Statistical methods for the forensic analysis of striated tool marks

by

Amy Beth Hoeksema

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Max D. Morris, Major Professor
L. Scott Chumbley
William Q. Meeker
Alyson G. Wilson
Huaiqing Wu

Iowa State University

Ames, Iowa

2013

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to many people who helped me with various aspects of my research and thesis writing. First and foremost, my major professor, Dr. Max D. Morris, for being an inspiring professor with a fascinating research background and an interesting project that led me to the work I did. Furthermore, for his guidance, patience and support over several years as the work was being done. Not only do I feel that I grew as a statistician during this time working with him, but I also enjoyed all the work I was able to do.

I would further like to thank Dr. L. Scott Chumbley and the other members of our collaborative forensics research team who helped answer questions and provide data along the way. I would also like to thank Jim Kreiser, former forensic examiner from the Illinois Crime Lab for providing the tool marks in this study.

I would like to thank my committee members for their efforts and time during this process: Dr. L. Scott Chumbley, Dr. William Q. Meeker, Dr. Alyson G. Wilson and Dr. Huaiqing Wu.

Last, but certainly not least, I would like to thank my friends in the Statistics Department at Iowa State, who are unfortunately too many to name, and my family who supported me through the ups and downs of classes and research and provided either guidance or just an ear to listen on numerous occasions. And a special thanks to my fiance, Dennis, who stuck with me through it and endured the most of my complaining and was there to celebrate with me when it was all through.

This work was performed at the Ames Laboratory under contract number DE-AC02-

07CH11358 with the U.S. Department of Energy. The document number assigned to this thesis/dissertation is IS-T 3087. The work was partially funded by the National Institute of Justice Grant # 2009-DN-R-119.

ABSTRACT

In forensics, fingerprints can be used to uniquely identify suspects in a crime. Similarly, a tool mark left at a crime scene can be used to identify the tool that was used. However, the current practice of identifying matching tool marks involves visual inspection of marks by forensic experts which can be a very subjective process. As a result, declared matches are often successfully challenged in court, so law enforcement agencies are particularly interested in encouraging research in more objective approaches. Our analysis is based on comparisons of profilometry data, essentially depth contours of a tool mark surface taken along a linear path. In current practice, for stronger support of a match or non-match, multiple marks are made in the lab under the same conditions by the suspect tool.

We propose the use of a likelihood ratio test to analyze the difference between a sample of comparisons of lab tool marks to a field tool mark, against a sample of comparisons of two lab tool marks. Chumbley et al. (2010) point out that the angle of incidence between the tool and the marked surface can have a substantial impact on the tool mark and on the effectiveness of both manual and algorithmic matching procedures. To better address this problem, we describe how the analysis can be enhanced to model the effect of tool angle and allow for angle estimation for a tool mark left at a crime scene. With sufficient development, such methods may lead to more defensible forensic analyses.

We then consider the effect of using multiple tool marks made in the lab. Specifically, we consider how flaws in the mark surface or error in the mark making process make

it is possible for tool marks to be made under the same conditions using the same tool that do not resemble one another. Thus it is necessary to incorporate a quality control step in the tool mark matching process. Toward this end, we describe a method that could be used to verify that all the lab marks made do in fact match each other well enough to be considered reliable for comparing to a field tool mark, or to identify those that should be eliminated.

Finally, we return to the proposed use of a likelihood ratio test to compare multiple tool marks made in the lab to a single field tool mark. In that analysis, a one-sided hypothesis test was used for which the null hypothesis states that the means of the two samples are the same, and the alternative hypothesis states that they are different and appropriately ordered. The weakness of this approach is that the hypotheses are reversed from the desired analysis; we must assume that the null hypothesis is true until we can prove otherwise, which equates to assuming the tool marks were made by the same tool (i.e. the evidence supports the suspect's guilt) until we can prove otherwise. Using synthetic tool marks generated from a statistical model fitted to the lab tool marks, we propose a method for comparing marks that reverses the hypotheses to achieve the desired test.

CHAPTER 1. INTRODUCTION

In forensics, patterns of fingerprints can be used to uniquely identify suspects in a crime. Similarly, the striae on a tool mark can be used to identify a tool. The grinding process used in manufacturing tools such as screwdrivers creates fine-grain parallel scratch marks on the tool called *striae*. If a striated tool mark is used to commit a crime, the negative impression of the tool surface left at a crime scene can be used to identify the tool that was used. Currently, trained forensic examiners use comparison microscopes to examine the details of a tool mark found in the field and compare it to a tool mark made in the lab to determine the match status. However, in 2009 a report by the National Research Council stated that "With the exception of nuclear DNA analysis... no forensic method has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source," (N.R.C. 2009, p. 7). This finding suggests that expert testimony alone is no longer sufficient to classify match status of tool marks and there is a need to make this process more objective.

When tool marks are compared, forensic examiners take into account both large scale signal of the tool mark, known as *class characteristics* and *subclass characteristics*, and the small scale microscopic characteristics known as *individual characteristics*. Class and subclass characteristics are thought to be those that are common to tools made consecutively by the same manufacturing process. Traits such as the length of the tool mark and the general shape go into the class and subclass characteristics. Individual characteristics are thought to be those that are unique and specific to the tool, such as

the fine details that occur during the final grinding process of the tool. When forensic examiners compare tool marks, they first verify that they share the same class and subclass characteristics, then verify microscopically that they match in individual characteristics before concluding that the tool marks match. The same considerations go into the statistical comparison of tool marks; methods must take into account the class, subclass and individual characteristics when computing a numeric comparison value.

Many scientific researchers have proposed methods for statistically comparing tool marks, some of which are summarized more thoroughly in Chapter 2. Several of these algorithms compare two tool marks and return a single numerical index of similarity to be compared. Each approach not only specifies a unique comparison method, but also calls for particular analysis techniques; there are currently "no standard methods for the application of probability and statistics to the analysis of tool mark evidence," (Petraco et al. 2012, p. 901). Most methods rely on comparison of a single field tool mark to a few tool marks made in the laboratory to determine the match status. However, even with the multiple tool marks available, single comparisons are made between lab and field tool marks and are used individually to determine a match. This is an area of concern since researchers would prefer not to rely on a sample of size one to conclude a match. There is also no method of controlling the error rates for these analyses since it is still not known how similar any two consecutively made tools could be. In this thesis, we address the common concern of a single data value and propose adjustments that can be made to resolve other potential concerns with tool mark comparisons.

In Chapter 2, we begin with a summary of some of the analyses that have been proposed by other researchers. For demonstration throughout the thesis, we use data values from Chumbley's algorithm proposed in Chumbley et al. (2010), so we also describe the details of this algorithm. We then introduce a method of analysis that uses multiple lab tool marks and their pairwise comparisons, as well as the pairwise comparisons of lab tool marks to the field tool mark, and compares the samples of data values. A likelihood

ratio test compares the two samples (lab with field comparisons, and lab with other lab comparisons) and returns evidence of a match based on whether or not the samples have the same mean. We provide examples of known matches and known non-matches to demonstrate how the proposed analysis can successfully distinguish between the two samples of tool marks.

We then illustrate how attributes of the tool marks, such as the angle at which the tool is held, have a significant impact on the appearance of the tool mark. That is, tool marks made at different angles with respect to the surface on which they are made can appear as non-matching marks although they were made by the same tool. This suggests that when making tool marks in the lab with the suspect tool, marks should be made at multiple angles since the angle of the field tool mark is always unknown. We propose a model that accounts for the angle of the lab tool marks and predicts the angle of the field tool mark. Examples are provided using tool marks made at 30°, 45°, 60°, 75°, and 85° from both known matching and known non-matching tool marks. The angle predictions and the results of the likelihood ratio tests are given for each example.

Another attribute of tool marks that can affect the results of a comparison is the quality of the marks that have been made in the laboratory; this is addressed in Chapter 3. Although mark-to-mark variation is to be expected due to the difference in individual characteristics, we provide examples in the beginning of this paper of marks that are known matches but differ enough to be falsely identified as non-matches. If there is a flaw in the surface the mark is made on, or only a partial segment of the mark is transfered for various reasons, the mark-to-mark variation can be large enough so as to affect the comparison values of the lab tool marks to one another. Thus, we propose the addition of a quality control step to the tool mark comparison process in which the lab tool marks are tested against one another, checking for outliers, before being compared to the lab tool mark. The model and analyses are provided, as well as examples of both matching and non-matching lab tool marks showing the varying degrees of match status.

Unfortunately the data set available to us contains only four tool marks per matching set, so although the quality model seems to be able to identify poorly made tool marks, the variance between lab tool marks has a significant effect on the analysis; tool marks that are very similar can falsely determine the existence of an outlier and those with a larger variation that you would expect to be different can falsely return a match.

