutiae from the fingerprint and fingermark for the analysis. The k minutiae chosen from the fingerprint are the configuration of k minutiae out of all possible k minutiae configurations from Mr. X that most closely aligns with the fingermark. We will call the data from the k minutiae configurations $Y^{(k)}$ and $X^{(k)}$. Using this data, they begin with the following basic form for the likelihood ratio to quantify the weight of evidence in support of H_0.

$$LR = \frac{P(Y^{(k)}|H_0, v=1)P(v=1|H_0) + P(Y^{(k)}|H_0, v=0)P(v=0|H_0)}{P(Y^{(k)}|H_1, v=1)P(v=1|H_1) + P(Y^{(k)}|H_1, v=0)P(v=0|H_1)} \tag{5.2}$$

Let V be an indicator variable, such that $V = 1$ if the print from the individual is sufficiently similar to the fingermark, and $V = 0$ otherwise. "Sufficiently similar" is determined based on the matching algorithm from the 3M Cogent AFIS system used by Neumann et al. [41]. In general, under H_0, X and Y will be deemed "sufficiently similar"; therefore, they make the simplifying assumption $P(v = 1|H_0) = 1$ and $P(v = 0|H_0) = 0$. Additionally, the relevant population used to calculation the denominator of (5.2) is chosen based on AFIS's matching algorithm, so they make another simplifying assumption that $P(v = 0|H_1) = 0$. Incorporating these assumptions, (5.2) becomes

$$LR = \frac{P(Y^{(k)}|H_0, v=1)}{P(Y^{(k)}|H_1, v=1)} \times \frac{1}{P(v=1|H_1)} \tag{5.3}$$

To select the relevant population, AFIS selects the k-minutiae configuration from an individual if it determines that is similar to the mark. Therefore, to calculate $P(v = 1|H_1)$, they use the number of individuals selected by AFIS divided by the total number of fingerprints in the database. To calculate $\frac{P(Y^{(k)}|H_0,v=1)}{P(Y^{(k)}|H_1,v=1)}$, they use the shape, direction, and type information described in Section 1.4. Using the assumption that within a specified location of a fingerprint, the shape, direction and type are considered independent. Therefore, the likelihood ratio in (5.3) can be calculated as

$$\begin{aligned} LR &= \frac{P(Y_S^{(k)}, Y_D^{(k)}, Y_T^{(k)}|H_0, v=1)}{P(Y_S^{(k)}, Y_D^{(k)}, Y_T^{(k)}|H_1, v=1)} \times \frac{1}{P(v=1|H_1)} \\ &= \frac{P(Y_S^{(k)}|H_0, v=1)}{P(Y_S^{(k)}|H_1, v=1)} \times \frac{P(Y_D^{(k)}|H_0, v=1)}{P(Y_D^{(k)}|H_1, v=1)} \times \frac{P(Y_T^{(k)}|H_0, v=1)}{P(Y_T^{(k)}|H_1, v=1)} \times \frac{1}{P(v=1|H_1)} \end{aligned} \tag{5.4}$$

Based on empirical analysis, they assume independence between each of the k minutiae when calculating the density for each component. The form of the likelihood ratio then becomes

$$LR = \prod_{i=1}^{k} \frac{P(Y_{S,i}^{(k)}|H_0, v=1)}{P(Y_{S,i}^{(k)}|H_1, v=1)} \times \prod_{i=1}^{k} \frac{P(Y_{D,i}^{(k)}|H_0, v=1)}{P(Y_{D,i}^{(k)}|H_1, v=1)} \times \prod_{i=1}^{k} \frac{P(Y_{T,i}^{(k)}|H_0, v=1)}{P(Y_{T,i}^{(k)}|H_1, v=1)} \times \frac{1}{P(v=1|H_1)} \tag{5.5}$$

As previously described, the shape component is based on k triangles rather than the k minutiae. To calculate the numerator of the shape component in (5.5), [41] uses a univariate Gaussian density for the first triangle and bivariate Gaussian densities for the subsequent triangles. Kernel density estimation is used to calculate the denominator of the shape component. They use a non-parametric distribution that is based on von-Mises kernels to calculate the numerator and denominator of the direction component. Lastly, to calculate the type component of the likelihood ratio, they establish a table of probabilities based on a survey of 200 latent print examiners in which each examiner was asked to identify the type of a series of minutiae on fingermarks. The probability is the probability the minutia is type l given an examiner marked it as m. They assume there is no uncertainty in an examiner's determination of minutia type when examining a fingerprint.

We now discuss some of the differences between our demonstration and the use of the metric to that in Neumann et al. [41]. The first difference is in the relevant set from which the k minutiae used in the examination are selected. Neumann et al. [41] starts with an individual and considers all of the k-minutiae configurations across that individual's ten fingers (sources). Therefore, there are $\sum_{i=1}^{10} \binom{n_i}{k}$ possible k-minutiae configurations, where n_i is the number of minutiae marked on the i^{th} source from the individual. It is then the responsibility of the examiner to choose the k-minutiae configuration that most closely matches to the k-minutiae configuration identified in the fingermark Y during comparison. We begin with a source rather than an individual, thus we start with only $\binom{n_i}{k}$ possible $k-$minutiae configurations. Our framework could be modified to consider the individual rather than the source by marking fingermarks based on the unique individual instead of the unique source.

Neumann et al. [41] quantifies the strength of evidence using what they call an approximate likelihood ratio. To account for the variability that exists between multiple marks produced by the same source in the numerator, a distortion model [40] (see section 1.2 for detail about the model) is used to simulate multiple impressions from the same source that have different appearances due to causes of variability such as those presented in Chapter 4. The evidence that the print and mark are from different sources is calculated in the denominator in which the variability between multiple prints produced by different sources is understood based on a distribution created using prints from different sources collected from the database. Because the relevant population is determined by a matching algorithm in AFIS, the set of prints used to calculate the evidence in the denominator consists of those most similar to the mark in question.

This process of calculating the approximate likelihood ratio contains the two biggest differences to account for when considering the context in which the shape metric is used. Neumann et al. [41] uses the fingerprint database to obtain a set of relevant prints that can be used to quantify the between source variability. Moreover, because AFIS scores are used to identify these prints, they are naturally the ones that most closely match the fingermark. In our framework, prints in the database are used to calculate parameters for the prior distributions on the variance parameters connected to the between source variability. The sources used to understand variability are not necessarily those in the database that most closely match the $k-$minutiae configuration from the fingermark being investigated. As discussed earlier, our aim is to obtain information from fingermarks that were produced under situations that are representative of what would be seen in casework and provide a good representation of the variations of surface and investigator that would be seen in casework. Lastly, the denominator of (5.5) does not account for within fingerprint variability in order to simplify the calculations. Because our database requires there be multiple prints from each source, the within source variability can be understood using the information from the database.

5.4 Preparing Data for Analysis

Because our current framework is not designed for categorical data, we will use the shape (aspect ratio and form factor) and direction components with a mind towards quantifying the weight of evidence that a mark and print were produced by the same source. The underlying assumption of our proposed framework described in Chapter 3 is that the data follow a multivariate Gaussian distribution; therefore, we will asses how well the data in our database meet this assumption and apply an appropriate transformation if necessary.

To determine if the data follow a multivariate Gaussian distribution, we will use the multivariate distribution comparison test and examine the summary statistics as those described in Section 3.4.3. To obtain the multivariate Gaussian distribution in each iteration, we will simulate multivariate Gaussian data using the mean and covariance estimated from the data in the database.

We begin by examining the form factor. Based on the test statistics and p-values in Table 5.1, it is clear the data do not follow a multivariate Gaussian distribution and a transformation is required.

	Q_1	Q_2	Q_3	Middle 95%
T_A	9.193	10.689	13.042	(5.603, 17.883)
P-value	0.000	0.000	0.000	(0.000, 0.001)

Table 5.1: Test statistic, T_A, and p-value calculated using the test in [37] comparing the distribution of the form factor data to a multivariate Gaussian distribution.

To transform the data, we apply a univariate Box Cox transformation [9] to each component of the form factor. Figure 5-4 shows the distributions of the original and transformed data for each component along with λ, the power used to transform the data. We use these plots to visualize differences between the original and transformed data.

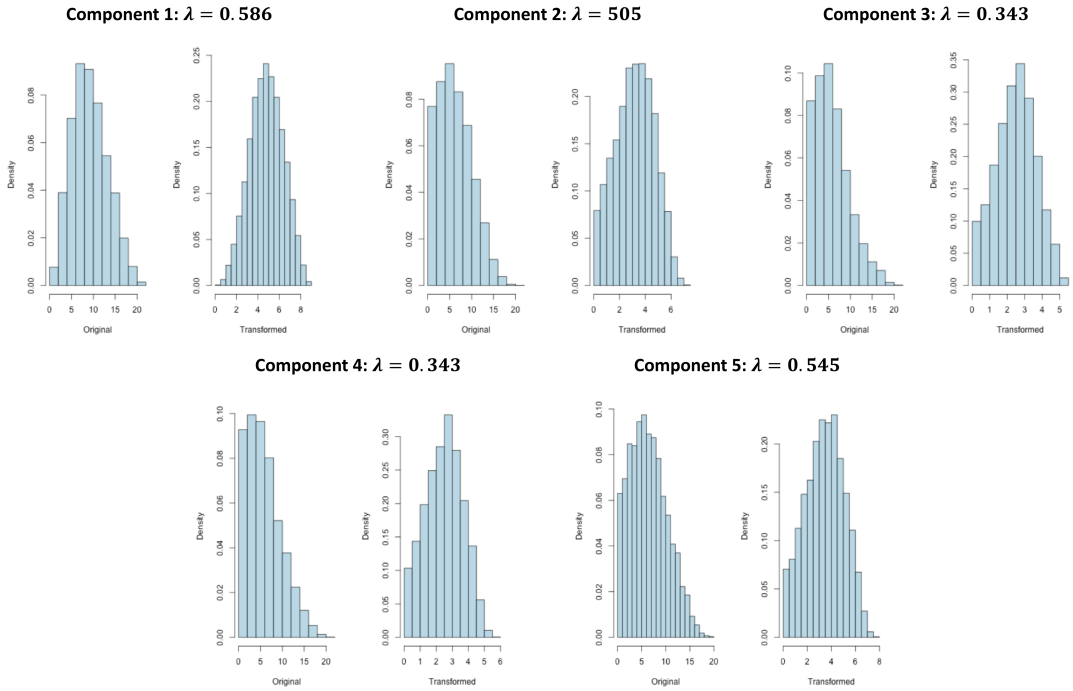

Figure 5-4: Original and transformed values of form factor.

We now apply the multivariate distribution comparison test to the transformed data. Based on the summary of the test in Table 5.2, we conclude that the transformed data meets the multivariate Gaussian assumption.

	Q_1	Q_2	Q_3	Middle 95%
T_A	0.320	0.617	1.046	(0.133, 2.139)
P-value	0.374	0.617	0.826	(0.091, 0.950)

Table 5.2: Test statistic, T_A, and p-value calculated using the test in [37] comparing the distribution of the form factor data to a multivariate Gaussian distribution.

We now repeat the same process to examine the aspect ratio data. Based on the results of the multivariate comparison test in Table 5.3, the original aspect ratio data does not follow a multivariate Gaussian distribution and should be transformed.

	Q_1	Q_2	Q_3	Middle 95%
T_A	27.150	29.588	32.535	(20.765, 40.583)
P-value	0.000	0.000	0.000	(0.000, 0.000)

Table 5.3: Test statistic, T_A, and p-value calculated using the test in [37] comparing the distribution of the form factor data to a multivariate Gaussian distribution.

We apply the Box Cox transformation and assess the whether the transformed distribution follow the multivariate Gaussian distribution in Table 5.4.

	Q_1	Q_2	Q_3	Middle 95%
T_A	2.360	3.131	4.036	(1.189, 5.654)
P-value	0.003	0.015	0.053	(0.000, 0.313)

Table 5.4: Test statistic, T_A, and p-value calculated using the test in [37] comparing the distribution of the form factor data to a multivariate Gaussian distribution.

Based on these results, though the distribution of the transformed data appears to be a little closer to the multivariate Gaussian distribution, important departures indicate that even the transformed data likely cannot be treated as multivariate Gaussian. Since the form factor and aspect ratio are both describing the shape of the minutiae configurations, we conduct a test to determine whether the true correlation between form factor and shape is significantly different from zero. The estimated correlation, test statistic, and p-value calculated from a t distribution with $n-2$ degrees of freedom are shown in Table 5.5. Since there is non-zero correlation between the form factor and aspect ratio, the additional information gained from including the aspect ratio in the model is not worth the additional uncertainty in the results due to the violations of the model assumptions. Therefore, we will use the form factor to represent shape in the calculation in the weight of evidence.

r	t	p-value
-0.452	-64.006	$< 2e$-16

Table 5.5: Estimated correlation: r, Test statistic: t, and p-value.

Lastly, we examine the distribution of the direction data. From the results of the multivariate distribution comparison test show in Table 5.6, we determine that a transformation is required. We

use a transformation from Yeo and Johnson [56], an extension of the Box Cox transformation that can accommodate observations that are less than or equal to 0. Given the original observation Y_i and the power parameter λ, the transformed value $Y_i^{(\lambda)}$ is calculated using (5.6).

$$Y_i^(\lambda) = \begin{cases} \frac{(Y_i+1)^\lambda - 1}{\lambda} & \lambda \neq 0, Y_i \geq 0 \\ \log(Y_i + 1) & \lambda = 0, Y_i \geq 0 \\ -\frac{(-Y_i+1)^{2-\lambda}-1}{(2-\lambda)} & \lambda \neq 2, Y_i < 0 \\ -\log(-Y_i + 1) & \lambda = 2, Y_i < 0 \end{cases} \tag{5.6}$$

The values of λ for each component are stated above the histograms in Figure 5-5. When the original observation is positive, the transformed value from Yeo and Johnson [56] is obtained by applying the usual Box Cox transformation. Based on the results of the multivariate comparison test in Table 5.7 and the size of the database, we determine that the transformed data can be used in our framework. Figure 5-5 shows the original and transformed direction data by component.

	Q_1	Q_2	Q_3	Middle 95%
T_A	13.755	15.396	17.248	(10.981,20.142)
P-value	0.000	0.000	0.000	(0.000,0.000)

Table 5.6: Test statistic, T_A, and p-value calculated using the test in [37] comparing the distribution of the form factor data to a multivariate Gaussian distribution.

	Q_1	Q_2	Q_3	Middle 95%
T_A	1.020	1.395	2.027	(0.450, 3.223)
P-value	0.074	0.224	0.402	(0.007, 0.807)

Table 5.7: Test statistic, T_A, and p-value calculated using the test in [37] comparing the distribution of the form factor data to a multivariate Gaussian distribution.

We will use the transformed values of the form factor and direction to fit the parameters to specify the prior distributions in Section 3.3 and calculate the weight of evidence. Given the results of the comparison tests shown in this section and the size of the database used in this demonstration, we can use the transformed data to calculate of the weight of evidence in our framework.

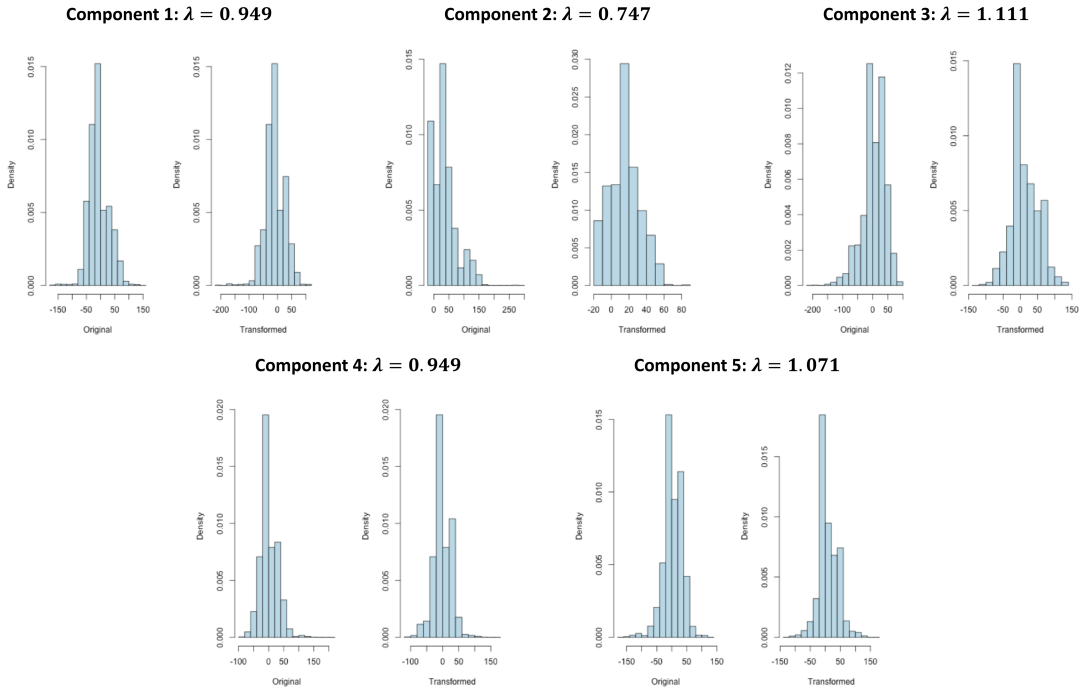

Figure 5-5: Original and transformed values of direction.

5.5 Results

5.5.1 One-Sample Framework

In chapters 2 and 3, we described the statistical foundation based on a two-sample problem in which the numerical summaries $\mathbf{X}_1$ and $\mathbf{Y}_1$ both had variability that needed to be accounted for in our calculations. As discussed in section 4.5, there is negligible variability in fingerprints produced in controlled environments; therefore, we are not accounting for the variability in fingerprints in our database. This not only has the benefit of making a database more feasible to build and maintain, but this also simplifies the formulation of the statistical framework from a two-sample framework to a one-sample one. We will describe the pertinent changes to our framework and then show the results of a demonstration using the sub- and fully-resourced databases in the next two sections.

Suppose $\mathbf{X}_1$ is the multivariate numerical summary from a fingerprint produced under controlled conditions and $\mathbf{Y}_1$ is the multivariate numerical summary from a fingermark recovered from the

scene. As before, we are testing the following hypotheses

$$\begin{aligned} H_0 &: \mathbf{X}_1 \text{ and } \mathbf{Y}_1 \text{ were produced by the same source} \\ H_1 &: \mathbf{X}_1 \text{ and } \mathbf{Y}_1 \text{ were produced by different sources} \end{aligned} \tag{5.7}$$

We have a similar model set up as the one shown in Section 3.1. Let $\mathbf{Y}_1 \sim G(\boldsymbol{\theta}_Y, \boldsymbol{\Sigma}_Y)$. As before, we assume the source means are generated from the distribution $\boldsymbol{\theta}_Y \sim G(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_0)$.

We use the same separation strategy described in Section 3.1 to obtain prior information about the covariance matrix $\boldsymbol{\Sigma}_Y$ in the fully-resourced case. In both the sub- and fully-resourced cases, we calculate the relevant estimated values and fitted parameters using the ANOVA approach presented in sections 3.2 and 3.3. As expected, the only difference in the ANOVA calculations under this set up, is the parameters $\boldsymbol{\theta}_0$ and $\boldsymbol{\Sigma}_0$ are fit using only data from databases S_Y (in the sub-resourced case) and S_Y^0 (in the fully-resourced case).

There is also a change in the formulation of the Bayes factor used to calculate the weight of evidence in support of H_0 in (5.7). The Bayes factor in the sub-resourced case is shown in (5.8).

$$BF_{01}(\mathbf{X}_1, \mathbf{Y}_1, \mathbf{D}_X, \mathbf{D}_Y) = \frac{\pi(\mathbf{Y}_1|\mathbf{X}_1, \hat{\boldsymbol{\Sigma}}_Y)}{\int \pi(\mathbf{Y}_1|\boldsymbol{\theta}_Y, \hat{\boldsymbol{\Sigma}}_Y)\pi(\boldsymbol{\theta}_Y)d\boldsymbol{\theta}_Y} \tag{5.8}$$

The Bayes factor in the fully-resourced case is shown in (5.9).

$$BF_{01}(\mathbf{X}_1, \mathbf{Y}_1, \mathbf{D}_X, \mathbf{D}_Y) = \frac{\int \pi(\mathbf{Y}_1|\mathbf{X}_1, \boldsymbol{\Sigma}_Y)p(\boldsymbol{\Sigma}_Y|\mathbf{D}_Y)d\boldsymbol{\Sigma}_Y}{\int\int \pi(\mathbf{Y}_1|\boldsymbol{\theta}_Y, \boldsymbol{\Sigma}_Y)\pi(\boldsymbol{\theta}_Y)p(\boldsymbol{\Sigma}_Y|\mathbf{D}_Y)d\boldsymbol{\theta}_Y d\boldsymbol{\Sigma}_Y} \tag{5.9}$$

5.5.2 Fitted Values: Sub-Resourced Case

For this demonstration, we use $k = 5$ minutiae. Since this is the sub-resourced case, we use information from S_Y, the set of marks that were collected using the same instrument as the one in the current investigation. We begin by looking at the average fitted values of the aspect ratio in Table 5.8 to check that the minutiae configurations are as expected. The aspect ratio is used to orient each configuration, so we expect for the first value to be the smallest on average. This is true in the table; therefore, we have confidence that the minutiae configurations are as expected for this

demonstration.

1	2	3	4	5
2.876	4470.515	11738.851	1968.611	939.716

Table 5.8: Average fitted values of the aspect ratio.

We will be using the transformed values of form factor and direction to calculate the weight of evidence, so we show the fitted values of $\boldsymbol{\theta}_0$ and $\boldsymbol{\Sigma}_0$ and the estimates $\hat{\boldsymbol{\Sigma}}_X$ and $\hat{\boldsymbol{\Sigma}}_Y$ for the transformed data. The fitted values for $\boldsymbol{\theta}_0$ for the form factor and direction are in Table 5.9.

Component	1	2	3	4	5
Form Factor	4.659	3.071	2.405	2.412	3.315
Direction	-5.747	16.376	12.133	0.361	9.732

Table 5.9: Fitted values of $\boldsymbol{\theta}_0$.

In Table 5.10, we have the fitted values for standard deviation of $\boldsymbol{\Sigma}_0$ and in Tables 5.11 and 5.12, we have the fitted values of the correlation structure $\boldsymbol{\Sigma}_0$. The values of standard deviation provide an indicator of the variation in the mean value of form factor for each source, which provides a measurement of the variability between the different fingermarks in S_Y. The fitted values of correlation indicate how the k form factor measurements relate to one another. We see similar results in the estimated values of standard deviation and correlation from $\hat{\boldsymbol{\Sigma}}_Y$ in tables 5.13, 5.14, and 5.15.

One benefit of using our framework to obtain the fitted and estimated values for $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_Y$ is that we don't need to make any assumptions about the correlation structure beforehand. We can use the relevant data from the database to obtain all information about these variance and covariance matrices. This is especially useful for the correlation structure, which often times assumptions about the structure are made based on prior knowledge or numerical data exploration. In Neumann et al. [41], they assume independence in the form factor values for non-adjacent triangles based on numerical results with $k = 12$ minutiae. We have found that for smaller k, there is non-negligible correlation between non-adjacent triangles, and thus the assumption does not always hold.

Component	1	2	3	4	5
Form Factor	0.979	0.907	0.677	0.708	0.980
Direction	25.078	12.177	27.708	17.736	22.875

Table 5.10: Fitted standard deviations from $\boldsymbol{\Sigma}_0$.

	1	2	3	4	5
1	1.000				
2	0.491	1.000			
3	0.261	0.515	1.000		
4	0.216	0.214	0.458	1.000	
5	0.559	0.281	0.211	0.477	1.000

Table 5.11: Form Factor: Correlation matrix from $\boldsymbol{\Sigma}_0$

	1	2	3	4	5
1	1.000				
2	-0.106	1.000			
3	-0.244	0.297	1.000		
4	0.041	0.028	0.110	1.000	
5	0.204	-0.101	-0.128	-0.074	1.000

Table 5.12: Direction: Correlation matrix from $\boldsymbol{\Sigma}_0$

1	2	3	4	5	
Form Factor	1.416	1.395	1.067	1.086	1.499
Direction	30.864	13.765	34.376	25.838	29.403

Table 5.13: Fitted standard deviations from $\hat{\boldsymbol{\Sigma}}_Y$.

	1	2	3	4	5
1	1.000				
2	0.483	1.000			
3	0.249	0.452	1.000		
4	0.203	0.160	0.480	1.000	
5	0.550	0.216	0.195	0.449	1.000

Table 5.14: Form Factor: Correlation matrix from $\hat{\boldsymbol{\Sigma}}_Y$

5.5.3 Fitted Values: Fully-Resourced Case

With a fully-resourced database, we have the database S_Y^0 that include fingermarks from instruments similar to *Instrument Y* from which we can gather more complete information about measurement variability. Using the data contained in these databases, we fit the values for $\boldsymbol{\theta}_0$ and

	1	2	3	4	5
1	1.000				
2	0.106	1.000			
3	0.045	0.145	1.000		
4	0.067	-0.024	0.141	1.000	
5	0.096	-0.094	-0.032	-0.038	1.000

Table 5.15: Direction: Correlation matrix from $\hat{\boldsymbol{\Sigma}}_Y$

$\boldsymbol{\Sigma}_0$ as before in the sub-resourced case, and we also fit values for $\boldsymbol{\mu}_Y$, $\boldsymbol{\tau}_Y^2$, ν_Y, and $\boldsymbol{\Lambda}_{Y,0}^{-1}$ for both the shape and direction components.

Table 5.16 contains the fitted values for $\boldsymbol{\theta}_0$. Similar to the fitted values in the sub-resourced case, on average, the first triangle has the largest form factor and the remaining triangles have relatively the same average value for form factor. We also see the phenomenon in the direction data that showed in the sub-resourced case; on average, the direction of the second minutiae differed from the reference print much more than for the other minutiae.

	1	2	3	4	5
Form Factor	4.736	3.115	2.450	2.431	3.385
Direction	-7.418	15.859	15.259	-0.002	9.761

Table 5.16: Fitted value for $\boldsymbol{\theta}_0$ calculated using fully-resourced database.

Next, we look at the standard deviation and correlation structures of $\boldsymbol{\Sigma}_0$ shown in tables5.17, 5.18, and 5.19. From the fitted values we can conclude that knowing the triangle number does not provide much information about the variability between sources, since the values are about the same for each triangle. The correlation structure in Tables 5.18 and 5.19 are similar to what we observed in the sub-resourced case, where we observed non-negligible correlation between non-adjacent triangles. In general, we also see the strongest correlation between adjacent triangles.

1	2	3	4	5	
Form Factor	0.980	0.885	0.676	0.705	0.991
Direction	23.438	12.790	26.180	17.079	20.340

Table 5.17: Fitted values for the standard deviations of $\boldsymbol{\Sigma}_0$ calculated using a fully-resourced database.

Tables 5.20 and 5.21 contains the fitted values for the parameters of the logNormal distribution

	1	2	3	4	5
1.000					
0.512	1.000				
0.287	0.507	1.000			
0.275	0.174	0.446	1.000		
0.609	0.281	0.237	0.509	1.000	

Table 5.18: Form Factor: Correlation structure of $\boldsymbol{\Sigma}_0$ calculated using the fully-resourced database

	1	2	3	4	5
1	1.000				
2	-0.097	1.000			
3	-0.129	0.224	1.000		
4	-0.037	0.036	0.194	1.000	
5	0.251	-0.057	-0.088	-0.119	1.000

Table 5.19: Direction: Correlation structure of $\boldsymbol{\Sigma}_0$ calculated using the fully-resourced database

that describes the form of the standard deviations. The scale matrix of the inverse Wishart prior distributions for the correlation structure is in Table 5.22. The correlation structure in the scale matrices reflect what we've seen in other correlation matrices in this demonstration. Using these parameter values, we can now fully describe the measurement variability of an instrument. Being able to fit the the values in tables 5.20 and 5.22 show the main advantage of the fully-resourced database over the sub-resourced database. We will see how this affects the interpretation of the weight of evidence in Section 5.5.4.