Although the model and method described in Chapter 2 is able to correctly predict the angle of the field mark, and the likelihood ratio test can often distinguish between matching and non-matching samples of comparisons, there is an inherent issue with the use of this test based on the hypotheses used. In Chapter 2, a standard hypothesis test for the difference in means of two samples is used in which the null hypothesis states the means of the samples are the same, and the alternative hypothesis states the mean of the comparisons between two lab marks is larger. Since we assume the null hypothesis is true until we can provide sufficient evidence to support the alternative hypothesis, this means we must assume the two samples of tool marks are the same until we can provide evidence to support they are not. This is equivalent to assuming the suspect is guilty (the suspect tool was used to make the field mark), until we can provide evidence to support he is not guilty (the suspect tool was not used to make the field mark). Since this is the reverse of the hypotheses we would like to test, we address this concern in Chapter 4.

To "reverse" the hypotheses, in Chapter 4 we propose using the lab tool mark to create synthetic tool marks which are statistically generated to match the lab mark in class and subclass characteristics and vary subtly in individual characteristics. Using a Loess smooth, we model the class characteristics and generate residuals for the individual characteristics. Parameters for the synthetic tool marks are chosen such that comparison between two synthetic tool marks and comparison between a synthetic mark and a lab mark are indistinguishable from one another as determined by a Kolmogorov-Smirnoff statistic. We then compare the field tool mark to the lab tool mark and conclude there

is evidence of a match if the data value from comparing the field tool mark to the lab tool mark is an outlier in the sample of comparisons between the field tool mark and the synthetic tool marks. Again, examples of both matching and non-matching field and lab tool marks are provided to demonstrate the analyses.

Chapter 5 provides a summary of the results presented in Chapters 2, 3 and 4. We also provide suggestions for future research.

CHAPTER 2. SIGNIFICANCE OF ANGLE IN THE STATISTICAL COMPARISON OF FORENSIC TOOL MARKS

A paper Submitted to *Technometrics*

Amy B. Hoeksema [1] [2] and Max D. Morris[3] [4]

Abstract

In forensics, fingerprints can be used to uniquely identify suspects in a crime. Similarly, a tool mark left at a crime scene can be used to identify the tool that was used. However, the current practice of identifying matching tool marks involves visual inspection of marks by forensic experts which can be a very subjective process. As a result, declared matches are often successfully challenged in court, so law enforcement agencies are particularly interested in encouraging research in more objective approaches. Our analysis is based on comparisons of profilometry data, essentially depth contours of a tool mark surface taken along a linear path. Chumbley et al. (2010) point out that the angle of incidence between the tool and the marked surface can have a substantial impact on the tool mark and on the effectiveness of both manual and algorithmic matching procedures. To better address this problem, we describe how the analysis can

[1]Graduate student, Department of Statistics, Iowa State University
[2]Primary researcher and author
[3]Department of Statistics, Iowa State University
[4]Department of Industrial and Manufacturing Systems Engineering, Iowa State University

be enhanced to model the effect of tool angle and allow for angle estimation for a tool mark left at a crime scene. With sufficient development, such methods may lead to more defensible forensic analyses.

2.1 Introduction

When a crime is committed using a machined metal tool, such as breaking into a house with a screwdriver, evidence is often left behind in the form of a striated tool mark. Once a suspect is identified and found to own such a tool or have one in his possession, it is up to forensic examiners to determine whether or not that tool was the one used to create the mark found at the crime scene. Using a process analogous to that of matching fingerprints, forensic examiners compare the tool mark left at the scene to others made in the lab using the suspect tool to look for similar microscopic characteristics with the goal of determining whether the tool marks match, i.e., were made by the same tool. The marks made by any tool that leaves a striated pattern can be compared using these techniques; however, for this paper we will focus on marks made by screwdrivers.

Although visual comparison of firearms and tool marks has been performed since the early 1900s, in recent years the process has come under scrutiny. In 2009, a National Research Council (N.R.C.) Report was published stating that "With the exception of nuclear DNA analysis... no forensic method has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source," (N.R.C. 2009, p. 7). This, along with other proceedings such as the Daubert Case (Daubert v. Merrell Dow Pharmaceuticals, Inc., 1993) has led to numerous instances of expert testimony on tool mark matching being disallowed as inadmissible in court. The current process is considered to be subjective and thus it is necessary to find a way to make it more objective and conclusive.

Many researchers, both in the forensics and statistics communities, have taken on this challenge and proposed numerical methods for comparing tool marks, some of which we will discuss in Section 2.1.1. Although these algorithms provide a way to quantify the similarity between two tool marks, little work has been done to examine the effects of attributes of the tool mark on the matching process, such as the angle at which the tool was held when making the mark. In this article, we will address the issue of the significance of tool angle to the tool mark matching process. We will first give a brief background of the tool mark comparison process and describe the quantitative process and the resulting data we will be using (Section 2.2). We describe the basic statistical model we use in forensic applications in Section 2.3. We will then show the important influence that angle has on tool mark matching and suggest a modified model that incorporates these effects (Section 2.4). Finally we will show some results using the new model (Section 2.5) and discuss conclusions and future research directions (Section 2.6).

2.1.1 Tool Mark Comparison Background

When making a tool such as a screwdriver, one step is a grinding process during which the end of the tool is ground down creating fine-grain parallel scratch marks that are called *striae*. The "negative" impression of the striae is left behind when the relatively hard tool comes in contact with a softer metal surface such as a metal window frame. Similar marks are left on bullets by the "rifling" pattern cut into gun barrels, and the forensic inspection methods used for tool marks and firearms are closely related. Any striated tool mark is made up of what are known as *class characteristics*, *subclass characteristics* and *individual characteristics*. According to the N.R.C., class characteristics are "distinctive features that are shared by many items of the same type... such as the width of the head of the screwdriver" and individual characteristics are "the fine microscopic markings and textures that are said to be unique to an individual tool or firearm. Between these two extremes are 'subclass characteristics' that may be com-

mon to a small group of firearms and that are produced by the manufacturing process, such as when a worn or dull tool is used to cut barrel rifling" (N.R.C. 2009, p. 152). Nichols summarized it well when he distinguished between the different types as "class characteristics which are intentional; subclass characteristics which are unintentional but common to a select group; and individual characteristics which are accidental and unique" (Nichols 1997, p. 466).

The current practice of examining striated tool marks is based on a forensic examiner comparing the evidence mark to the reference tool mark side-by-side under a comparison microscope, such as that shown in Figure 2.1. Before the tool marks are placed on the two stages of the microscope, the examiner must confirm that the tool marks resemble one another closely enough to justify a more detailed comparison. By doing so, he can visually discount similarity due to both class and subclass characteristics. The National Institute of Justice (NIJ) states in their on-line Firearm Examiner Training module that "Examination of the tool allows the examiner to assess the level of subclass characteristics...The examiner compares the class characteristics of the two objects; if all class characteristics correspond, the examiner proceeds to compare the individual characteristics" (National Institute of Justice). It is the individual characteristics of the tool marks that are critical in quantifying and analyzing for evidence of a match.

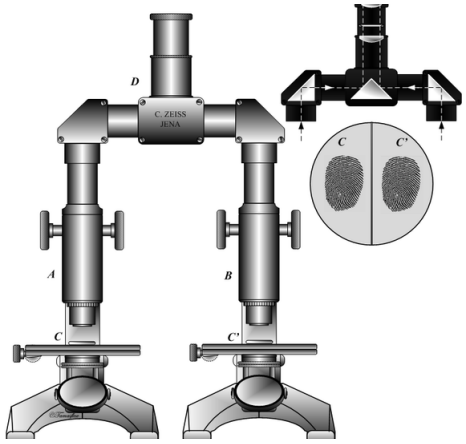


Figure 2.1: Comparison microscope, Tamasflex (2012).

The first quantitative method for comparing tool marks was proposed by Biasotti (1959). His method, based on a concept known as consecutive matching striae (CMS), provided a way for examiners to describe the extent of matching striae that could be understood universally, and thus removed some of the subjectivity from the process. In his initial empirical study, he counted the number of striae that lined up between non-matching bullets all fired from the same kind of gun, and found that in the material he examined there were no runs of CMS that exceeded four. Biasotti and Murdock (1997) proposed a conservative quantitative criterion for identification which is used for matching in the forensic community that states there is evidence of identification "when at least two different groups of at least three consecutive matching striae appear in the same relative position, or one group of six consecutively matching striae are in agreement in an evidence tool mark compared to a test tool mark." Since then, many other studies have been done confirming this criterion and using the CMS as a quantitative comparison tool. Neel and Wells (2007) propose using the relative frequency of CMS runs to perform a test of proportions. In this test they compare the "most conservative known match" to the "best known non-match" and determine a match if the proportions of CMS in

both situations are different.

More recently, as newer technology has come available, researchers have begun to develop algorithms to compare entire marks digitally rather than having to rely on microscopic counting methods. Using a stylus profilometer or optical profilometer, or a confocal microscope, the depths of the grooves of striated tool marks can be measured resulting in a quantified tool mark that we will refer to as a *digitized mark*. Quantifying the entire tool mark opens the door to more objective methods of mark comparison which result in a single numerical index of similarity between two marks. Bachrach et al. (2010) proposed using the cross correlation function, which he calls the relative distance between two marks, to quantify the degree of match. Chumbley et al. (2010) suggested using a Mann-Whitney U-statistic, which is described in the next subsection, to quantify the match. Another quantitative approach proposed by Petraco et al. (2012) quantifies tool marks into binary matrices, which identify where the lines of the striae start and stop, and uses principle components analysis to determine groupings of tools.