	1	2	3	4	5
$\boldsymbol{\mu}_{Y,0}$	0.414	0.396	0.112	0.129	0.442
$\boldsymbol{\tau}_{Y,0}$	0.163	0.163	0.163	0.163	0.163

Table 5.20: Form Factor: Fitted values for $\boldsymbol{\mu}_{X,0}$ and $\boldsymbol{\tau}_{Y,0}$ calculated using a fully-resourced database.

In order to quantify the weight of evidence that a print and mark are from the same source, we use Carlin and Chib [14]. Before describing the details of the algorithm, we discuss the calculation of $p(\boldsymbol{\Sigma}_Y|\boldsymbol{D}_Y)$, the updated prior information for $\boldsymbol{\Sigma}_Y$ for a component.

The form for $p(\boldsymbol{\Sigma}_Y|\boldsymbol{D}_Y)$, such that $\boldsymbol{D}_Y$ numerical data from fingermarks in S_Y, is the same as shown in (3.7) described in Chapter 3. We use a variation on the method in Barnard et al. [7] that takes advantage of the separation strategy (described in detail in Section 3.3) in a Gibbs sampling

	1	2	3	4	5
$\boldsymbol{\mu}_{Y,0}$	3.532	2.701	3.603	3.313	3.407
$\boldsymbol{\tau}_{Y,0}$	0.359	0.359	0.359	0.359	0.359

Table 5.21: Direction: Fitted values for $\boldsymbol{\mu}_{,0}$ and $\boldsymbol{\tau}_{Y,0}$ calculated using a fully-resourced database.

	1	2	3	4	5
1	1.000				
2	0.469	1.000			
3	0.240	0.481	1.000		
4	0.182	0.131	0.420	1.000	
5	0.561	0.210	0.171	0.437	1.000

Table 5.22: Form Factor: Fitted value for $\boldsymbol{\Lambda}_{Y,0}^{-1}$, the scale matrix of the inv-Wishart prior distribution of the correlation structure of $\boldsymbol{\Sigma}_Y$.

framework. As a starting value for the algorithm, we calculate $\hat{\boldsymbol{\Sigma}_Y}$, the estimate of the covariance matrix calculated using data in S_Y. Then, for iterations $i = 1, \ldots, n$, where n is the total number of iterations, we do the following:

1. Use the separation strategy from Section 3.3 to identify the diagonal matrix of standard deviations. Let $diag(\mathbf{S}) = [S_1, \ldots, S_k]$.

2. For $i = 1, \ldots, k$,

   i. Update the i^{th} element in $diag(\mathbf{S})$ by sampling $S_{i^*} \sim \log N(\mu_{Y,0i}, \tau^2_{Y,0i})$. $\mu_{Y,0i}$ is the i^{th} element of the fitted value $\mu_{Y,0}$, and $\tau^2_{Y,0i}$ is the i^{th} element of the fitted value of $\tau^2_{Y,0}$.

   ii. Calculate the updated covariance matrix

   $$\boldsymbol{\Sigma}_{Y^*} = \mathbf{SRS}$$

   iii. Calculate $p(\boldsymbol{\Sigma}_{Y^*}|\boldsymbol{D}_Y)$.

3. Using the updated covariance matrix, $\boldsymbol{\Sigma}_{Y^*}$ (the most current covariance matrix after updating all standard deviation values), use the separation strategy to identify the correlation matrix

	1	2	3	4	5
1	1.000				
2	0.071	1.000			
3	0.017	-0.004	1.000		
4	0.031	-0.034	0.136	1.000	
5	0.029	-0.037	-0.014	-0.016	1.000

Table 5.23: Direction: Fitted value for $\mathbf{\Lambda}_{Y,0}^{-1}$, the scale matrix of the inv-Wishart prior distribution of the correlation structure of $\mathbf{\Sigma}_Y$.

R. Let

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & r_{13} \dots & r_{1k} \\ r_{21} & 1 & \dots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \dots & 1 \end{bmatrix}$$

4. We want to ensure that the covariance matrix stays positive definite, therefore we must carefully update the values in the correlation matrix. To ensure that each updated matrix remains positive definite, we follow the sampling scheme from Barnard et al. [7]. Let $R(r)$ be the updated correlation matrix such that $R[i,j] = R[j,i] = r$ and $f(r) = |R(r)|$.

   For $i = 1, \dots, k$ and $j = 1, \dots, k$ such that $j > i$,

   i. Calculate $f(0)$, $f(1)$, and $f(-1)$.

   ii. Find the roots of $ar^2 + br + c$ such that

   $$a = [f(1) + f(-1) - 2f(0)]/2$$
   $$b = [f(1) - f(-1)]/2$$
   $$c = f(0)$$

   These roots define the interval from which the new correlation r_{ij^*} can be drawn to ensure the covariance matrix remains positive definite.

   iii. Draw updated value r_{ij^*} from the a uniform distribution defined on the interval specified in the previous step. Let $\mathbf{R}[i,j] = \mathbf{R}[j,i] = r_{ij^*}$.

iv. Calculate the updated covariance matrix

$$\boldsymbol{\Sigma}_{Y^*} = \mathbf{SRS}$$

v. Calculate $p(\boldsymbol{\Sigma}_{Y^*}|\boldsymbol{D}_Y)$.

Once the prior information for the covariance matrix $\boldsymbol{\Sigma}_Y$ is updated, we use Carlin and Chib [14] to calculate the Bayes factor. This method was chosen because it does not require integration and it runs efficiently, which is important when considering implementing our framework in practice. In this algorithm, we have two models that align with our hypotheses and write the Bayes factor as shown below

$$\begin{aligned} &M_0 : \boldsymbol{\theta}_Y = \mathbf{X} \text{ vs. } M_1 : \boldsymbol{\theta}_Y \neq \mathbf{X} \\ &BF_{01} = \frac{P(M_0|\mathbf{Y})/P(M_1|\mathbf{Y})}{P(M_0)/P(M_1)} \end{aligned} \tag{5.10}$$

To calculate the Bayes factor in 5.10, we use Gibbs sampling to calculate $P(M_0|\mathbf{Y})$. For each iteration, we do the following:

1. Generate the model parameters for M_0. Under this model, the parameters are $\boldsymbol{\theta}_Y = \mathbf{X}$ and $\boldsymbol{\Sigma}_Y$. The value of $\mathbf{X}$ is determined from the process described in Section 5.2 and $\boldsymbol{\Sigma}_Y$ is drawn from the following:

$$\begin{aligned} P(\boldsymbol{\theta}_Y, \boldsymbol{\Sigma}_Y|M_0) &= \pi(\boldsymbol{Y}|\boldsymbol{\theta}_Y, \boldsymbol{\Sigma}_Y, M = 0)p(\boldsymbol{\theta}_Y, \boldsymbol{\Sigma}_Y|M = 0) \\ &= G(\boldsymbol{Y}|\mathbf{X}, \boldsymbol{\Sigma}_Y)p(\boldsymbol{\Sigma}_Y|\mathbf{D}_Y) \end{aligned} \tag{5.11}$$

2. Generate the model parameters under M_1. Under this model,

$$\begin{aligned} P(\boldsymbol{\theta}_Y, \boldsymbol{\Sigma}_Y|M_1) &= \pi(\boldsymbol{\theta}_Y, \boldsymbol{\Sigma}_Y|M = 1) \\ &= G(\boldsymbol{\theta}_Y|\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_0)p(\boldsymbol{\Sigma}_Y|\mathbf{D}_Y) \end{aligned} \tag{5.12}$$

3. Calculate the conditional posterior probability for M_0.

$$P(M = 0|\boldsymbol{\theta}_Y, \boldsymbol{\Sigma}_Y, \boldsymbol{Y}) = \frac{G(\mathbf{Y}|\mathbf{X}, \boldsymbol{\Sigma}_Y)p(\boldsymbol{\Sigma}_Y|\mathbf{D}_Y)P(M_0)}{\pi(\mathbf{Y}|\mathbf{X}, \boldsymbol{\Sigma}_Y)p(\boldsymbol{\Sigma}_Y|\mathbf{D}_Y)P(M_0) + G(\mathbf{Y}|\boldsymbol{\theta}_Y, \boldsymbol{\Sigma}_Y)G(\boldsymbol{\theta}_Y|\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_0)p(\boldsymbol{\Sigma}_Y|\mathbf{D}_Y)P(M_1)} \tag{5.13}$$

such that $\pi(M_0)$ and $\pi(M_1)$ are the prior probabilities for the respective models. In this demonstration, we use $P(M_0) = P(M_1) = 0.5$; however, in practice, these prior probabilities could be determined using non-fingerprint evidence from the investigation.

4. Randomly draw a model M_0 or M_1 based on the probabilities $P(M = 0|\boldsymbol{\theta}_Y, \boldsymbol{\Sigma}_Y, \boldsymbol{Y})$ and $1 - P(M = 0|\boldsymbol{\theta}_Y, \boldsymbol{\Sigma}_Y, \boldsymbol{Y})$

In this demonstration, we used 10,000 iterations; we did not see sensitivity in the result from changing the number of iterations. Once all the iterations of the Gibbs sampling is complete, the posterior probability of M_0 is estimated as the following:

$$\hat{P}(M_0|\mathbf{Y}) = \frac{\#\text{ occurences of } M_0}{\#\text{ of iterations}} \tag{5.14}$$

Using this value, the Bayes factor can be calculated using the formula in (5.10).

5.5.4 Weight of Evidence

We now examine the weight of evidence calculated using sub- and fully-resourced databases. Table 5.24 shows the weight of evidence in support of the hypothesis that a print and mark are produced by the same source. Under each scenario, the fingerprint under investigation is the same; it is the left-most impression in figures 5-6 and 5-7. Fingermarks 1 and 2 shown in Figure 5-6 were produced by the same source as the fingerprint. Fingermarks 3 and 4 shown in Figure 5-7 were produced by sources that differ from the print.

For each comparison, we calculate the weight of evidence using three different models based on the numerical summaries calculated in sections 5.3 and 5.4: form factor only, direction only, and form factor and direction. The results from these three models provide an indication of how much

Figure 5-6: Left to right: Fingerprint produced under controlled conditions and two fingermarks recovered from an uncontrolled condition such as a crime scene. All three impressions were produced by the same source.

the weight of evidence is dependent on the metrics used in the model. From the results in the Table 5.24, we see that the weight of evidence can change depending on the model. This phenomenon is most prevalent when a sub-resourced database is used. We can use the information from each of the three models to quantify insights about some of the similarities and differences between the mark and print that may be highlighted by examiners in the comparison phase of ACE-V. For example, the minutiae directions in Fingermark 3 and the fingerprint are much more similar than the shape of the triangles in their minutiae configurations. The weight of evidence from the individual components provide additional support for some of the considerations (such as the one in the example) examiners use to make their final conclusions.

Though we can gain insights from the individual models, the model that includes both the form factor and direction provides a more holistic measure of how the fingerprint and fingermark compare. Thus, when considering which metric to use as an input into the modeling framework, we suggest using one that includes multiple numerical descriptions of a fingerprint impression. We also suggesting using a metric that is calculated in a transparent way, so that its components can be used to better understand any interesting similarities and/or differences observed between the print and mark. For example, AFIS scores are often proprietary, and it is unclear how they are calculated. Therefore, they cannot be used to glean additional information about the comparison of a mark and print aside from how closely they match. If such a metric must be used, we recommend also calculating the weight of evidence using a more transparent metric to provide more statistical backing to the examiners' conclusions.

Figure 5-7: Left to right: Fingerprint produced under controlled conditions and two fingermarks produced under uncontrolled conditions such as a crime scene. All three impressions were produced by different sources.

	Sub-Resourced			**Fully-Resourced**		
	Form Factor	**Direction**	**Both**	**Form Factor**	**Direction**	**Both**
Fingermark 1	-14.798	4.091	-10.707	1.933	8.647	5.042
Fingermark 2	-7.155	2.859	-4.296	3.195	11.421	5.583
Fingermark 3	-15.930	2.069	-13.86	1.017	5.3006	3.245
Fingermark 4	-2.010	-23.315	-25.325	3.686	1.2411	1.810

Table 5.24: Weight of evidence ($2 \log BF_{01}$) in support of the hypothesis that the fingermark was produced by the same source as the fingerprint.

In general, the magnitude of the weight of evidence is less when calculated using the fully-resourced rather than the sub-resourced database. This is due to the fact that the variability in the fingermarks is more completely accounted for when using a fully-resourced database. Another thing to note is that we calculated positive weight of evidence even when the mark and print were from different sources. This result is not surprising since the minutiae configurations examined from the fingerprint and fingermark are ones that most closely match. In practice, we recommend using a numerical summary that more clearly highlights the unique features in a fingerprint and fingermark.

Choosing a Numerical Summary

In addition to the transparency of the calculation, measures of performance can be used to choose the model (determined by the numerical summary (or summaries) included in the framework) used to quantify the weight of evidence. We recommend using the following performance characteristics from Meuwly et al. [36]:

- *Accuracy*: How well the weight of evidence aligns with the truth. For example, if the fingerprint and fingermark are from the same source, the model with the highest value for weight of evidence would be deemed the most accurate.
- *Discriminating Power*: How well the model distinguishes between fingerprint and fingermark comparisons in which different conclusions are true.
- *Calibration*: How meaningfully the model results can be interpreted regardless of which conclusion is true. Models that have better discrimination power have better calibration.

One way to evaluate the performance characteristics is by using a log-likelihood-ratio cost (Cllr) [10], a measure of how well the weight of evidence corresponds with a correct conclusion. The models should be evaluated using a set of real fingerprints and marks with known sources that represent what is seen in casework [36]. The Cllr for accuracy can be decomposed into the Cllr for discriminating power and the Cllr for calibration, so accuracy can be used to choose a model. Because the Cllr is a cost measure, the model with the lowest value of Cllr is the one that is most accurate. If one wishes to include multiple models in their assessment of the weight of evidence, we recommend using a weighted average of $2\log BF_{01}$ in which the weights are based on the model accuracy as determined by the Cllr. The weights should be assigned such that the most accurate model has the largest when in the aggregate calculation of weight of evidence.