In their paper, Petraco et al. state "There are no standard methods for the application of probability and statistics to the analysis of tool mark evidence," (Petraco et al. 2012, p. 901). The method that we propose in Section 2.3 assumes we have a single numerical index of similarity for every tool mark comparison available. Although many of the methods mentioned above would suffice for our research, we have chosen to use the U-statistic proposed by Chumbley et al. since it meets the requirements for our technique and mirrors the current practice that examiners use for comparison.

2.1.2 Chumbley's U-Statistic

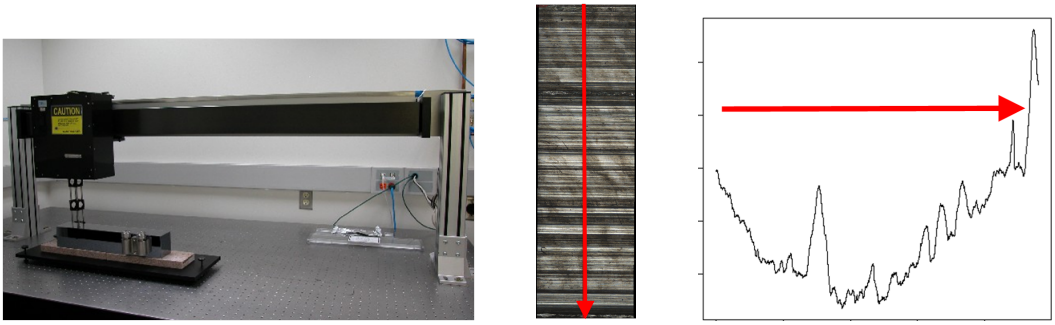
To fully understand the approach of the Chumbley algorithm, we will first look more closely at the current practice of tool mark matching followed by forensic examiners. The NIJ's on-line Firearm Examiner Training module breaks down the process that examiners use; we have summarized the process here to remove some of the technical

details. Once an examiner has a mark made by a suspect tool in the lab to compare to an evidence mark, he begins by visually comparing the marks to verify they merit more careful microscopic comparison. He then places both marks on a comparison microscope and identifies a small area of the lab specimen that seems to have the "best" marks and indexes this area.

The next step is to align the two marks to confirm the "consistency of class characteristics." At this point, if the class characteristics match, he will proceed to move the evidence mark on the microscope stage to identify an area of the evidence mark that matches the "best" area of the test mark. If such an area is found, so the individual striae of the two marks align in the best marked regions, he locks the two tool marks in this relative orientation and examines the remaining areas of the two tool marks in question to verify that the surrounding areas are also similar. Because the striae are essentially parallel patterns of scratches in each mark, the emphasis is on finding small areas in each mark for which the "cross striae" patterns are similar. It is important to reiterate that once the marks have been aligned by identification of segments that appear to match, further examination is made along corresponding segments of the tool mark to confirm (or otherwise) this apparent match. The examiner's conclusion is based on how well these additional segments match after the tool marks have been aligned.

As was previously mentioned, we chose to use Chumbley's algorithm as it was specifically designed to mimic the examiner's process. To quantify the comparison of the tool marks, they must first be digitized. The stylus of a surface profilometer, shown in Figure 2.2(a), is used to trace the surface of the tool mark along a linear path perpendicular to the striae as indicated in Figure 2.2(b). Arrows on both figures illustrate the direction of movement by the stylus. The depths of the grooves are recorded at a set of "pixel" locations of fixed separation. When the numeric depths are plotted by pixel location, the result is a digital tool mark like that shown in Figure 2.2(c). The vertical axis of this graph is on a dramatically different scale than the horizontal axis to amplify the

depth of the striae across the tool mark. Because the striated surface is essentially a set of parallel ridges, most of the useful information about the individual characteristics of the tool can be characterized by this single-index data series.



(a) (b) (c)

Figure 2.2: Using a profilometer to digitize a tool mark. (a) Stylus profilometer, (b) magnified tool mark showing the location of a profilometer scan and (c) the resulting digitized tool mark.

The key to the forensic examiner's process is to find an area of "best" match in both tool marks being compared and verify a match based on how well the surrounding areas align. In an analogous strategy, the algorithm proposed by Chumbley et al. (2010) uses numerical optimization to determine the "best" matching subset of digitized tool mark in each profilometer trace. An index of similarity is computed for all possible pairs of windows of a set length between the two marks along the entirety of both tool marks. The length of this best match window is chosen as part of a preliminary experiment which maximizes the Chumbley statistic for a pair of tool marks. The best matching windows are chosen to be the two areas of tool marks for which the index of similarity is maximized; Chumbley et al. refer to this first step as the Optimization step and they use Pearson's correlation coefficient as the index of similarity. Due to the large number of pixels present in both marks (around 9000 in each), and the relatively small size of the best matching window (usually 300 to 500 pixels), even marks made by different tools tend to include areas with large correlation.

Once the best matching windows have been found, the algorithm performs a Validation step. This step begins by moving corresponding windows a random distance from the best match window but the same distance on each mark, and computing the index of similarity (correlation) for these windows. We will refer to these corresponding windows as *coordinated shifts*. Although this correlation is lower than the best match window, if the two marks were made by the same tool, it should be relatively high since these windows should physically correspond. Finally, a third set of windows is identified at random locations in each tool mark which we will call *independent shifts*. That is, windows are shifted independently of one another on each tool mark and likely in different distances from the best match. Since these windows do not correspond, there is no reason to believe they should return a large correlation value. Both the coordinated shifts and the independent shifts are repeated many times using different random shift amounts, and correlations between tool marks are computed for each set of windows. These validation windows are chosen to be much smaller than the best match window; usually 50 pixels each. A primary reason for the relatively small size of the optimization and validation windows is, again, the intent that this process mimic what tool mark examiners currently do. Current practice does not require that striae patterns correspond across the entire width of the tool marks, but focuses on matching relatively small sub-regions and follow-up examination of segments in the (physical) vicinity of that match.

Figure 2.3 illustrates the optimization and validations steps of Chumbley's algorithm on a pair of matching tool marks. Although the entire tool mark is used for both steps in the algorithm, we have zoomed in on the portions of tool mark that contain the best match windows, which are represented by the solid connecting lines; the correlation for this pair of windows is 0.999. Two sets of coordinated windows are shown by the dashed connecting lines; the first pair of windows have a correlation of 0.997 and the second pair of have a correlation of 0.967. The boxes connected by dotted lines represent two sets of

independent windows; the first windows have a correlation of 0.557 and the second pair have a correlation of 0.202. In this example, both coordinated windows have a higher correlation than the independently shifted windows, which is what we would expect since the tool marks are known matches.

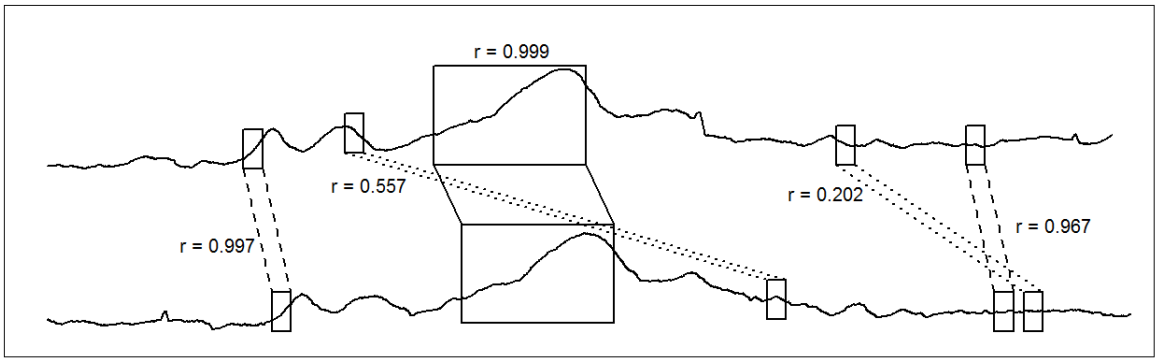


Figure 2.3: Matching segments of two tool marks showing the best match window (solid lines), two coordinated shifts (dashed lines) and two independent shifts (dotted lines) and their correlations.

Finally, a single similarity index is computed, specifically, a Mann-Whitney U-statistic (e.g., Kowalski and Tu (2007)) is the basis of a non-parametric two-sample test used to determine whether the two samples of correlation coefficients (from coordinated and independent shifts) were drawn from a common distribution. The null hypothesis for this test states that the two samples have the same distribution, so the probability of an element from one sample being larger than an element from the other sample is 0.5. The alternative hypothesis states that the underlying distributions are different, so the probability that an observation from one sample is larger than one from the other sample is greater than or less than 0.5. Using the correlations from the coordinated shifts and independent shifts as the two samples, if the tool marks match, we would expect that the coordinated shifts would yield larger correlations than the independent shifts, as illustrated in Figure 2.3. However if the marks do not match, then none of the validation windows should physically correspond, and there should be

no systematic difference between the correlations generated through coordinated and independent shifts. As a final step, the U-statistic is standardized so that under the null hypothesis it asymptotically follows a standard normal distribution; Chumbley et al. refer to this value as T1. For these data, the null hypothesis that the two samples come from the same distribution simplifies to "no match" between the two tool marks.