5.6 Conclusion

From the results of this demonstration, we recommend that examiners use a fully-resourced database to quantify the weight of evidence when reporting their conclusions. This would provide more context around their conclusion than just a binary "match/no match". In addition to their conclusion and the important similarities and differences in the features that led to such conclusion, examiners could report the weight of evidence $2\log BF_{01}$ in support of the print and mark being produced by the same source. By presenting this weight of evidence, those interpreting the evidence to make a decision (such as a jury or judge, for example), can use this weight of evidence to assess their confidence in the examiner's conclusion. This helps provide more transparency in the process, and provides a mechanism in which examiners can provide "probative" value for the conclusions as suggested by Dror and Mnookin [16].

As shown in this demonstration, using a sub-resourced database can make the weight of evidence appear artificially strong. As discussed in Chapter 2, the sub-resourced database is the type most commonly found in practice at this point. Therefore, in order to truly quantify the weight of evidence, the fully-resourced database should be used. We have shown that the weight of evidence as measured by the fully-resourced database more accurately accounts for the variability that exist in the measurements and therefore could be more accurately interpreted by a party using the results, such as a jury. Additionally, since we have confidence that the weight of evidence calculated from a fully-resourced database incorporates more aspects of measurement variability, we can be even more confident in an examiner's conclusion when the magnitude of the weight of evidence is high.

Thus far in this text, we have only considered the case in which one candidate is investigated as the potential source of a fingermark recovered from a crime scene. However, in practice, examiners often analyze multiple candidates as the potential source of a mark. In the next chapter, we explore the statistical properties of this phenomenon with the goal of providing more practical guidelines examiners can use when analyzing multiple candidates.

Chapter 6

Dependencies in the Weight of Evidence for Multiple Candidates

6.1 Introduction

On March 11, 2004, terrorist bombings occurred on commuter trains in Madrid, Spain. On March 19, 2004, Brandon Mayfield, an attorney living in Portland, Oregon, was identified by the Federal Bureau of Identification (FBI) as the source of a fingermark recovered from the scene of the attack, and on May 6 he was arrested as a material witness to the attack. On May 19, the Spanish National Police (SNP) identified Ouhnane Daoud as the source of the fingermark. On May 20, Mayfield was released from prison after the FBI compared Daoud's fingerprints to the mark, and on May 24, all charges against Mayfield were officially dropped [42]. This is one of the most infamous instances of an erroneous identification in modern forensic science. Due to the magnitude of the error by the Latent Print Unit in the FBI, the Office of Inspector General (OIG) in the U.S. Department of Justice conducted a review to investigate how the erroneous identification was made. The Office of Inspector General concluded that some of the main sources of error included the large amount of similarities between Mayfield's and Daoud's fingerprints, the original examiner's reliance on minute Level III detail to make the identification, and the lack of adequately explaining the differences between the mark and Mayfield's print [42]. In addition to the investigation by the OIG, the FBI implemented corrective actions to the Latent Print Unit after a review of the Mayfield

investigation by an international committee of latent print examiners and forensic experts [48].

Identification errors this one have caused increase criticism about the scientific validity of methods used to analyze forensic evidence [51, 21]. Though understanding the source of the errors that occurred is important for assessing the reliability of forensic evidence, we are interested in exploring another aspect of the Mayfield case, specifically the weight of evidence in support of the hypothesis that Mayfield was the source of the fingermark found at the scene. As more evidence about Daoud was examined by the FBI, the weight of evidence in support of Mayfield as the source of the mark decreased. In fact, it decreased to the point that he was eventually cleared as the source of the mark. We propose a way to quantify this phenomenon by measuring the weight of evidence in a way that accounts for the dependencies that exist when multiple candidates are examined as potential sources.

We begin by by exploring different ways to consider dependencies in the weight of evidence of multiple candidates and how these could be applied to the Mayfield case in section 6.2. Then, we demonstrate our proposed method mimicking a two candidate scenario in section 6.3.1 and extend the analysis three or more candidates in sections 6.3.2 and 6.3.3. In section 6.4 we show some of the asymptotic properties of our proposed method, and in section 6.5 we conclude with guidelines for interpreting the weight of evidence when multiple candidates are examined.

6.2 Interpreting Evidence in the 2004 Madrid Train Attack

Brandon Mayfield was the fourth candidate retrieved in a search using the FBI's Integrated Automated Fingerprint Identification System (IAFIS). The search was conducted using seven minutiae points marked on an image of one of the fingermarks recovered from the crime scene by SNP [48]. Before comparing Mayfield's fingerprint to the mark, the initial latent print examiner had likely conceptualized high prior odds that Mayfield was the source of the mark, given that he was ranked high on the candidate list produced by an IAFIS search in a database containing millions of prints. Examiners have expectations of the system's efficiency and ability to identify matches [15] and have the potential to be influenced in their decision making based on a candidate's rank on the list of

IAFIS candidates [16, 17]. Given the large size of the FBI database, it would not have been unusual to retrieve a non-match very similar to the fingermark, yet it is unclear how much this possibility was included in the examiner's assessment.

Once the initial identification was made, Mayfield's prints were examined by a second latent print examiner who verified the original conclusion. The second examiner was informed that the original examiner made an identification to Mayfield [48]. This piece of information coupled with the fact that the original examiner was an experienced supervisor who was highly respected [48], likely increased the second examiner's estimation of $\frac{P(H_0)}{P(H_1)}$. After the second examiner verified the original conclusion of identification, the Latent Print Unit chief officer confirmed the identification [42]. Because of the FBI's identification determination, the SNP requested Mayfield's prints for examination in April 2004. They reached a "negativo" (negative) conclusion in regards to Mayfield being the source of the mark [42]; [48] specifies that the conclusion was "inconclusive." Members from the FBI Latent Print Unit met with the SNP to discuss the differences in their analysis; however, about a week prior to the meeting, the FBI stated that they were "absolutely confident" in their original conclusion that Mayfield was the source of the fingermark [42].

The original FBI examiners did not have detailed knowledge about Mayfield's personal and professional life when the original identification was made. However, by the time the SNP released the *April 13 Negativo Report*, the examiners in the FBI Latent Print Unit were aware of more personal details about Mayfield obtained from the Portland division of the FBI. They had been informed that Mayfield practiced Islam, had contacts with suspected terrorists, and had once acted as an attorney for a convicted terrorist [42]. Office of the Inspector General, Oversight and Review Division [42] concludes that this knowledge about Mayfield's personal life wasn't the primary cause for the FBI's lack of careful review after the SNP report; however, there is evidence that it could have been an influence. An examiner admitted that had the details about Mayfield been more "mundane", there may have been more urgency to more carefully review the original conclusions and possibly catch the errors made [42]. For the examiners in the FBI Latent Print Unit who hadn't seen the original fingerprint analysis, the combination of the FBI's original identification conclusion, the SNP's "negativo" conclusion and the additional details about Mayfield's personal would lead them to assess the prior odds differently than the original examiner.

At each stage of the Mayfield investigation, the new examiners reviewing the evidence had a

different set of information that influenced their determination of the prior odds of Mayfield being the source of the fingermark. Lund and Iyer [34] describe a similar scenario in which multiple people are independently deriving their own estimates of the prior odds. Though they talk about "decision makers" mainly in reference to non-experts (e.g. jurors, judges, attorneys) who use the information from forensic experts to make decisions, many of the challenges they present about assigning weights to prior probabilities can also be true of forensic experts. Therefore, we use their framework can be used to describe the Mayfield case. There is a *relevant population*, a set of scenarios, $H_0 : \{H_{0i}\}_{i=1}^{a}$ in which the examiner may consider Mayfield the source of the fingermark and a set of scenarios $H_1 : \{H_{1j}\}_{j=1}^{b}$ in which the examiner would not consider Mayfield the source. Before comparing Mayfield's print to the mark, each examiner assigned a probability to each scenario regarding their belief in its plausibility. As demonstrated in the Mayfield case, can be problematic, because each examiner may use different information to define their relevant population and prior belief about each scenario.

These issues were demonstrated in the Mayfield case, as each examiner had different information on which to base their relevant population. Similarly, when the SNP conducted its original analysis, Mayfield was not in the examiners' relevant population, since his fingerprints were not in their database. One could argue this affected the amount to which the examiners in the SNP believed he was the source of the mark. Similarly, the eventual culprit, Daoud, was not in the FBI's relevant population, since he was not in the FBI database. The original FBI examiners would have assigned the scenario of Daoud being the source of the fingermark zero in the original analysis. However, [34] argue that any situation could be indisputable if sufficient data is provided, so no scenario is truly assigned prior probability of exactly zero. This is true in the case of Daoud who the FBI eventually determined was the source of the mark after the data derived from analyzing his fingerprint.

In Chapter 2, we proposed a a framework for using a database to systematically derive prior information rather than rely on information from practitioners. In the Mayfield case, this could have involved deriving information from international terror cases involving a bomb in which evidence was retrieved from a textured surface, such as a bag. An advantage to this is that each examiner would have started their analysis with the same prior information. One limitation; however, would be in which information in the database could be used to inform prior belief. It has the potential to lead to scenarios in which examiners are given information which could introduce bias into the examination. This type of *contextual bias* leads examiners to make decisions that confirm their prior

beliefs, a type of *confirmation bias* [17]. Confirmation bias leads an examiner to find information consistent with their prior-held beliefs, potentially leading the examiner to base their conclusion on evidence that is of poor quality or disputable or to miss information that could possibly contradict those prior beliefs [17]. In Mayfield's case, the original FBI examiners placed over-sized emphasis on Level III details which are often considered too variable to rely on for identification. In the process, they discounted other Level III details showing discrepancies between Mayfield's print and the fingermark [42], which in part lead to the false identification.

While systematically deriving prior information could have helped reduce some of the factors that contributed to the error, there is still another substantial element in this investigation to consider: the relationship in the weight of evidence between Mayfield and Daoud.

6.3 Dependencies in the Weight of Evidence for Multiple Candidates

6.3.1 Two Candidates

We now consider the challenge of interpreting the weight of evidence in a way that takes into account the fact that multiple candidates are examined. Because Mayfield was the fourth ranked candidate produced by IAFIS, there were at least three other candidates examined by the original FBI examiners. Considerations for interpreting evidence when multiple candidates are examined (for example, the list of candidates derived by IAFIS) will be discussed in section 6.3.2. Our current focus, then, is on the relationship in the weight of evidence derived from the examinations of two candidates, such as in the case of Mayfield and Daoud.

Only one unique source can be the one that produced a particular fingermark, so there exists some dependency in the weight of evidence when multiple candidates are considered as potential sources. Given the model in (1.2) to quantify the weight of evidence for each candidate, there are two places in which we can account for this dependency: in the prior odds, $\frac{P(H_0)}{P(H_1)}$, or the Bayes factor, $\frac{P(X,Y|H_0)}{P(X,Y|H_1)}$. In the previous section, we discussed the limitations of updating the prior odds as new information becomes available. Given these limitations, we propose a framework that accounts

for the dependencies in the Bayes factor. This approach provides a structure in which changes in the weight of evidence when new candidates are introduced are accounted for in the quantification and interpretation of evidence, rather than depending on examiners to change their interpretation and decision criteria. It also better aligns the real-world scenario in which the introduction of an additional candidate does not occur until after the print from the original person of interest has been examined, thus accounting for dependencies in a way in which the data from original candidate does not change even when the additional candidate is introduced. This approach is more objective, thus working towards a goal for improving forensic sciences by yielding "greater accuracy, repeatability, and reliability" [21] in how evidence is interpreted.

The key difference our proposed framework and other multiplicity methods is that the primary objective of our framework is not to draw conclusions but rather provide an interpretation for the weight of evidence in support of H_0. For example, frequentist techniques such as the Bonferonni correction, are aimed at reducing the Type I error rate when multiple comparisons are made in hypothesis testing. Moreover, these methods assume H_0 is true and the data is statistically significant if it provides sufficient evidence against H_0. In contrast, we do not impose a notion of truth on either hypothesis and thus will interpret the results as evidence in support of H_0.

We demonstrate our framework using a univariate scenario; however, it can be easily extended to the multivariate case. An extension to the multivariate case would only affect the continuous portion of the model set up, not the discrete portion on which we will impose a constraint. Suppose we know the measurement of an unknown source. We call the measurement Y. If we collected multiple measurements from the source using the same instrument, we expect there to be variability in the values, since every instrument has some amount of measurement error. Therefore, we say that repeated measurements follow some distribution centered at the true measurement θ_Y with a variance σ_Y^2. For this demonstration, we will use $Y \sim G(\theta_Y, \sigma_Y^2)$. We identify two candidates and measure each to determine if they are the source that was originally measured. Let X_1 and X_2 be the measurements of the two candidates, such that $X_1 \sim G(\theta_{X_1}, \sigma_X^2)$ and $X_2 \sim G(\theta_{X_2}, \sigma_X^2)$. To account for the variability in measurements between sources, the values θ_{X_1} and θ_{X_2} are each generated from a Gaussian distribution centered at θ_0 with variance σ_0^2.

When the two candidates are examined, there are four possible outcomes: (1) Candidate 1 is the source of the mark, (2) Candidate 2 is the source, (3) neither Candidate 1 nor Candidate 2 are

the source, (4) both Candidate 1 and Candidate 2 are sources of the mark. We can represent each scenario as a separate model:

$$\begin{aligned} M_1 &: \theta_{X_1} = \theta_Y \text{ and } \theta_{X_2} \neq \theta_Y \\ M_2 &: \theta_{X_1} \neq \theta_Y \text{ and } \theta_{X_2} = \theta_Y \\ M_3 &: \theta_{X_1} \neq \theta_Y \text{ and } \theta_{X_2} \neq \theta_Y \\ M_4 &: \theta_{X_1} = \theta_Y \text{ and } \theta_{X_2} = \theta_Y \end{aligned} \tag{6.1}$$

We quantify the weight evidence in support of model M_k representing the true outcome using a Bayes factor [32]. Given $H_0 : M_k$ is the true model vs. $H_1 : M_k$ is not the true model, the Bayes factor

$$BF_{01} = \frac{\pi_k(X_1, X_2)}{1 - \pi_k(X_1, X_2)} = \frac{P(M_k|X_1, X_2)/(1 - P(M_k|X_1, X_2))}{P(M_k)/(1 - P(M_k))} \tag{6.2}$$

such that $\frac{P(M_k)}{1-P(M_k)}$ are the prior odds and $\frac{P(M_k|X_1,X_2)}{1-P(M_k|X_1,X_2)}$ are the posterior odds. From [32], the posterior probability for model, M_k, $k = 1, \ldots, 4$ is

$$P(M_k|X_1, X_2) = \left[\sum_{t=1}^{4} \frac{P(M_t)\pi_t(X_1, X_2)}{P(M_k)\pi_k(X_1, X_2)} \right]^{-1} \tag{6.3}$$

and the marginal density of the data under model M_t is

$$\pi_t(X_1, X_2) = \int P(X_1, X_2|\theta_{X_1}, \theta_{X_2}, M_t) P(\theta_{X_1}, \theta_{X_2}|M_t) d\theta_{X_1} d\theta_{X_2} \tag{6.4}$$

We can interpret BF_{0k} in a similar way as the likelihood ratio that is often used in the forensic science literature as described in [6].

In our model framework, we propose introducing a constraint on (6.1) by setting $P(M_4) = 0$. In order to understand how this constraint affects the interpretation of the weight of evidence, we begin with the interpretation of evidence under (6.1) with no constraints.

Given the two candidates under examination, suppose X_i $(i = 1, 2)$ is the measurement from the primary candidate of interest. We interpret the weight of evidence in support of the i^{th} candidate being the source of the mark, $H_0 : \theta_{Xi} = \theta_Y$. Suppose $i = 1$, then using the models in (6.1) the weight of evidence in support of H_0 is

$$BF_{01} = \frac{P(X_1, X_2|H_0)}{P(X_1, X_2|H_1)} = \frac{\pi_1(X_1, X_2) + \pi_4(X_1, X_2)}{\pi_2(X_1, X_2) + \pi_3(X_1, X_2)} \tag{6.5}$$

and to interpret BF_{01}, we use the scale of evidence in Table 2.1.

Model (6.5) is the weight of evidence under the "No Constraints" condition in which there are no restrictions imposed on the model structure in (6.1), in other words $P(M_k) > 0$ for all $M_1, \ldots, M_4$. This is equivalent to the likelihood ratios described in [6, 40, 41] where the weight of evidence in support of each candidate is interpreted independently of any other candidates under consideration.

Because we assume that Candidate 1 and Candidate 2 are two unique sources, it is not possible for both to be the source that produced the measurement Y. We represent this constraint in our model framework (6.1) by setting $P(M_4) = 0$. $P(M_k), k = 1, 2, 3$ are derived as before. This is similar to Neath and Cavanaugh [39]'s approach for eliminating impossible outcomes in the multiple comparisons problem. Under this constraint, $P(H_0|X_1, X_2) = P(M_1|X_1, X_2)$ and the weight of evidence in support of H_0 is now

$$BF_{01} = \frac{\pi_1(X_1, X_2)}{\pi_2(X_1, X_2) + \pi_3(X_1, X_2)} \tag{6.6}$$

Table 6.1 shows the weight of evidence for two candidates interpreted under the model framework with no constraints and our proposed framework with the constraint. Each column shows the the strength of evidence in support of the i^{th} candidate being the source of the mark calculated as $2 \log BF$. Under the "No Constraints" condition, the Bayes factor BF is calculated using (6.5), and BF is calculated using (6.6) under the "Constraints" condition. Each row of the table represents a different scenario in which two candidates are examined.

Scenarios 1 and 20 illustrate two interesting phenomena that occur under the constraint. In the first scenario, the evidence for X_2 is strongly against X_2 being the source of the mark when it is interpreted independently. When this evidence is interpreted accounting for the evidence from X_1, a very strong candidate, the evidence against X_2 becomes stronger while there is little impact on the interpretation of the evidence for X_1. Scenario 20 illustrates how the interpretation of evidence changes when both candidates under consideration are strong (similar to the case of Mayfield and Daoud). Though the evidence for each candidate is independently interpreted as strong, the evidence

	No Constraints		Constraints	
Scenario	X_1 Strength	X_2 Strength	X_1 Strength	X_2 Strength
1	6.000	-5.889	5.897	-11.986
2	6.000	-4.394	5.789	-10.491
3	6.000	-3.469	5.675	-9.566
4	6.000	-2.773	5.554	-8.870
5	6.000	-2.197	5.424	-8.294
6	6.000	-1.695	5.286	-7.792
7	6.000	-1.238	5.138	-7.335
8	6.000	-0.811	4.978	-6.908
9	6.000	-0.401	4.804	-6.498
10	6.000	0.000	4.614	-6.097
11	6.000	0.401	4.403	-5.696
12	6.000	0.811	4.167	-5.286
13	6.000	1.238	3.900	-4.859
14	6.000	1.695	3.592	-4.402
15	6.000	2.197	3.227	-3.900
16	6.000	2.773	2.781	-3.324
17	6.000	3.469	2.206	-2.628
18	6.000	4.394	1.395	-1.703
19	6.000	5.889	0.008	-0.208
20	6.000	6.000	-0.097	-0.097

Table 6.1: X_i Strength $= 2 \log BF_{i0}$. Each row of the table represents a new scenario in which two candidates are examined. Columns 2 and 3 ("No Constraints") show the $2 \log BF_{i0}$ when the evidence for each candidate in the scenario is interpreted independently. Columns 4 and 5 ("Constraints") show the interpretation of the evidence when the candidates are examined accounting for the existence of the other candidate.

for both candidates is interpreted as weakly against under the constrained framework. This last scenario is especially important to consider when thinking about the interpretation of evidence in cases in which the candidates examined are selected based on a match score such as AFIS. In these cases, all the candidates tend to closely match the fingermark and thus would each have fingerprint evidence that is independently interpreted as strong.

6.3.2 Three Candidates

Examiners use programs such as AFIS to retrieve a candidate list of potential matches for a fingermark. Dror and Mnookin [16] and Busey et al. [12] discuss this process extensively pointing out how the size of the database increases the potential for retrieving non-matches that are very similar

to the fingermark. As we observed in the previous section, examining multiple candidates that closely match the mark has a significant effect on the interpretation of evidence for each candidate. We now extend our analysis to the case in which three candidates are examined, focusing on the effect of examining multiple candidates that are similar to the fingermark in question.

Following the two candidate scenario, we assume that at most one unique candidate can be the source of the mark. The possible models are in (6.7).

$$\begin{array}{lll} M_1 : \theta_{X_1} = \theta_Y & \theta_{X_2} \neq \theta_Y & \theta_{X_3} \neq \theta_Y \\ M_2 : \theta_{X_1} \neq \theta_Y & \theta_{X_2} = \theta_Y & \theta_{X_3} \neq \theta_Y \\ M_3 : \theta_{X_1} \neq \theta_Y & \theta_{X_2} \neq \theta_Y & \theta_3 = \theta_Y \\ M_4 : \theta_{X_1} \neq \theta_Y & \theta_{X_2} \neq \theta_Y & \theta_3 \neq \theta_Y \end{array} \tag{6.7}$$

Table 6.2 shows similar phenomena as seen in Table 6.1. The interpretation of evidence for each candidate changes once the constraint is imposed. Additionally, this impact on the interpretation of evidence depends on the strength of the other two candidates. Scenario 16 shows the largest effect on the interpretation of evidence for the candidates being examined. When all the candidates are independently interpreted as close matches, the interpretation of evidence then goes against each candidate under the constrained model. When candidates are generated using a program such as AFIS, they are likely close matches to the fingermark under investigation. Therefore, this scenario in Table 6.2 and scenario 20 in Table 6.1 are ones that are likely to occur in practice. To better understand how the weight of evidence can be interpreted for a list of candidates produced by AFIS, we now generalize our framework to K candidates.

6.3.3 K Candidates

When K candidates are examined, the set of possible models is

	No Constraints			Constraints		
Scenario	X_1 Strength	X_2 Strength	X_3 Strength	X_1 Strength	X_2 Strength	X_3 Strength
1	6.000	-5.889	-5.889	5.800	-11.991	-11.991
2	6.000	-2.197	-5.889	5.347	-8.299	-12.017
3	6.000	-2.197	-2.197	4.978	-8.326	-8.326
4	6.000	0.000	-5.889	4.562	-6.102	-12.079
5	6.000	0.000	-2.197	4.305	-6.128	-8.387
6	6.000	0.000	0.000	3.803	-6.190	-6.190
7	6.000	0.000	2.773	2.416	-6.444	-3.417
8	6.000	2.773	-5.889	2.760	-3.329	-12.333
9	6.000	2.773	-2.197	2.652	-3.356	-8.642
10	6.000	2.773	2.773	1.605	-3.672	-3.672
11	6.000	4.394	-5.889	1.384	-1.708	-12.697
12	6.000	4.394	-2.197	1.329	-1.734	-9.005
13	6.000	4.394	0.000	1.204	-1.795	-6.808
14	6.000	4.394	2.773	0.722	-2.050	-4.035
15	6.000	4.394	4.394	0.111	-2.414	-2.414
16	6.000	6.000	6.000	-1.435	-1.435	-1.435

Table 6.2: X_i Strength $= 2\log BF_{i0}$. Each row of the table represents a new scenario in which three candidates are examined. Columns 2 - 4 ("No Constraints") show the $2\log BF_{i0}$ when the evidence for each candidate in the scenario is interpreted independently. Columns 5-7 ("Constraints") show the interpretation of the evidence when the candidates are examined accounting for the existence of the other two candidates.

$$\begin{aligned} &\text{For } i = 1, \ldots, K, \\ &M_i : \begin{cases} \theta_{X_i} = \theta_Y \\ \theta_{X_j} \neq \theta_Y \quad j = 1, \ldots, K \text{ such that } i \neq j \end{cases} \\ &M_{K+1}\text{: } \theta_{X_1} \neq \theta_Y \quad \theta_{X_2} \neq \theta_Y \quad \ldots \quad \theta_{X_K} \neq \theta_Y \end{aligned} \tag{6.8}$$

When interpreting AFIS scores, the absolute score is not necessarily the most important indicator but rather the score relative to other candidates [12]. Therefore, if the i^{th} candidate is the primary candidate of interest, we will measure the other candidates in relation to Candidate i. Let $r_k = \frac{X_i}{X_k}$ be the ratio between measurements X_i and X_k. Generalizing (6.6) to K candidates, the general

form of BF_{10}, i.e. evidence against $H_0 : \theta_{X_i} = \theta_Y$ is

$$BF_{10} = \exp\left\{ -\frac{1}{2}\left[\frac{(X_i-\theta_0)^2}{\sigma_X^2+\sigma_0^2} - \frac{(X_i-\theta_Y)^2}{\sigma_X^2}\right]\right\} \times \left((\sigma_X^2+\sigma_0^2)^{-\frac{1}{2}} + \sum_{k=1}^{K} \exp\left\{ -\frac{1}{2}\left[\frac{\left(\frac{X_i}{r_k}-\theta_Y\right)^2}{\sigma_X^2} - \frac{\left(\frac{X_i}{r_k}-\theta_0\right)^2}{\sigma_X^2+\sigma_0^2}\right]\right\}\right) \tag{6.9}$$

Suppose there is strong evidence in support of Candidate i being the source of the mark. Using (6.9), we can calculate the evidence in support of $H_0 : \theta_{X_i} = \theta_Y$ as $\frac{1}{BF_{10}}$. In addition to Candidate i, suppose we obtain a list of 15 candidates produced by AFIS. Since AFIS is designed to obtain the sources in the database that most closely matches the fingermark, we let $r_k = [0.70, 0.95]$ for $k = 1, \ldots, 15$. AFIS ranks the candidates in decreasing order of similarity to the mark, so we presume the level of similarity to Candidate i mirrors the AFIS rankings, thus $r_1 > r_2 > \cdots > r_{15}$. Figure 6.3.3 shows how the interpretation of evidence from the i^{th} candidate changes as an each additional candidate is introduced. This is similar to what was observed in sections 6.3.1 and 6.3.2 - even if the initial evidence support of H_0 can be interpreted as strong, this changes substantially if other sources similar to Candidate i are examined. The figure shows that if at least three candidates from a list of AFIS results are examined, any conclusion that Candidate i is the source of the fingermark should be made with caution. 6.5.

6.4 Theoretical Properties

We have illustrated a framework in which the weight of evidence can be interpreted in way that accounts for the evidence from multiple candidates. In this section, we explore more underlying properties the phenomena we have observed in the previous two sections in an effort to provide recommendations for how this framework can be applied in practice.