The intent of Chumbley et al. was that a single U-statistic, calculated in this way, might be used to assess the similarity of two tool marks, by comparison to the standard large-sample distribution theory for this statistic. While empirical work showed that these indices were approximately normally distributed (due to their linear form), and statistics computed using tool marks made with the same tool were typically larger than those made with different tools (as would be hoped), the moments of the apparent null distribution were not always as would be suggested by the standard theory. This is likely due, at least in part, to a lack of independence between correlations in each sample stemming from the finite population of pixels in each tool mark, and the correlation pattern evident within each trace. In addition, experimental work indicated that physical factors, such as the overall smoothness of the marked surfaces, had some effect on the moments of the null distribution.

The present research was carried out primarily to construct a "self-calibrating" test to overcome this weakness in the Chumbley et al. proposal. While we continue to use the U-statistics described by Chumbley et al. in comparing individual pairs of tool marks, we expand the comparison by relying on multiple lab marks, so that the analysis can be based on multiple lab-to-lab comparisons and multiple lab-to-evidence comparisons. In the illustrative calculations that follow, individual U-statistics were computed using "best match" windows of either 300, 400 or 500 pixels (out of about 9000 in each series) and validation windows of 50 pixels. Fifty correlations from coordinated shifts and 50 from independent shifts were compared within each U-statistic. Because we will not depend on the asymptotic moments of the U-statistics here, the datum representing

each comparison of two tool marks was computed as the average of 200 such U-statistics to minimize variation in the analyzed index values.

2.2 Data

An important physical characteristic of this problem is that there is generally only one field tool mark available. Depending on the circumstances of the crime, additional evidence tool marks may occasionally be available, but in any case, this cannot be controlled by the forensic examiner, who must work with what is available. The restriction to a single field tool mark is an obvious limitation on the available information, but one that is inherent in the physical problem. On the other hand, a large number of comparison tool marks may be made in a forensics laboratory using a suspect tool; in fact, this is common practice among tool mark examiners. Multiple laboratory tool marks may be made under the same conditions or under different conditions, depending on what is known or suspected regarding the particular use of the tool at the crime scene. Again, it should be recognized that while this is useful, no amount of replication in the laboratory sample can entirely compensate for the limit imposed by the physical constraint of a single field mark. Because information is in this sense unavoidably incomplete, it is important to understand that it will not be possible to effectively detect all true matches; there will always be cases in which the suspect tool actually was used at the crime scene, but that the limited evidence will not support a definitive conclusion that the marks were made by the same tool.

We focus here on the case in which a single field tool mark is available, and a suspect tool has been used to produce multiple tool marks in the laboratory. In this section and in Section 3, the lab tool marks are regarded as experimental replicates, in that they are made under the same conditions. In Section 4 this will be expanded to a setting in which an important covariate, tool angle, is systematically varied.

2.2.1 Data for Multiple Lab Marks

Suppose a single tool mark was found at the crime scene. A suspect tool is obtained and forensic examiners make several marks using the suspect tool in the lab under controlled conditions. As previously mentioned, comparing multiple marks all known to be made by the same tool under the same conditions in the lab, we can create a sample of matching mark-pairs from that tool. The same lab tool marks can be compared to the field mark to create a smaller sample of mark-pairs with an unknown matching status. Rather than evaluating only one mark-pair, we now have two samples that we can compare. If there is no apparent systematic difference between these two samples, this supports the argument that the marks were all made by the same tool, i.e., that the crime scene tool mark and lab tool marks "match."

A single data value, as we will refer to it in this paper, is the numerical index of similarity that results from comparing two tool marks. Let x_0 represent the tool mark that was found at the crime scene. Let $x_1, ..., x_n$ represent the n tool marks that were made by the suspect tool in the lab. Note that all comparisons of the marks not including x_0 are known matches, since all were made by the same tool. Let y_{ij}, $i < j$, represent the numerical index of similarity that results from comparing x_i to x_j. Once we make all pairwise comparisons of available tool marks we will have two different types of data. The first set, y_{0j} for $j = 1, ..., n$, includes all comparisons of the field mark to the lab marks. We will call this type of data field-lab comparisons. The second set is y_{ij} for $i, j = 1, ..., n$ and $i < j$. These data values represent indices of similarity for a known match from the suspect tool which we will call lab-lab comparisons.

The collection of all data values will be denoted by the vector $\mathbf{y}$ which is of length N. For purposes of organization, $\mathbf{y}$ will always be ordered such that $i < j$ and each index follows standard numerical ordering, 0 through n. When all possible comparisons are made there are n field-lab comparisons and $\binom{n}{2}$ lab-lab comparisons, so $N = n + \binom{n}{2}$.

When describing data, we will discuss a dataset in terms of the number of lab marks, such as a dataset of size n. Note that this means there are $n + 1$ tool marks under comparison since the complete set includes the field mark as well as lab marks.

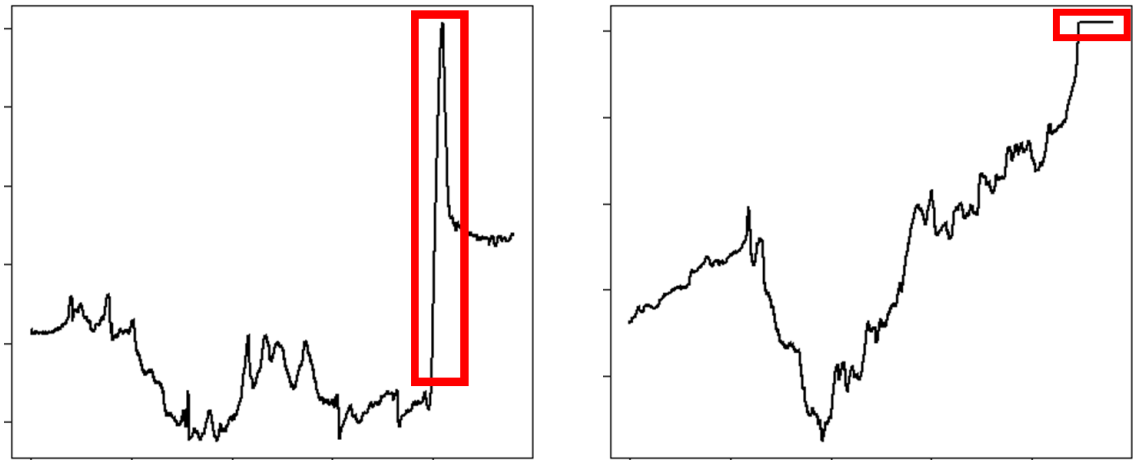
2.2.2 Data Used In This Study

The data used for demonstration in this paper were not actual forensic data, but were generated under laboratory conditions using similar tools. All tool marks were made in a lab by a professional forensic examiner. Hence, "matching" data sets were generated by comparisons of tool marks known to have been made by the same tool, while "non-matching" data sets were generated from n tool marks made by one tool and a field mark made by another. For our study, six screwdrivers were chosen out of a batch of fifty sequentially manufactured screwdrivers. For each screwdriver, four physical scrapes were made at set lab conditions on a piece of lead. This was repeated under five different sets of lab conditions (the exact settings are explained in Section 2.4.1). Ten profilometer traces were taken along the each of the marks and were then visually observed as digitized tool marks looking for flaws in the individual traces. The first of the ten profilometer traces of each was chosen to represent that physical scrape once it was verified that all ten were the same in appearance and there were no visual distinctions or flaws between the traces.

A further word of explanation should be offered concerning the rationale for how tools were selected for this study. As just noted, the six screwdrivers used were selected from a set of consecutively manufactured tools; this was done to minimize variation in manufacturing conditions (including wear of the manufacturing tools) so that the resulting screwdrivers are as *identical as possible*. For other similar studies, custom tools are sometimes produced in more tightly controlled conditions than can be obtained in a manufacturing environment, again to minimize variation among the tools. This may at first seem to be an odd approach to the design of these studies, since the eventual

aim is to produce techniques that can be used across a much broader population of tools. However, the single most important consideration in forensic examination is that false "match" declarations be made as seldom as possible. (As noted at the beginning of this section, the physical limits of forensic evidence available make it inevitable that some true matches cannot be confirmed.) Experiments conducted with near-identical tools offer the most severe test of whether an analytical method can be trusted to yield few false "match" conclusions. Because credible methods *must* pass this test, it is a generally safe bet that they will not lead to erroneous "match" determinations when the tools are more physically different. Of course, it is also important that studies based on different kinds of near-identical tools eventually be undertaken, to demonstrate that good performance can be expected across a variety of relevant tool types. But given the need to effectively discriminate between tool marks made by very similar tools, the common use of sequentially produced tools in forensic testing is easily understood.

Upon inspection of the digitized tool marks, it was noted that a few anomalies were occurring on the edges of several of the marks. Occasionally during the profilometer scan, part of the flat lead plate that does not contain information about the physical tool mark was read resulting in a sharp peak at the edges of the digitized mark from the stylus transitioning between the flat lead surface to the start of the tool mark. We refer to this as a sharp peak anomaly in the data. Another anomaly occurred when the stylus reached the maximum or minimum allowable depth while it was scanning the tool mark. As a result, the maximum or minimum depth value was recorded for the length of tool mark for which the mark was too deep or too shallow. In these areas, the digitized mark shows a flat horizontal line on this segment of tool mark, which we refer to as a flatline anomaly. Examples of both of these anomalies are shown in Figure 2.4 where a sharp peak is visible on the right edge of the tool mark in Figure 2.4(a) and a flatline is visible on the right edge of the tool mark in Figure 2.4(b); both are shown in the boxes.



(a) A sharp peak anomaly.

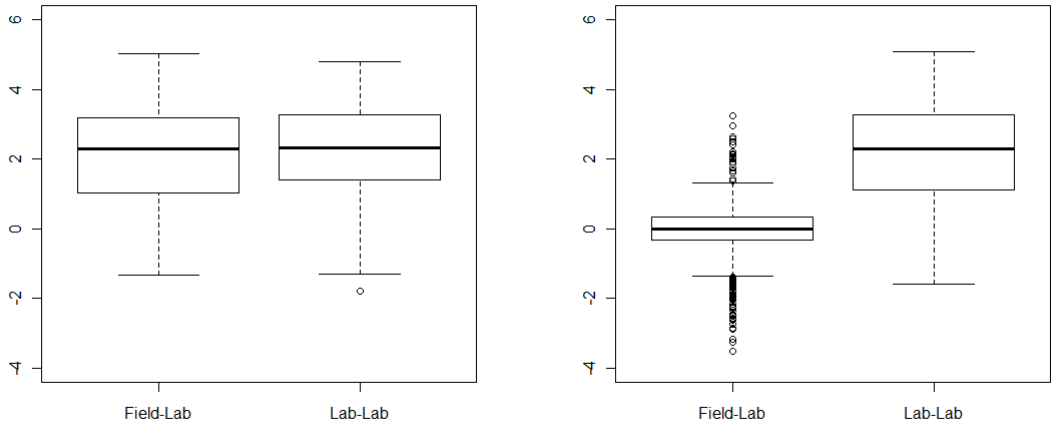
(b) A flatline anomaly.

Figure 2.4: Examples of anomalies on the edges of tool marks before pre-processing.

The primary concern of anomalies such as these is that even though they are not a part of the actual tool mark we are interested in comparing, their features make them likely for matching other similar parts of other tool marks resulting in large correlation. The best match window could be chosen to include these areas which would result in false information from the corresponding coordinated and independent validation windows. For this reason, and since these areas are not part of the signal from the tool mark, we included a pre-processing step in which the portion of any tool mark from a detected anomaly to the nearest end was deleted, so long as this left at least 80% of the tool mark intact. For these purposes, a sharp peak was defined to be a section of 200 consecutive pixel locations that have a variance greater than 100. Flatlines were defined to be sections of tool mark that reached the profilometer's maximum or minimum measurement value.

After pre-processing, pairs of tool marks were evaluated using the Chumbley procedure, resulting in the data to be analyzed (y_{ij}). Visualizing the available data, we can see how they might be used to determine match status. Recall for this study we had six tools available and made four tool marks under each lab setting; a total of 30 sets of

four tool marks made under the same conditions. To examine the distributions of y_{ij}, we consider a few examples. For the distribution of matching y_{ij}s, a single tool mark from each set of matching marks was chosen to be the field tool mark and compared to the remaining three in the set. Thus for a full analysis, there were three field-lab comparisons and three lab-lab comparisons that resulted from data known to match. The field-lab and lab-lab comparisons for each example are shown in the boxplots of Figure 2.5. Note there were 90 y_{ij}s in each boxplot, three from each of the thirty matching sets. We can see from the plot that the samples of field-lab comparisons are indistinguishable from the samples of lab-lab comparisons when the field mark was made under the same set of conditions and by the same tool as the lab marks.



(a) Known match. (b) Known non-match.

Figure 2.5: Boxplots of field-lab comparisons with lab-lab comparisons under known match (a) and known non-match conditions (b).

For a non-match distribution, we compared all four tool marks within a matching set to a field mark which is a mark chosen from a set made by another tool. Non-matching data sets have $n = 4$ - four field-lab comparisons and six lab-lab comparisons. For each match set, a non-matching set was made using one mark from the each of the five sets made by each of the other five screwdrivers as the "field mark;" thus 25 non-matching sets were made for each match set. The aggregated comparison values for all of the sets

are shown in the boxplots in Figure 2.5(b) and have again been grouped by field-lab comparisons and lab-lab comparisons. Since the a single match set was compared to 25 different field tool marks, there is repetition in the lab-lab $y_{ij}s$, so the duplicates have been removed. There are a total of 180 distinct comparisons in the lab-lab boxplot and 2,625 distinct comparisons in the field-lab boxplot. We can see from these boxplots that the the field-lab comparisons tend to be smaller than the lab-lab comparisons, and they are centered at zero.

Examining the boxplots in Figure 2.5, we can make a few conclusions about how to determine evidence of a match using these data. If the field-lab comparisons are indistinguishable from the lab-lab comparisons, such as shown in Figure 2.5(a), then there is no evidence that the tool marks are different, and the data are therefore consistent with the hypothesis that all marks were created using the same tool. However, if the field-lab comparisons are relatively small compared to the lab-lab comparisons, as in Figure 2.5(b), then there is evidence that the field mark and the lab marks were created using different tools.

2.3 The Basic Model

Since we would like to determine whether or not tool marks were made by the same tool, a hypothesis test can be used to compare the two samples of field-lab comparisons and lab-lab comparisons. Toward development of such a test, let μ_0 be the mean for a field-lab comparison, that is $E(y_{0j}) = \mu_0$ for $j = 1, ..., n$. Let μ_1 be the mean for a lab-lab comparison, so $E(y_{ij}) = \mu_1$ for $i, j = 1, ..., n$ and $i < j$. We will assume that all data values have a common variance defined as $Var(y_{ij}) = \sigma^2$ for $i, j = 0, 1, ..., n$ and $i < j$. We further assume that each y_{ij} is normally distributed. The similarity index of Chumbley et al. (2010) is a standardized U-statistic, for which an assumption of approximate normality is justifiable. The assumption may also be reasonable for other

similarity measures.

To verify the assumption of normality is reasonable for the data we have available, we refer back to the field-lab comparisons that produced the boxplots in Figure 2.5. The U-statistic theory states that under the null hypothesis of the two samples being from the same distribution, the large-sample distribution of the standardized U-statistic is standard normal. For the Chumbley algorithm, the two samples are the coordinated shifts and the independent shifts so if the null hypothesis is true, there is no difference in the two types of shifts which indicates evidence of a non-match. A histogram of the non-matching field-lab data is shown in Figure 2.6(a). We can see the plot is symmetric, centered at zero and has a standard deviation slightly less than one. However, the tails of the distribution are longer than would be expected of a standard normal distribution. The Q-Q plot is shown in Figure 2.6(b) verifies this concern, even though the middle appears very linear.