6.4.1 Consistency of the Bayes Factor

Various authors have written about the asymptotic properties of Bayes factors [26], including their consistency in measuring the weight of evidence. If a true model for the data exists, and it is

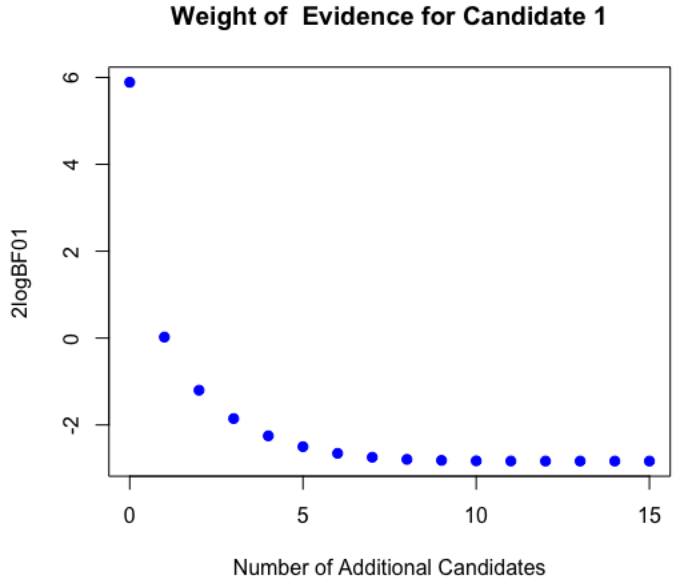

Figure 6-1: Weight of evidence in support of $H_0 : \theta_{X_1} = \theta_Y$ vs. the number of candidates from AFIS considered. The first point is the independent case, in which Candidate 1 is the only candidate examined. For each additional candidate, we assume all higher ranked candidates have also been examined. All of the additional candidates were simulated so that they have a ratio to X_1 between 0.7 and 0.95. We treat Candidate 1 as the closest match to the fingermark Y, therefore it is the first candidate produced by AFIS. The additional candidates are in order of their similarity to Candidate 1.

model M_0, then the Bayes factor is consistent if BF_{01} ∞ as $n \to \infty$. Under our model set up, we show this holds true when we assume independence between candidates. The Bayes factor is

$$BF_{01} = \frac{\left(\frac{n}{2\pi}\right)^{\frac{1}{2}}\left(\frac{1}{2}\right)e^{-\frac{n}{2\sigma_X^2}(\bar{X}-\theta_Y)^2}}{\left(\frac{n}{2\pi}\right)^{\frac{1}{2}}\left(\frac{1}{2}\right)(\sigma_X^2+n\sigma_0^2)^{-\frac{1}{2}}e^{-\frac{n}{2}\left(\bar{X}^2-\frac{\sigma_0^2\bar{X}^2}{\sigma_X^2+n\sigma_0^2}\right)}} = (\sigma_X^2+n\sigma_0^2)^{\frac{1}{2}}e^{-\frac{n}{2}\left(\frac{-2\bar{x}\theta_Y+\theta_Y^2}{\sigma_X^2}+\frac{\sigma_0^2}{\sigma_X^2+n\sigma_0^2}\bar{X}^2\right)} \quad (6.10)$$

As $n \to \infty$, $(\sigma_X^2 + n\sigma_0^2)^{\frac{1}{2}} \to \infty$ and $\frac{\sigma_0^2}{\sigma_X^2+n\sigma_0^2}\bar{X}^2 \to 0$. When M_0 is the true model, $\bar{X} \to \theta_Y$ as $n \to \infty$ by the Law of Large Numbers. Therefore, $e^{-\frac{n}{2}\left(-2\bar{X}\theta_Y+\theta_Y^2+\frac{\sigma_0^2}{n\sigma_0^2+1}\bar{X}^2\right)} \to e^{\frac{n}{2}\theta_Y^2}$ as $n \to \infty$. This quantity approaches ∞ at a much faster rate than $(n\sigma_0^2 + 1)^{\frac{1}{2}}$ approaches zero; therefore, $BF_{01} \to \infty$ as $n \to \infty$ when M_0 is the true model. $BF_{10} = \frac{1}{BF_{01}}$. When M_0 is the true model, we have shown that BF_{01} approaches ∞. Therefore, $BF_{10} = \frac{1}{BF_{01}} \to 0$ as $n \to \infty$. It easily follows that $BF_{10} \to 0$ as $n \to \infty$, since $BF_{10} = \frac{1}{BF_{01}}$.

We generalize (6.9) to have K candidates and thus the Bayes factor for evidence against H_0 is

$$\begin{aligned} BF_{10} = \exp\left\{ -\frac{n_1}{2}\left[\frac{(\bar{X}_1-\theta_0)^2}{\sigma_X^2+\sigma_0^2} - \frac{(\bar{X}_1-\theta_Y)^2}{\sigma_X^2}\right]\right\} \\ \times\left((\sigma_X^2+n\sigma_0^2)^{-\frac{1}{2}} + \sum_{k=1}^{K}\exp\left\{-\frac{n_k}{2}\left[\frac{\left(\frac{\bar{X}_1}{r_k}-\theta_Y\right)^2}{\sigma_X^2} - \frac{(\frac{X_1}{r_k}-\theta_0)^2}{\sigma_X^2+n_k\sigma_0^2}\right]\right\}\right) \end{aligned} \tag{6.11}$$

By the Law of Large Numbers, $(\bar{X}_1-\theta_Y)^2 \to 0$ as $n\to\infty$. Additionally, $(1+n\tau^2)\to\infty$ and $e^{-\frac{n}{2}}\to 0$ as $n\to\infty$. Therefore, $P(X_1,\ldots,X_K|M_1)\to 1$ as $n\to\infty$ and $1-P(X_1,\ldots,X_K|M_1))\to 0$. Similarly, $BF_{10}=\frac{1}{BF_{01}}\to 0$ as $n\to\infty$. Thus, we have shown that the Bayes factor is a consistent estimator of the weight of evidence in support of M_1.

6.4.2 Accounting for Multiple Candidates in the Bayes Factor

Suppose K candidates are examined. Let $\mathbf{X}=X_1,\ldots,X_K$ be numerical summaries of evidence from the K candidates. The i^{th} candidate is the one of primary interest, therefore in this section we will look at the interpretation of the weight of evidence for X_i. Before exploring the properties of the weight of evidence, we define some quantities that will be used throughout.

Suppose there are only two models $M_i:\theta_{X_i}=\theta_Y$ and $M_0:\theta_{X_i}\neq\theta_Y$, then the Bayes factor for X_i is

$$BF_{i0} = \frac{P(M_i|\mathbf{X})/P(M_0|\mathbf{X})}{P(M_i)/P(M_0)} \tag{6.12}$$

As seen before, $2\log BF_{i0}$ is the interpretation of evidence for X_i when the evidence from each candidate is interpreted independently. If we suppose there are $K+1$ models $M_0,\ldots,M_K$, then the Bayes factor for model M_i is

$$BF_i = \frac{P(M_i|\mathbf{X})/(1-P(M_i|\mathbf{X}))}{P(M_i)/(1-P(M_i))} \tag{6.13}$$

In the exploration of BF_i, the form of $P(M_i|\mathbf{X})$ in (6.14) will be used.

$$
\begin{aligned}
P(M_i|\mathbf{X}) &= \left[1 + \sum_{\substack{k=0 \\ k\neq i}}^{K} \frac{P(M_k|\mathbf{X})}{P(M_i|\mathbf{X})}\right]^{-1} \\
&= \left[1 + \frac{P(M_0|\mathbf{X})}{P(M_i|\mathbf{X})} + \sum_{\substack{k=1 \\ k\neq i}}^{K} \frac{P(M_k|\mathbf{X})}{P(M_i|\mathbf{X})}\right]^{-1} \\
&= \left[1 + \frac{P(M_0|\mathbf{X})}{P(M_i|\mathbf{X})} + \sum_{\substack{k=1 \\ k\neq i}}^{K} \frac{P(M_k|\mathbf{X})/P(M_0|\mathbf{X})}{P(M_i|\mathbf{X})/P(M_0|\mathbf{X})}\right]^{-1} \\
&= \left[1 + \frac{P(M_0|\mathbf{X})}{P(M_i|\mathbf{X})}\left(1 + \sum_{\substack{k=1 \\ k\neq i}}^{K} \frac{P(M_k|\mathbf{X})}{P(M_i|\mathbf{X})}\right)\right]^{-1}
\end{aligned}
\tag{6.14}
$$

Using (6.12) - (6.14), we will decompose BF_i into two components: BF_{i0} and DF_i. BF_{i0} quantifies the weight of evidence for X_i as seen in (6.12), and DF_i quantifies the effect of evidence from the additional $K-1$ candidates under examination.

Starting with the numerator of (6.13), we can write the posterior odds as

$$
\begin{aligned}
\frac{P(M_i|\mathbf{X})}{(1-P(M_i|\mathbf{X}))} &= \frac{P(M_i|\mathbf{X})}{\left[P(M_i|\mathbf{X}) + P(M_0|\mathbf{X}) + \sum\limits_{\substack{k=1 \\ k\neq i}}^{K} P(M_k|\mathbf{X})\right] - P(M_i|\mathbf{X})} \\
&= \frac{P(M_i|\mathbf{X})}{P(M_0|\mathbf{X}) + \sum\limits_{\substack{k=1 \\ k\neq i}}^{K} P(M_k|\mathbf{X})} \\
&= \frac{P(M_i|\mathbf{X})/P(M_0|\mathbf{X})}{1 + \sum\limits_{\substack{k=1 \\ k\neq i}}^{K} P(M_k|\mathbf{X})/P(M_0|\mathbf{X})} \\
&= \frac{[P(M_i)/P(M_0)]BF_{i0}}{1 + \sum\limits_{\substack{k=1 \\ k\neq i}}^{K} \frac{P(M_k)}{P(M_0)} BF_{k0}}
\end{aligned}
\tag{6.15}
$$

The prior odds in the denominator of (6.13) can be written as

$$\begin{aligned}\frac{P(M_i)}{(1-P(M_i))} &= \frac{P(M_i)}{\left[P(M_i)+P(M_0)+\sum\limits_{\substack{k=1\\k\neq i}}^{K}P(M_k)\right]-P(M_i)}\\ &= \frac{P(M_i)}{P(M_0)+\sum\limits_{\substack{k=1\\k\neq i}}^{K}P(M_k)}\\ &= \frac{P(M_i)/P(M_0)}{1+\sum\limits_{\substack{k=1\\k\neq i}}^{K}P(M_k)/P(M_0)}\end{aligned} \tag{6.16}$$

Thus from (6.15) and (6.16), BF_i is equivalent to the following:

$$\begin{aligned}BF_i &= \frac{P(M_i|\mathbf{X})/(1-P(M_i|\mathbf{X}))}{P(M_i)/(1-P(M_i))}\\ &= \frac{[P(M_i)/P(M_0)]BF_{i0}\left[1+\sum\limits_{\substack{k=1\\k\neq i}}^{K}\frac{P(M_k)}{P(M_0)}BF_{k0}\right]^{-1}}{[P(M_i)/P(M_0)]\left[1+\sum\limits_{\substack{k=1\\k\neq i}}^{K}\frac{P(M_k)}{P(M_0)}\right]^{-1}}\\ &= BF_{i0}\left[1+\sum_{\substack{k=1\\k\neq i}}^{K}\frac{P(M_k)}{P(M_0)}\right]\Bigg/\left[1+\sum_{\substack{k=1\\k\neq i}}^{K}\frac{P(M_k)}{P(M_0)}BF_{k0}\right]\\ &= BF_{i0}\times DF_i\end{aligned} \tag{6.17}$$

such that

$$DF_i = \left[1+\sum_{\substack{k=1\\k\neq i}}^{K}\frac{P(M_k)}{P(M_0)}\right]\Bigg/\left[1+\sum_{\substack{k=1\\k\neq i}}^{K}\frac{P(M_k)}{P(M_0)}BF_{k0}\right] \tag{6.18}$$

By decomposing BF_i as shown in (6.17), the strength of evidence from X_i can be interpreted in a way that precisely indicates how closely X_i matches Y, the measurements from the fingermark in question, and the effect of evidence from other candidates as quantified in the *dependency factor* in (6.18).

The dependency factor measures the relationship between the prior information and the strength of evidence of the additional $K-1$ candidates. Therefore, it is worth noting the interpretation of

DF_i may be understood as the strength of the effect on BF_i from the additional candidates. For example, if $DF_i < -6$, then the presence of the additional candidates strongly weakens the strength of evidence for X_i, and the BF_{i0} is not a reliable interpretation of the evidence from X_i. It is not always the case that DF_i weakens BF_i. If the additional $K-1$ candidates have weak fingerprint evidence but were examined due to strong prior information from non-fingerprint evidence, then $DF_i > 1$ and BF_i increases thus strengthening the evidence for X_i.

From (6.17), the strength of X_i, i.e. the weight of evidence for Candidate i, can be interpreted as the following $2\log(BF_i) = 2\log BFi0 + 2\log DF_i.$, such that each component is interpreted using the scale in Table 2.1. Table 6.3 shows the decomposition of BF_1 from Table 6.2.

Scenario	BF_{20}	BF_{30}	BF_1	BF_{10}	DF_1
1	-5.889	-5.889	5.800	6.000	-0.200
2	-2.197	-5.889	5.347	6.000	-0.653
3	-2.197	-2.197	4.978	6.000	-1.022
4	0.000	-5.889	4.562	6.000	-1.438
5	0.000	-2.197	4.305	6.000	-1.695
6	0.000	0.000	3.803	6.000	-2.197
7	0.000	2.773	2.416	6.000	-3.584
8	2.773	-5.889	2.760	6.000	-3.240
9	2.773	-2.197	2.652	6.000	-3.348
10	2.773	2.773	1.605	6.000	-4.394
11	4.394	-5.889	1.384	6.000	-4.616
12	4.394	-2.197	1.329	6.000	-4.671
13	4.394	0.000	1.204	6.000	-4.796
14	4.394	2.773	0.722	6.000	-5.278
15	4.394	4.394	0.111	6.000	-5.889
16	6.000	6.000	-1.435	6.000	-7.435

Table 6.3: Each row represents a different scenario in which X_2 and X_3, the evidence for the two additional candidates, are of varying strength when interpreted independently. This is represented by BF_{20} and BF_{30} in columns two and three. The weight of evidence for X_1 is broken into components BF_{10}, the weight of evidence from X_1 that doesn't account for the other candidates, and DF_1, the dependency factor that quantifies the effect of candidates two and three.

6.4.3 Candidate Selection Criteria

The next area of interest is in understanding how BF_i changes as the number of candidates K increases. Suppose N candidates (N large) are selected using a database search such as AFIS. Out of the N candidates, the examiner chooses to examine only the K candidates who meet some threshold, ϵ_K, that is based on prior information largely determined by non-fingerprint evidence. Suppose the K candidates are arranged in decreasing order based on $B_{k0}, k = 1, \ldots, K$, where B_{k0} is the strength of X_k not accounting for other candidates. Using this ordering, the Bayes factor for the k^{th} strongest candidate be summarized using a parametric form

$$B_k = BF_{10}e^{-a(k-1)} \approx BF_{k0} \tag{6.19}$$

such that $a < 1$. From (6.19), the total weight of evidence from the K candidates can be approximated as the following

$$\sum_{k=1}^{K} BF_{k0} \approx \int_{1}^{K} B_k dt \tag{6.20}$$

If $\int_{1}^{\infty} B_k dt$ is finite, then it can be approximated by an infinite sum (shown in **??**) that quantifies the weight of evidence including candidates beyond those that meet the original threshold.

$$\lim_{K\to\infty} \sum_{k=1}^{K} BF_{k0} \approx \int_{1}^{\infty} B_k dt \tag{6.21}$$

We now use (6.21) to explore more precisely how the evidence changes as the threshold for prior information changes. Given each $X_k \sim G(\theta_{X_k}, \sigma_X^2)$, $Y \sim G(\theta_Y, \sigma_Y^2)$, and $\theta_k \sim G(\theta_0, \sigma_0^2)$, we show below what happens as K increases, i.e. as the threshold of criteria for which new candidates can be considered becomes more lax.