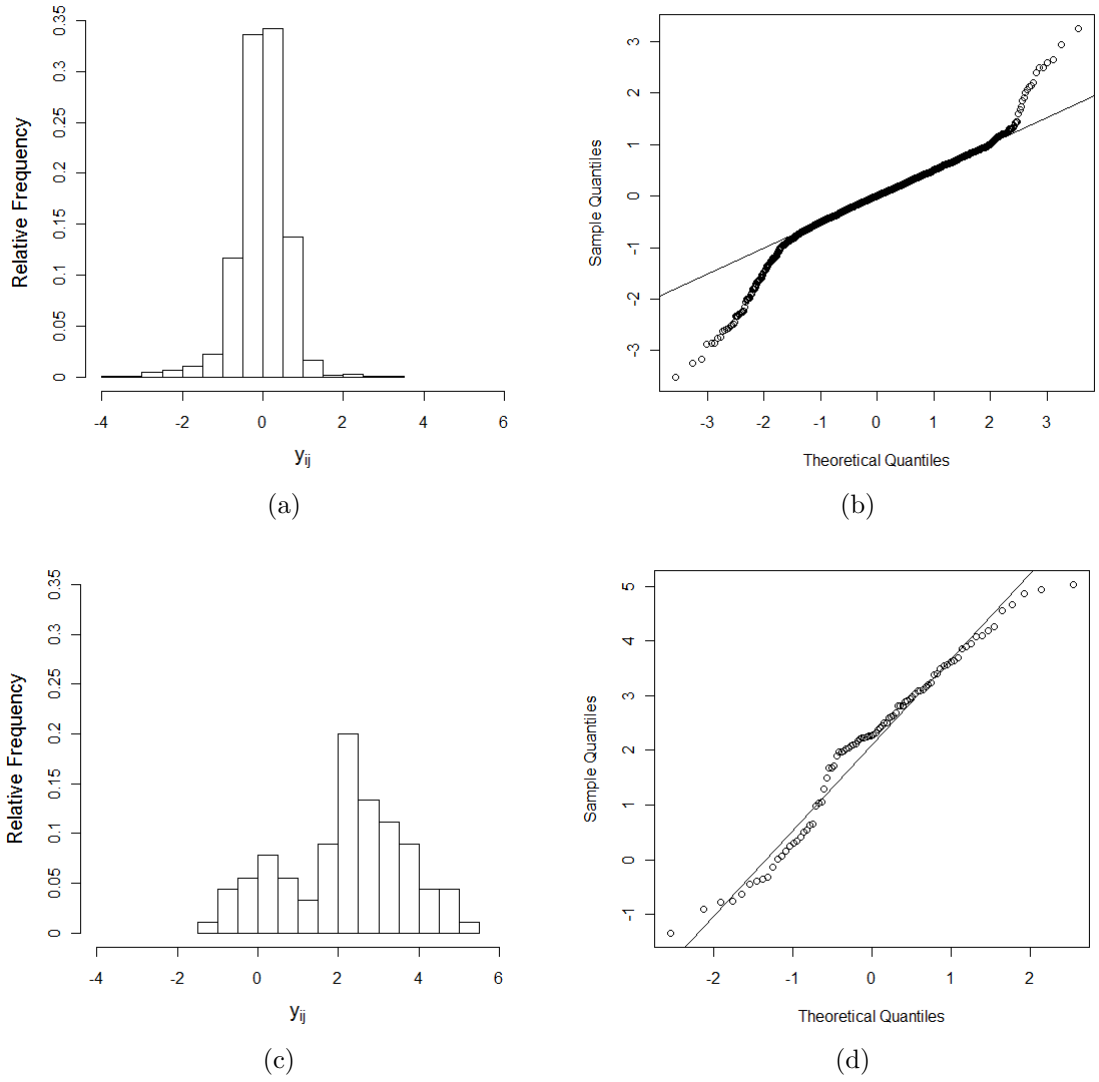


Figure 2.6: Relative frequency histograms of the (a) field-lab comparisons from non-matches and (b) it's associated Q-Q plot; (c) field-lab comparisons from matches and (d) it's associated Q-Q plot.

A histogram of the matching field-lab comparisons is shown in Figure 2.6(c). Here, the mean for the matching comparisons is much larger, around 2, and the spread is larger as well. This is due, in part, to the variability in the quality of matching tool marks. There were a handful of tool marks that are substantially different from the others within their match set and this results in the minor mode near 0. A Q-Q plot is shown next to the histogram in Figure 2.6(d) which confirms the data are relatively

normally distributed.

Although these specifications are sufficient to fully define the distribution for a single data value, we also need to address the joint distribution of all pairwise comparisons. To facilitate this, we will further assume that the joint distribution of $\mathbf{y}$ is multivariate normal. The mean of $\mathbf{y}$ is a vector of means μ_0 and μ_1 with the form $\boldsymbol{\mu} = (\mu_0 \mathbf{1}'_n, \mu_1 \mathbf{1}'_{N-n})'$ and the variance of each element of $\mathbf{y}$ is σ^2. To finish defining the joint distribution of $\mathbf{y}$, we need to develop an appropriate dependency structure reflecting the way the data are generated.

2.3.1 Correlation

Each y_{ij} is the result of comparing two tool marks, specifically x_i with x_j. Thus, at most four physical tool marks are involved in the consideration of covariance between two data values. Since these four tool marks are not necessarily distinct, we will say two data values are correlated with correlation ρ if a common tool mark is involved in both comparisons. That is, y_{ij} is correlated with y_{kl} if $i = k, i = l, j = k$ or $j = l$, but not $(i, j) = (k, l)$. Two comparisons with no marks in common are uncorrelated. We do not consider the case of two comparisons made on the same pair of tool marks because those similarity values would be identical, i.e., there is no measurement-specific "error" in this system, and so no point in replication. Let $\mathbf{R}$ be the $N \times N$ correlation matrix of $\mathbf{y}$ defined by the following entries

$$Corr(y_{ij}, y_{kl}) = \begin{cases} 0 \text{ if } i \neq k, i \neq l, j \neq k \text{ and } j \neq l \\ \rho \text{ if } i = k \text{ or } i = l \text{ or } j = k \text{ or } j = l, \text{ but not } (i, j) = (j, k) \\ 1 \text{ if } i = k \text{ and } j = l. \end{cases} \tag{2.1}$$

With this correlation structure in place, we can finish defining the joint distribution of all pairwise comparisons of tool marks. Let $\mathbf{y}$ be the vector of all pairwise comparisons ordered so that each y_{ij} is such that $i < j$. Then $\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{R})$ where

$\boldsymbol{\mu} = (\mu_0 \mathbf{1}'_n, \mu_1 \mathbf{1}'_{N-n})'$ and $\mathbf{R}$ is as defined in (2.1). The complete model for a data set of size $n = 4$ is shown in (2.2) through (2.4).

$$\mathbf{y} = (y_{01}, y_{02}, y_{03}, y_{04}, y_{12}, y_{13}, y_{14}, y_{23}, y_{24}, y_{34})' \tag{2.2}$$

$$E(\mathbf{y}) = \boldsymbol{\mu} = (\mu_0, \mu_0, \mu_0, \mu_0, \mu_1, \mu_1, \mu_1, \mu_1, \mu_1, \mu_1)' \tag{2.3}$$

$$Var(\mathbf{y}) = \sigma^2 \mathbf{R} \text{ with } \mathbf{R} = \begin{pmatrix} 1 & \rho & \rho & \rho & \rho & \rho & \rho & 0 & 0 & 0 \\ \rho & 1 & \rho & \rho & \rho & 0 & 0 & \rho & \rho & 0 \\ \rho & \rho & 1 & \rho & 0 & \rho & 0 & \rho & 0 & \rho \\ \rho & \rho & \rho & 1 & 0 & 0 & \rho & 0 & \rho & \rho \\ \rho & \rho & 0 & 0 & 1 & \rho & \rho & \rho & \rho & 0 \\ \rho & 0 & \rho & 0 & \rho & 1 & \rho & \rho & 0 & \rho \\ \rho & 0 & 0 & \rho & \rho & \rho & 1 & 0 & \rho & \rho \\ 0 & \rho & \rho & 0 & \rho & \rho & 0 & 1 & \rho & \rho \\ 0 & \rho & 0 & \rho & \rho & 0 & \rho & \rho & 1 & \rho \\ 0 & 0 & \rho & \rho & 0 & \rho & \rho & \rho & \rho & 1 \end{pmatrix} \tag{2.4}$$

We require that ρ be non-negative because correlation is used to model the effect of a tool mark common to two pairs. The structure of the correlation matrix presented in (2.1) forces stricter boundaries on the range of values for ρ. In particular for $n \geq 3$,

$$Var(y_{01} - y_{12} + y_{23} - y_{03}) = 4\sigma^2(1 - 2\rho) \tag{2.5}$$

which implies that $\rho < 0.5$. Combining this with our requirement that ρ not be negative, we can say $\rho \in [0, 0.5)$.

2.3.2 Likelihood Analysis

The model described previously suggests there could be separate means for the two available samples, field-lab comparisons and lab-lab comparisons, if they were made by different tools. We saw that this difference in means is supported by our data in Figures

2.5(b) since the field-lab comparisons are markedly smaller than the lab-lab comparisons. To test whether the same tool made the field and lab tool marks, we can set up a test for the hypotheses

$$\mathrm{H}_0 : \mu_0 = \mu_1 \text{ vs } \mathrm{H}_A : \mu_0 < \mu_1. \tag{2.6}$$

Using normal model theory and generalized least squares, maximum likelihood estimates (MLEs) can be easily derived under the null model for $\mu(= \mu_0 = \mu_1)$ and σ^2 given a value for ρ as follows

$$\hat{\mu}|\rho = (\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \tag{2.7}$$

$$\hat{\sigma}^2|\hat{\mu}, \rho = (\mathbf{y} - \mathbf{X}\hat{\mu})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mu})/N \tag{2.8}$$

where $\mathbf{X} = \mathbf{1}_N$. The constraint in the alternative model requires the use of constrained least squares to maximize $\boldsymbol{\mu} = (\mu_0, \mu_1)'$. The `R` function `pcls()` within the `MGCV` package does this for us (Wood 2012). To fit the alternative model for a fixed ρ and using $\mathbf{X} = \begin{bmatrix} \mathbf{1}_n & 0 \\ 0 & \mathbf{1}_{N-n} \end{bmatrix}$, we use `pcls()` to obtain estimates of μ that maximize the likelihood and use (2.8) to estimate the variance. Finally we use a grid search for ρ and can compute parameter estimates as the values that maximize the normal likelihood.

Using a likelihood ratio test (LRT) for the null and alternative models described in (2.6), the resulting p-value will determine whether or not there is evidence of a match. The likelihood ratio statistic is defined as $\lambda = \dfrac{\ell(\hat{\mu}, \hat{\sigma}^2, \hat{\rho})}{\ell(\hat{\mu}_0, \hat{\mu}_1, \hat{\sigma}^2, \hat{\rho})}$, where $\ell(\cdot)$ denotes the normal likelihood function for the indicated set of parameter estimates. Due to the inequality constraint in the alternative hypothesis, the asymptotic distribution of $2\ln(\lambda)$ is a chi-bar distribution Chernoff (1954). For this particular single constraint, it is a mixture with half of the density on a point mass at 0 and half of the density on a chi-squared distribution with 1 degree of freedom since there is only one parameter difference in the models.

To demonstrate this process, we return to the data that was used in Figures 2.5 and 2.6. In the known match examples, Figures 2.5(a) and 2.6(c), we aggregated many data sets that were each made by the same tool under the same set of lab conditions. One of those sets has been chosen to demonstrate the LRT. Recall we have $n = 3$ marks (one field mark and three lab marks), resulting in a total of three field-lab comparisons and three lab-lab comparisons. Table 2.1 shows the parameter values and maximized log likelihood for both the null and alternative models. For these data $-2\ln(\lambda) = 0.941$ resulting in a large p-value of 0.166. We fail to reject the null hypothesis that the means are equal and conclude there is no evidence that the tools are different.

	Null Model	Alternative Model
$\ln \ell(\hat{\boldsymbol{\mu}}, \hat{\sigma}^2, \hat{\rho})$	-4.457	-3.986
$\hat{\rho}$	0.450	0.450
$\hat{\mu_0}$	2.524	1.870
$\hat{\mu_1}$		3.178
$\hat{\sigma}^2$	2.949	2.520
$-2\ln(\lambda) = 0.941$, p-value $= 0.166$		

Table 2.1: MLEs for matching data displayed in Figure 2.5(a).

For the non-match data from Figures 2.5(b) and 2.6(a), the field tool mark was chosen from a different screwdriver than the lab marks it was tested against. One set of non-matching data has been chosen out of the data shown for this example. Now we have $n = 4$ so there was one field mark and four lab tool marks for a total of 10 data values. Table 2.2 shows the parameter estimates from maximizing the log likelihood for these data under the null and alternative hypotheses and the likelihood ratio statistic, $-2\ln(\lambda) = 8.378$. The small p-value of 0.002 indicates we should reject the null hypothesis that the means of the two samples are equal. From this we would

conclude there is strong evidence that the two samples were created using different tools.

	Null Model	Alternative Model
$\ln \ell(\hat{\boldsymbol{\mu}}, \hat{\sigma}^2, \hat{\rho})$	-12.508	-9.082
$\hat{\rho}$	0.100	0.050
$\hat{\mu_0}$	1.267	-0.669
$\hat{\mu_1}$		3.119
$\hat{\sigma}^2$	4.609	1.955
$-2\ln(\lambda) = 8.378$, p-value $= 0.002$		

Table 2.2: MLEs for non-matching data displayed in Figure 2.5(b)

The two examples provided show two estimates for the correlation coefficient which are both non-zero and happen to span the range of possible values. To show the significance of the correlation coefficient in these models, we consider all the examples of matches that were discussed previously. The correlation coefficient from all matching data sets are shown in Figure 2.7 grouped by the tool that was used. We can see that the correlation for examples made by a given tool are all very similar. We also note that all of the correlation values are within the range and different from zero.

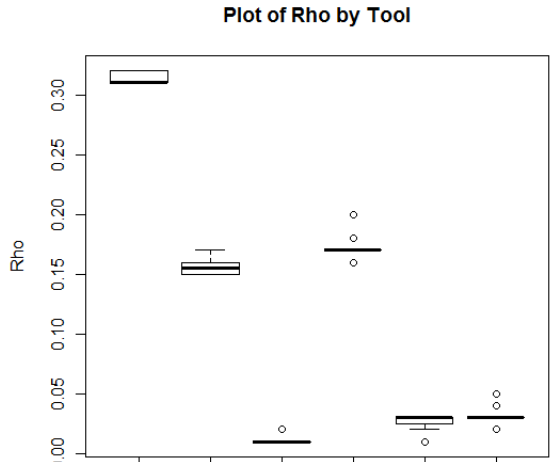
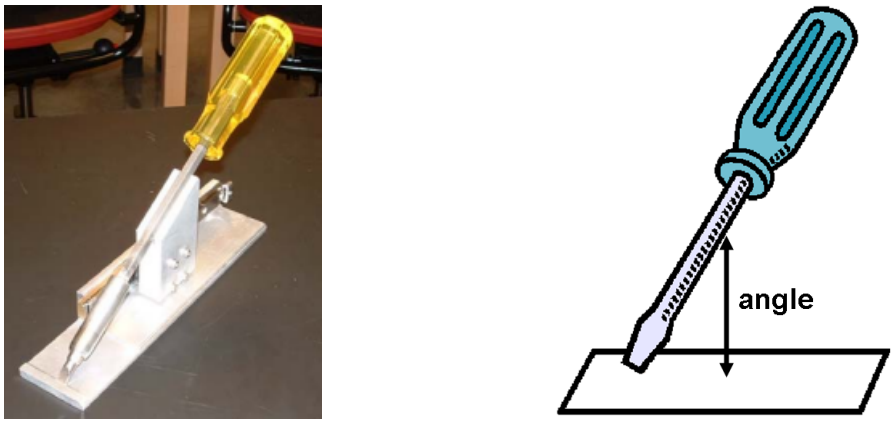


Figure 2.7: Correlation coefficients for all matching examples grouped by tool.

2.4 Angle Model

2.4.1 Angle Influence

We saw from the examples in Section 2.3.2 that the likelihood analysis is effective in providing evidence of a match between tool marks when the marks are made under controlled lab settings in which they are all produced the same way. However, to be useful in practice, it must also perform well when tool marks are made under different conditions or more specifically, for unknown field conditions. When a screwdriver is scraped against a metal surface, specific circumstances, such as the pressure exerted on the tool and the angle between the tool and the surface, can affect the appearance of the tool mark. Here we will consider the angle at which a mark is made, a measurable quantity that can be analyzed and accounted for to enhance the model and analysis described in Section 2.3. The jig shown in Figure 2.8(a) can be used to make controlled tool marks in the lab at various angles. For clarity, the angle at which a mark is made is measured as the smallest angle the tool makes with respect to the marked surface, illustrated in Figure 2.8(b).



(a) (b)

Figure 2.8: (a) Photograph of the jig used to make tool marks at specific angles in the lab. (b) Visual defining the angle between a screwdriver and a marked surface.

Although there are many attributes of a tool mark that could effect the tool mark comparisons, tool angle is the most consistently cited and accounted for in comparison models. The NIJ training manual lists a "number of variables [that] must be considered" when making tool marks in a lab to compare to a field mark. The variables they list are the action, amount of force, direction the tool moved, angle and physical circumstances. A North Carolina State Crime Lab procedural manual also lists physical features that must be considered which are type of mark, width/diameter of the tool, directoin of motion, angle, trace evidence and irregularities. Burd and Kirk (1942) state the factors that will influence the character of the mark are the degree of edge irregularity, vertical angle, horizontal angle, change of vertical or horizontal angle, change of direction, presence of debris and type of material the mark is made on. However, they further state "Of these factors only [vertical angle, horizontal angle and change of vertical or horizontal angle] need to be considered in detail," (p. 681). Their reasoning is that the other factors either will change over time, or complicate the comparison process but do not invalidate the match. In their study, Bachrach et al. (2010) empirically tested the effect of angle and medium that a mark is made on and concluded "the variation of the angle of attack has a significant effect on the resulting tool mark even if the medium is the same."

In the case of the Chumbley algorithm, Chumbley et al. (2010) demonstrated that when two tool marks are made by the same tool, similarity indices are generally much larger when the tool angle is the same in each case. To demonstrate this effect with our own data, we return to the data we have available. It was previously mentioned that for each tool, four marks were made under the same set of lab conditions and this was repeated for five different lab conditions. The five lab conditions correspond to different tool-surface angles, specifically 30°, 45°, 60°, 75° and 85°. To illustrate how the angle affects the comparison values, pairwise comparisons, y_{ij}, were made between all of the 120 available marks for each tool mark across the five different angles. This process was

repeated for all six screwdrivers available; Figure 2.9 displays boxplots of the comparison values for all pairs of marks made by a common tool grouped by the difference in the tool angles.