The weight of evidence <u>against</u> $M_i : \theta_{X_i} = \theta_Y$ is

$$BF_{0i} = \exp\left\{ -\frac{1}{2}\left[\frac{(X_i-\theta_0)^2}{\sigma_X^2+\sigma_0^2} - \frac{(X_i-\theta_Y)^2}{\sigma_X^2}\right]\right\} \times \left((\sigma_X^2+\sigma_0^2)^{-\frac{1}{2}} + \sum_{k=1}^{K} \exp\left\{ -\frac{1}{2}\left[\frac{(X_k-\theta_Y)^2}{\sigma_X^2} - \frac{(X_k-\theta_0)^2}{\sigma_X^2+\sigma_0^2}\right]\right\}\right) \tag{6.22}$$

such that X_k is the numerical summary of the k^{th} additional candidate. This is an example of the Bayes factor in (6.13) that accounts for the dependencies between models. K increases as $\epsilon \to 0$, thus more candidates are examined. Moreover, since the threshold ϵ_K is based on prior information that includes some fingerprint evidence (such as AFIS scores), N increases as $\epsilon_K \to 0$. Therefore, we explore theoretical properties under the scenario in which $\epsilon_K \to 0$, $N \to \infty$, and $K \to \infty$.

Property 1. *As $\epsilon_K \to 0$, the weight of evidence in support of M_i approaches 0.*

Proof. To prove the property, we will show $BF_{0i} \to \infty$ as $\epsilon_K \to 0$. To do so, it is sufficient to determine

$$\lim_{K\to\infty} \sum_{k=1}^{K} \exp\left\{ -\frac{1}{2}\left[\frac{(X_k-\theta_Y)^2}{\sigma_X^2} - \frac{(X_k-\theta_0)^2}{\sigma_X^2+\sigma_0^2}\right]\right\}$$

Let

$$f(X_k) = \exp\left\{ -\frac{1}{2}\left[\frac{(X_k-\theta_Y)^2}{\sigma_X^2} - \frac{(X_k-\theta_0)^2}{\sigma_X^2+\sigma_0^2}\right]\right\}$$

.

- **Case 1:** $f(X_k) > 1$. This occurs when $\left(\frac{(X_k-\theta_Y)^2}{\sigma_X^2} - \frac{(X_k-\theta_0)^2}{\sigma_X^2+\sigma_0^2}\right) < 0$. This could be true in the scenario in which X_k is close to θ_Y, i.e. there is strong evidence in support of X_k being the source of the mark.
- **Case 2:** $0 < f(X_k) < 1$. This occurs when $\left(\frac{(X_k-\theta_Y)^2}{\sigma_X^2} - \frac{(X_k-\theta_0)^2}{\sigma_X^2+\sigma_0^2}\right) > 0$. This could be true in the scenario in which X_k is far from θ_Y, i.e. there is strong evidence against X_k being the source of the mark.

In both cases, $f(X_k) > 0$, and in Case 1, $f(X_k)$ is unbounded. Therefore,

$$\lim_{K\to\infty} \sum_{k=1}^{K} \exp\left\{ -\frac{1}{2}\left[\frac{(X_k-\theta_Y)^2}{\sigma_X^2} - \frac{(X_k-\theta_0)^2}{\sigma_X^2+\sigma_0^2}\right]\right\} = \infty$$

. □

Using Property 1, we conclude that $DF_i \to 0$ and $2\log DF_i \to -\infty$ as more candidates are examined. The practical implication of this is when the criteria for examination becomes increasingly lax, the interpretation of evidence for Candidate i should only be made in the context of all the other candidates examined.

6.5 Guidelines for Interpreting the Weight of Evidence

We now turn from our exploration in the previous section to practical guidelines for examiners in regards to analyzing evidence from multiple candidates.

The first guideline is that examiners should report $2\log BF_i$ along with their stated conclusion about whether or not Candidate i is the source that produced the fingermark in question. Dror and Mnookin [16] suggest that examiners include probative value when reporting their conclusions to express some level of uncertainty in the results. By stating $2\log BF_i$ along with the decision, examiners are providing an indication of the strength of evidence that was used to make the decision. Additionally, because BF_i is calculated by taking into account the total number of candidates examined, this statistic provides a more holistic assessment of the uncertainty in the examiner's decision than the likelihood ratio statistics in [40], [41], and [6] (see section 1.2 for a discussion of these statistics).

Building upon the first guideline, in addition to reporting $2\log BF_i$, examiners should report each of its components - $2\log BF_{i0}$ and $2\log DF_i$. The statistic $2\log BF_{i0}$ provides a measure of strength of the evidence from Candidate i that was used by the examiner to make the decision. This measure is necessary, because if the fingerprint evidence is used as part of a larger case, it will be important to understand how strong that fingerprint evidence is on its own without taking into account all the other candidates examined. As a complement to that, reporting $2\log DF_i$ is important to provide a clear measure of the strength of fingerprint evidence from the other candidates. This statistic also takes accounts the issues with multiplicity that arise when multiple candidates are examined. Dror and Mnookin [16] state that examiners should adjust their decision criteria when there are multiple candidates to account for these dependencies (similar to a Bonferroni correction).

Reporting $2\log DF_i$ provides a way to account for these dependencies using one standard quantity rather than changing decision-making criteria as the number of candidates changes.

There are multiple scales to consider for interpreting the Bayes factors in Section 6.4.2. Jeffreys [30] introduced a scale for interpreting the Bayes factor in which $BF_{01} > 100$ is considered decisive evidence in support of H_0; however, Evett [20] argued that BF_{01} should be at least 1,000 in order to be considered decisive in the forensic context. We recommend interpreting the Bayes factors as $2\log BF$, since they can be interpreted on the same scale as some likelihood ratio test statistics [32], such as $-2(\log L(\theta|H_0) - \log L(\theta|H_1)$ where $L(\theta|H_i)$ is the likelihood function under hypothesis H_i. Given the development of forensic methods based on likelihood ratios [40, 41, 6], having the common scale can provide a link to the methods we've proposed to existing ones. At this point, we recommend examiners use Kass and Raftery [32] (shown in Table 2.1) as a guide for interpreting the Bayes factors, since it is an established scale that can provide common language between the statistics and forensic communities. Since this scale was not originally developed to be used in a forensic context, the accuracy, discrimination, and calibration performance characteristics from Section 5.5.4 could be used to modify the scale specifically for forensic evidence. Ultimately, a common scale should be established in the forensic community, so that evidence can be interpreted in a consistent way.

In section 6.4, we showed the effect lowering the threshold of prior information required to seriously consider a candidate for examination. While this idea of shrinking the threshold for viable prior information to zero is not sensible in practice, it does provide insight that can lead to a value of ϵ_K that sufficiently accounts for the relevant prior information but is not so inclusive that the candidate list is too large to be manageable in practice. The threshold for ϵ_K should be based on two components: fingerprint evidence and non-fingerprint evidence. The fingerprint evidence should be based on a match score, such as the score produced by AFIS. Often, this is the only metric used to determine the candidates that will be examined, which increases the chance of an erroneous decision. This is due to the fact that candidates selected by AFIS are those who fingerprints most closely match the fingermark and thus many are close non-matches [16].

To reduce this chance of erroneous conclusions, in addition to fingerprint evidence, non-fingerprint evidence should be included in determining the threshold ϵ_K. Past casework can examined to assess the strength of the non-fingerprint evidence from cases that are similar to the one in question.

Since this non-fingerprint evidence largely relies on other investigative work, it is recommended that this assessment be done within each forensic agency in order to get a true measure of the prior information available for that agency. For example, it would be expected that the FBI has more access to non-fingerprint evidence than a local police station; therefore, the FBI may have a different threshold for ϵ_K than the local police station.

Given the effect considering multiple candidates could have on the interpretation of evidence from the candidate of interest, there is the potential for examiners to "game the system" and examine as few candidates as possible. There are practical constraints that limit the number of candidates an examiner can analyze; however, we should eliminate situations in which an examiner exhibits confirmation bias and only considers one or very few candidates in order to interpret evidence in way that suits a prior belief about the primary candidate of interest. By using a threshold of prior information,ϵ_K, to generate the candidate list, there is less subjectivity on the part of the examiner in determining which candidates to examine.

Finally, based on the insights from the demonstration in Chapter 5, examiners should consider multiple models when quantifying the weight of evidence. Using the performance characteristics in Section 5.5.4, examiners can determine which model most accurately quantifies the weight of evidence.

Chapter 7

Conclusion and Future Research

7.1 Conclusion

Using Lindley [33] as a guide, we have proposed an objective framework for obtaining prior information and interpreting the strength of fingerprint evidence. In Chapter 2, we presented a theoretical framework for a fully-resourced database that can be used for statistical research and obtaining prior information. Through the use of simulated results, we showed how an ideal, fully-resourced database can be used to understand the instrument-to-instrument variability among similar instruments that can not be understood in a sub-resourced database. We extended this framework to multivariate measurements in Chapter 3, so it can be used with a wide variety of metrics used to convert fingerprint images into numerical summaries of data.

In Chapter 4, we used the crime scene investigation processes to explore the potential causes of variability that exist between multiple prints and marks produced by the same source. We determined the categories of factors that are controllable and could be accounted for in a database versus the uncontrollable factors that are included in the measurement error. Using these factors, we provided recommendations for the definition of "instrument" based on the investigator expertise and qualities of the surface such as texture and shape. We recommended the use of broad categories when defining "instrument" in order to keep the database manageable and widely applicable.

We combined the results from chapters 2 - 4 in a demonstration using a database created from

simulated fingerprints. Using the shape metric from [41], we quantified the weight of evidence that a print and mark were produced by the same source using information from a sub-resourced and fully-resourced database. Through this demonstration, we showed the value of using a fully-resourced database to quantify the weight of evidence along with the flexibility of our framework. We also demonstrated a solution to the practical challenge of transforming different types of data to fit into a multivariate Gaussian framework.

In Chapter 6, we closely examined the interpretation of the weight of evidence and the dependencies in the weight of evidence that exist when multiple candidates are considered the potential source of a mark. Using a Bayesian approach, we proposed accounting for these dependencies by imposing the realistic constraint on the model structure that at most on candidate can be the source of a mark. We explored some of the properties of this framework using the case of two candidates, and then generalized it to a scenario with K candidates. After discussing some of the theoretical properties that arise as a result of these dependencies, we proposed guidelines regarding how examiners can carefully account for the changes in the weight and interpretation of evidence as new candidates are introduced. The recommendations are written primarily for the scenario in which a latent print examiner obtains a list of candidates from a database such as AFIS.

7.2 Future Research

There are future directions to explore based on the results from this research. The first area of future research is to extend the framework presented in chapters 2 and 3 beyond the Gaussian data context. For example, minutiae type (ridge ending or bifurcation) is often included in a data summary, so the framework could be extended to the context of statistical inference using the multinomial model. Angular data is also commonly used in numerical summaries of fingerprints, so this framework could also be formulated to better accommodate inference for data that follow von-Mises distributions.

In regards to the dependencies in the weight of evidence, there are more questions to be explored. In our analysis, we assumed that each candidate was unique; however, this may not necessarily be true (for example, candidates selected from a database containing only fingermarks). New model structures can be explored that impose constraints in a systematic way that allow for such

possibilities. Understanding this model structure could be useful in the context of analyzing data produced by "latent-to-latent" searches in which the original source of the mark in the database is not necessarily known.

The "latent-to-latent" phenomenon could be extended to the scenario in which multiple fingermarks with unknown sources are recovered. Clustering could be used to group the fingermarks in order to determine whether any of the marks were produced by the same source.

Finally, this research has been conducted in the context of examining fingerprint evidence; however, the ideas from this work can be more widely applied to other areas of forensic evidence, specifically other areas of pattern evidence. An immediate step in this direction would be in exploring the feasibility of our approach in other forensic contexts, such as shoe prints and tool marks, given the data generating techniques in these areas and sources of variability in the impressions from shoes and tools, respectively.

Appendix A

MINDTCT

A.1 Overview

To translate the fingerprint images into data that can be used for analysis, we use the Biometric Image Software (NBIS) developed by NIST. NBIS is a free and open-source software that performs a variety of functions such as fingerprint pattern classification, feature detection and image quality assessment, and image quality assessment. For the purpose of this project, we are most interested in feature detection, so we'll focus on the package `MINDTCT`, whose main task is to detect minutiae.

`MINDTCT` reads in an image of a fingerprint and outputs files containing quantified information about the print including an assessment of fingerprint quality, coordinates and orientations of minutiae, and a ridge flow map. Once a fingerprint image is read in, the program takes the following general steps:

$$\textbf{Assess Image Quality} \Rightarrow \textbf{Binarize Image} \Rightarrow \textbf{Detect Minutiae}$$
$$\Rightarrow \textbf{Remove False Minutiae} \Rightarrow \textbf{Output Results}$$

Since image quality is important for detecting minutiae, we describe this element of the program in detail. We guide the reader to the documentation ([54]) for the remaining details.

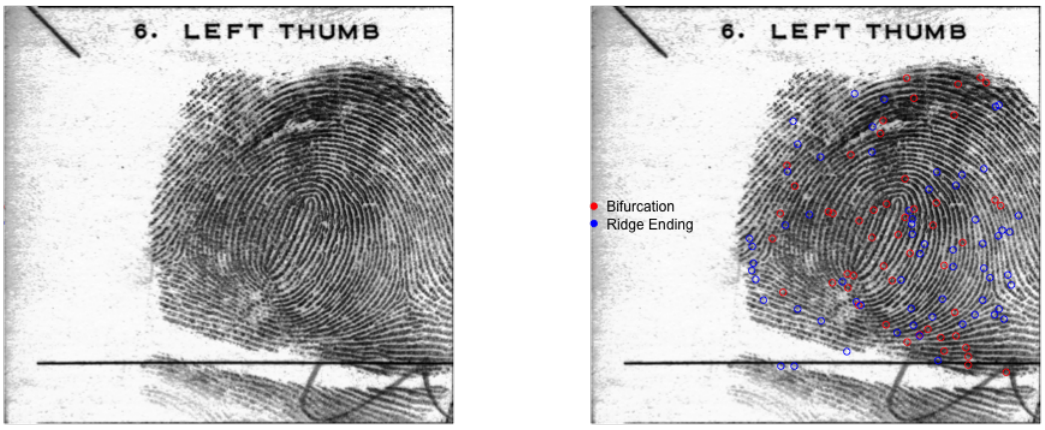

Figure A-1: Left thumbprint classified as whorl pattern produced by a male subject. Image input into MINDTCT (left); minutiae detected by MINDTCT (right)

.

A.2 Assessing Image Quality

Immediately after reading in a fingerprint image, MINDTCT assesses the quality of the image. The image quality assessment helps the user determine the reliability of the feature detection results. The image quality measure is also used to calculate a minutiae quality score. MINDTCT produces an image quality map that provides a quality for each block of pixels in the fingerprint image. The ratings range from "0" to "4", with "0" being the lowest quality and "4" the highest. The rating is computed using an algorithm that considers the ridge flow direction, color contrast, and curvature of the ridges in a fingerprint image.

Direction Map

Clearly locating the ridges in a fingerprint image is an important first step in detecting minutiae. Low ridge flow in a block of the image typically indicates that part of the image is part of the background or is too low quality to reliably detect minutiae. To detect the ridges in the image, MINDTCT produces a direction map that indicates the direction of the ridges in an 8-pixel $\times$ 8-pixel block that exists in a 24-pixel $\times$ 24-pixel window. The window is shifted as each block is assessed

to ensure that there is a smooth ridge flow pattern once all of the blocks have been analyzed. Additionally, all pixels in the same block are given the same ridge flow direction.

To determine the direction for a given block, the surrounding window is incrementally rotated 16 times using a Discrete Fourier Transform (each rotation is about 11.25°). At each orientation, the pixels along each row are summed and the vectors are convolved onto four waveforms that each have different frequencies. A measure consisting of the sine and cosine values of each waveform are added together to create a single measure of how well the orientation fits the waveforms. Once this value is calculated for each orientation, the orientation with the best fit, i.e. the highest measure, is the ridge flow direction for that block.

Once the program makes a determination of the ridge flow for a specified block, the ridge flow is captured in a separate copy of the original fingerprint image. The program repeats this process for each block in a way such that the windows highly overlap thus ultimately creating a smooth direction map of the ridge flow patterns.

Contrast Map

Areas in the fingerprint image with low color contrast are often the background or parts of the fingerprint that have been severely smudged, so `MINDTCT` will not attempt to detect minutiae in these areas. As part of the image quality assessment, `MINDTCT` creates a map that determines whether a part of the image has high or low color contrast. Pixel blocks are determined in a similar fashion as described in section A.2. The color intensity is calculated for each pixel, and a distribution of these intensities is computed for the block. The pixel intensity measures how much gray can be seen in a given pixel. For the 256 grayscale fingerprint images, a pixel with intensity of zero appears black, and a pixel with intensity 256 appears white.

To stabilize the distribution of pixel intensities within a block, only the central 80% of the distribution is used to determine level of contrast. If the range of the central 80% is less than five, then the block is labeled as having low color contrast. The threshold of five was determined empirically based on the 256 grayscale. If a block has a pixel intensity of five, then there are only ten shades of gray in the block, which would make it difficult to reliably detect meaningful features such as minutiae.

High Curve Map

The final element in assessing image quality is the ridge curvature. Identifying these areas is important, because areas with high ridge curvature, such as the core and delta, are often useful when comparing fingerprints. Additionally, it is more difficult minutiae in these areas. `MINDTCT` produces a high curve map that indicates such areas in the fingerprint image. The identification of an area having "high curvature" is based on two criteria: *vorticity* and *curvature*. The vorticity of a block measures the total change in ridge direction around all of its surrounding neighbors. The curvature of a block measures the maximum change in its ridge direction and that of its surrounding neighbors. Even if an area of the fingerprint is considered to have high curvature, `MINDTCT` is still able to detect minutiae. The minutiae detected in these areas, however, receive lower quality scores than minutiae detected in other areas.

All of these elements, ridge flow direction, color contrast and curvature, are used to compute the quality score between 0 and 4. In general, the map can be used to assess the reliability of the minutiae detected by the program. In the case that entire image is of sufficiently low quality, the program will not conduct the remaining steps to detect minutiae and instead produces an error message.

Bibliography

[1] Georgia Bureau of Investigation Division of Forensic Sciences: Latent Prints Overview, 2011. http://dofs.gbi.georgia.gov/sites/dofs-gbi.georgia.gov/files/imported/vgn/images/portal/cit_1210/1/18/180850381GBI-LatentPrints.pdf.

[2] NIST Ballistics Toolmark Research Database, 2016. https://tsapps.nist.gov/NRBTD/.

[3] Anguli: Synthetic Fingerprint Generator, 2018. http://dsl.cds.iisc.ac.in/projects/Anguli/.

[4] Center for Statistics and Applications in Forensic Science, 2018. https://forensicstats.org/our-research/statistical-foundations/.

[5] Requirements for the IAI Crime Scene Certification, 2018. https://www.theiai.org/certifications/crime_scene/requirements.php.

[6] Joshua Abraham, Christophe Champod, Chris Lennard, and Claude Roux. Modern statistical models for forensic fingerprint examinations. *Forensic Science International*, 232:131–150.

[7] John Barnard, Robert McCulloch, and Xio-Li Meng. Modeling covariane matrices in terms of standard deviations and correlations, with applications to shrinkage. *Statistica Sinica*, 10: 1281–1311, 2000.

[8] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:567–585, 1989.

[9] G.E. Box and D.R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological*, 26:211–252, 1964.

[10] Niko Brümmer and Johandu Preez. Application-independent evaluation of speaker detection. *Computer Speech & Langugage*, 20:230–275, 2006.

[11] G. Buckland and H. Evans. *Shots in the Dark: True Crime Pictures.* Bulfinch Press, New York,NY, 2001.

[12] Thomas Busey, Arch Silapiruti, and John Vanderkolk. The relation between sensitivity, similar non-matches and database size in fingerprint database searches. *Law, Probability and Risk*, 13: 151–168, 2014.

[13] Raffaele Cappelli, Matteo Ferrara, Annalisa Franco, and Davide Maltoni. Fingerprint verification competition 2006. *Biometric Technology Today*, 15:7–9, 2007.

[14] Bradley Carlin and Siddhartha Chib. Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society - Series B*, 3:473–484, 1995.

[15] Itiel Dror. Cognitive neuroscience in forensic science: understanding and utilizing the human element. *Philosophical Transactions of the Royal Society B*, 370:20140255, 2015.

[16] Itiel Dror and Jennifer Mnookin. The use of technology in human expert domains: challenges and risks arising from the use of automated fingerprint identification systems in forensic science. *Law, Probability and Risk*, 9:47–67, 2010.

[17] Gary Edmond, Alice Towler, Bethany Growns, Gianni Ribeiro, Bryan Found, David White, Kaye Ballantyne, Rachel Searston, Matthew Thompson, Jason Tangen, Richard Kemp, and Kristy Martire. Thinking forensics: Cognitive science for forensic practitioners. *Science & Justice*, 57:144–154, 2017.

[18] Nicole M. Egli. *Interpretation of Partial Fingermarks Using an Automated Fingerprint Identification System*. PhD thesis, University of Lausanne, 2009.

[19] N.M Egli, C. Champod, and P. Margot. Evidence evaluation in fingerprint comparison and automated fingerprint identification systems-modelling within finger variability. *Forensic Sciences International*, 167:189–195, 2007.

[20] I.W. Evett. Implementing bayesian methods in forensic science. *paper presented at the Fourth Valencia International Meeting on Bayesian Statistics*, 1991.

[21] Executive Office of the President's Council of Advisors on Science and Technology. *Report to the President Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. Exectuve Office of the President's Council of Advisors on Science and Technology, Washington, DC.

[22] Michael Fagert and Keith Morris. Quantifying the limits of fingerprint variability. *Forensic Science International*, 254:87–99, 2015.

[23] Federal Bureau of Investigation. Next Generation Identification (NGI).

[24] Federal Bureau of Investigation. Recording Legible Fingerprints, 2017. `https://www.fbi.gov/services/cjis/fingerprints-and-other-biometrics/recording-legible-fingerprints`.

[25] Federal Bureau of Investigation. FBI media, 2017. `https://multimedia.fbi.gov/?q=fingerprint&perpage=50&page=1&searchType=image`.

[26] Carmen Fernández, Eduardo Ley, and Mark Steel. Benchmark priors for bayesian model averaging.

[27] Jacqueline Fish, Larry Miller, Michael Braswell, and Edward Wallace. *Crime Scene Investigation (3rd edition)*. Anderson Publishing, Waltham, MA.

[28] Headlines on Human Hands. Do you have unusual fingerprints?, 2005. `http://handlines.blogspot.com/2005/09/do-you-have-unusual-fingerprints.html`.

[29] Eric Holder, Laurie Robinson, and John Laub. *The Fingerprint Sourcebook*. U.S. Department of Justice Office of Justice Programs, National Institute of Justice, 2011.

[30] Harold Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, U.K., 3 edition, 1961.

[31] M. Kam, G. Fielding, and R. Conn. Writer identification by professional document examiners. *Journal of Forensic Sciences*, 42:1–33, 1997.

[32] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[33] D.V. Lindley. A problem in forensic science. *Biometrika*, 64:207–213, 1977.

[34] Steven Lund and Hari Iyer. Likelihood ratio as weight of forensic evidence: A closer look. *Journal of Research of National Institute of Standards and Technology*, 122:1–32.

[35] Alice Maceo. Qualitative assessment of skin deformation: A pilot study. *Journal of Forensic Identification*, 59:390–440, 2009.

[36] Didier Meuwly, Daniel Ramos, and Rudolf Haraksim. A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Science International*, 276:142–153, 2017.

[37] Ullrich Munzel and Edgar Brunner. Nonparametric methods in multivariate factorial designs. *Journal of Statistical Planning and Inference*, 88:117–132, 2000.

[38] National Forensic Science Technology Center CSI Proficiency Test. National Forensic Science Technology Center CSI Proficiency Test, 2018. `https://www.nfstc.org/products/csi-proficiency-test/`.

[39] Andrew Neath and Joseph Cavanaugh. A bayesian approach to the multiple comparisons problem. *Journal of Data Science*, 4:131–146, 2006.

[40] Cedric Neumann, I.W. Evett, and J.E. Skerrett. Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm. *Journal of the Royal Statistical Society - Series A*, 175:371–415, 2012.

[41] Cedric Neumann, Christophe Champod, Mina Yoo, Thibault Genessay, and Glenn Langenburg. Quantifying the weight of fingerprint evidence through the spatial relationship, directions and types of minutiae observed on fingermarks. *Forensic Science International*, 248:154–171, 2015.

[42] Office of the Inspector General, Oversight and Review Division. *A Review of the FBI's Handling of the Brandon Mayfield Case*. U.S. Department of Justice, Washington, DC.

[43] Robert Olsen. *Scott's Fingerprint Mechanics*. Charles C Thoams, Springfield, IL, 1978.

[44] OSAC. Research Needs Assessment Form, 2015. `https://www.nist.gov/sites/default/files/documents/forensics/osac/FRS-Research-Need-Statistical-Modeling.pdf`.

[45] M.B. Rosson. The role of experience in editing. *Proceedings of INTERACT '84 IFIP Conference on Human-Computer Interaction*, pages 45–50, 1985.

[46] G.S. Sodhi and J. Kaur. Powder method for detecting latent fingerprints: a review.

[47] S. Sonnentag. Expertise in professional software design: a process study. *The Journal of Applied Psychology*, 83:703–715, 1998.

[48] Robert Stacey. A report on the erroneous fingerprint individualization in the madrid train bombing case. *Journal of Forensic Identification*, 54:706–718, 2004.

[49] David Stoney. Measurment of fingeprint individuality. In Henry Lee and R.E. Gaensslen, editors, *Advances in Fingeprint Technology, Second Edition*, chapter 9, pages 327–388. CRC Press, Boca Raton, 2001.

[50] Technical Working Group on Crime Scene Investigation. *Crime Scene Investigation: A Guide for Law Enforcement*. National Forensic Science Technology Center, Largo, FL.

[51] The National Academies Press. *Strengthening Forensic Science in the United States: A Path Forward*. The National Academies Press, Washington, DC.

[52] Bradford Ulrey, Austin Hicklin, Maria Roberts, and JoAnn Buscaglia. Changes in latent fingerprint examiners' markup between analysis and comparison.

[53] U.S. Department of Justice, Office of the Inspector General. *A Review of the FBI's Handling of the Brandon Mayfield Case*. Office of the Inspector General Oversight and Review Division, Washington, DC.

[54] Craig Watson, Michael Garris, Elham Tabassi, Charles Wilson, R. McCabe, Stanley Janet, and Kenneth Ko. User's Guide to NIST Biometric Image Software (NBIS), 2007. `http://ws680.nist.gov/publication/get_pdf.cfm?pub_id=51097`.

[55] D. White, P.J. Phillips, C.A. Hahn, M. Hill, and A.J. O'Toole. Perpetual expertise in forensic facial image comparison. *Proceedings of the Royal Society London B. Biological Sciences*, 282: 1814–1822, 2015.

[56] In-Kwon Yeo and Richard Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87:954–959, 2000.