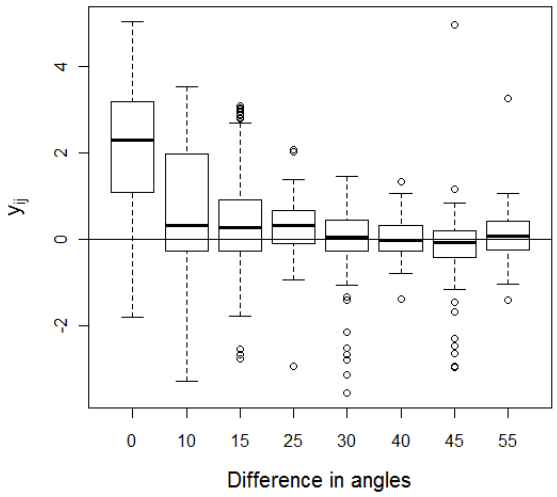
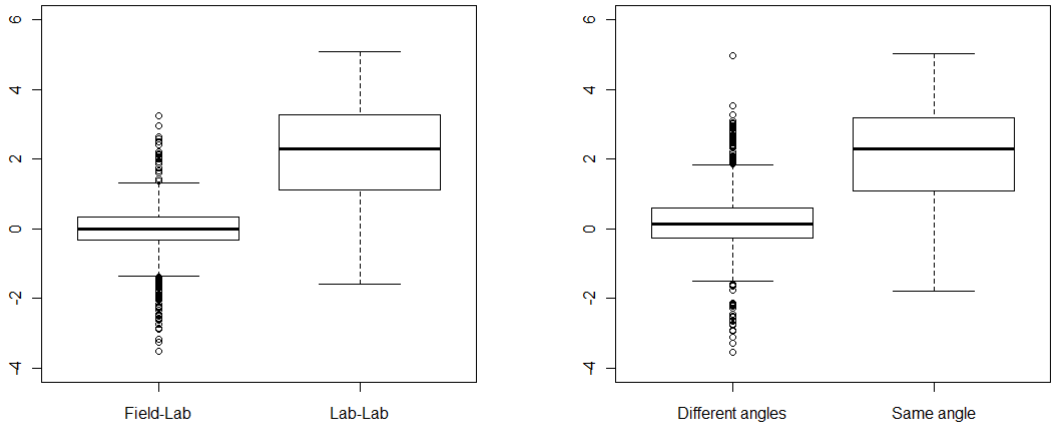


Figure 2.9: Boxplots for comparisons from the same tool made at different angles.

We can see that when tool marks are made at the same angle, the boxplot resembles that of the matches from Figure 2.5 since it is centered around 2 and has a standard deviation around 1. However, as the difference in angle gets larger, the boxplots all center around zero, resembling the non-matching boxplot, Figure 2.5(b). Specifically, this seems to be true for tool marks that were made at angles differing by more than 10°. In other words, even when marks are made by the same tool, if the tool angles differ by more than 10°, the comparison values resemble those from non-matching tool marks. Bachrach et al. (2010) found a similar result in their study when they state "the total [empirical] error rates are pronounced enough that comparison of tool marks created at 15° with those created at 45^{circ} is not better than guessing"

To better demonstrate the similarities between comparing matches to non-matches and comparing matches made at the same angle to those made at different angles, we

refer to Figure 2.10. For comparing matches to non-matches, we recall the non-match boxplot, Figure 2.5(b), which is shown again in Figure 2.10(a). We then took all the matching comparisons used in Figure 2.9 but grouped all the data made from the same angle together, and those made from different angles together; the boxplots are shown in Figure 2.10(b). Side-by-side, the similarities between comparing matches to non-matches, and comparing matches made at the same and differing tool angles is more apparent. We can see that there are more larger data values in the different angles boxplot of Figure 2.10(b) than in the field-lab comparisons of the non-match boxplot in Figure 2.10(a). This is not surprising since we noted in Figure 2.9 that some of the comparisons made by tool marks only differing by 10° still slightly larger than those differing by more degrees. Overall, these boxplots further show that for tool angles that differ by 10° or more, the data are no longer identifiable as a match.



(a) Boxplots of non-matches.

(b) Boxplots of marks made using the same tool, at the same and different angles.

Figure 2.10: Boxplots showing the similarities between data from different tools made at the same angle and data from the same tool made at different angles.

Knowing that tool angle has a significant effect on the data, it is important to generalize the approach described in Section 2.3 to account for these effects. One difficulty in incorporating angle information is that it is impossible to know the tool angle that was used to make a mark left at the crime scene. However, in a lab, tool marks can be

made by the suspect tool at any angle to try to better match a crime scene mark. Note here that with current procedures, tool mark examiners often do make marks at multiple angles for this purpose. If enough tool marks are made in the lab at angles differing by 10° or less, one or more of them should be expected to yield a high comparison value to the field mark if it was made with the same tool. Likewise, if the field mark does not match the lab marks well at any angle, we can conclude that the tool marks were made by different tools.

2.4.2 Model with Angle

Before we modify the basic model of Section 2.3 to account for angle information, we need to introduce more notation. Let a_i be the tool angle, in degrees, at which tool mark x_i is made for $i = 0, 1, ..., n$. Since we know that similar angles between two matching tool marks tend to produce relatively large values of y_{ij}, we will incorporate tool angles as a function of their difference.

We saw from Figure 2.9, that the mean response for data values is large when the angles are the same and approaches zero as the difference in angles increases. We also observed that comparisons of matching tool marks made at angles differing by more than 10° resemble non-matches. The function we chose to represent the difference in angles was $d(a_i, a_j) = \exp[-\theta(a_i - a_j)^2]$. We chose this function so that $d(a_i, a_j) = 1$ when tool angles match, and approaches zero as $|a_i - a_j|$ increases. Although we would prefer to estimate θ in the model, the relatively small amount of data available to us, in particular the small number of available angles, makes this difficult. For this reason, we chose to fix $\theta = 0.01$ which makes $d(a_i, a_j)$ much closer to zero when the difference in angles is 10 degrees or more. Figure 2.11 shows this behavior of $d(a_i, a_j)$ as a function of the difference in angles, $|a_i - a_j|$.

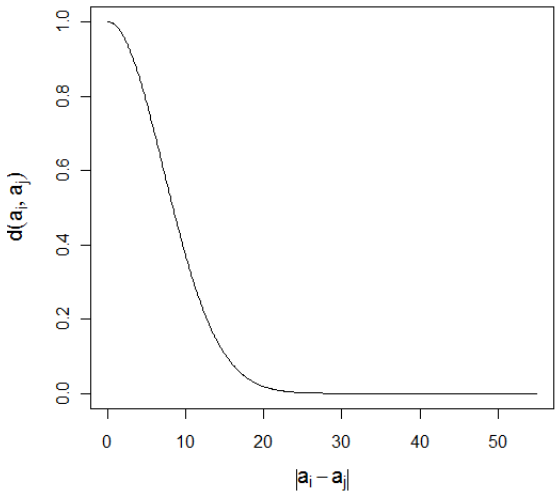


Figure 2.11: Plot of $d(a_i, a_j)$ as a function of the absolute distance between angles, $|a_i - a_j|$.

A modified data model incorporating tool angle is

$$
\begin{aligned}
&y_{ij} \sim N(\mu_{ij}, \sigma^2) \text{ where} \\
&\mu_{0j} = \mu + \alpha_0 d(a_0, a_j) \text{ for } j = 1, ..., n \\
&\mu_{ij} = \mu + \alpha_1 d(a_i, a_j) \text{ for } i, j = 1, ..., n \text{ and } i < j \\
&\text{subject to the constraint } \alpha_0 \leq \alpha_1.
\end{aligned} \tag{2.9}
$$

Each y_{ij} is normally distributed with a separate mean determined by whether the comparison is a field-lab comparison, μ_{0j}, or a lab-lab comparison, μ_{ij}. In both types of comparisons, μ represents the baseline mean of all the y_{ij}s computed from tool marks made at different angles. This mean tends to be near zero. The model also includes an additional mean component based on the similarity measure $d(\cdot, \cdot)$; α_0 represents the difference in means from the field-lab comparisons and the baseline mean, and α_1 represents the difference in means between the lab-lab comparisons made at the same angle and the baseline mean. If the field mark and lab marks were made by the same tool, both α_0 and α_1 should be around 2 or 3. However, note that the field-lab comparisons should fit as well as, but no better than the lab-lab comparisons. For this reason, we

have placed the restriction that $\alpha_0 \leq \alpha_1$. If the tool marks were not made by the same tool, α_1 should stay around 2 or 3, but α_0 should be considerably smaller than α_1, and close to zero.

Using this model, we can again make inference about whether or not the suspect tool made the crime scene marks with a likelihood ratio test. However, a difference in means now will be determined by whether α_0 is less than α_1. We are interested in comparing the hypotheses

$$\mathrm{H}_0 : \alpha_0 = \alpha_1 \text{ vs } \mathrm{H}_A : \alpha_0 \leq \alpha_1. \tag{2.10}$$

The alternative model was defined in (2.9); the null model does not differentiate between field-lab and lab-lab comparisons and is defined as

$$\begin{aligned} &y_{ij} \sim N(\mu_{ij}, \sigma^2) \text{ where} \\ &\mu_{ij} = \mu + \alpha d(a_i, a_j) \text{ for } i, j = 0, 1, ..., n \text{ and } i < j. \end{aligned} \tag{2.11}$$

Both the null and alternative models assume that we have known angles for every tool mark made in the lab; that is $a_1, ..., a_n$ are known. The angle of the mark made in the field, a_0, is unknown. Thus, a_0 is a parameter in the model along with μ, α, α_0, α_1, σ^2 and ρ. The correlation structure described in Section 2.3.1 remains for this model and ρ will still be chosen using a grid search between 0 and 0.5. Since we know the tool angle needs to be accurate within 10° to see evidence of a match, we will perform a grid search for a_0 in increments of 5° between 20° and 90°. These angle bounds were chosen as reasonable angles for which a viable tool mark could be made.

Maximum likelihood estimates for μ, α and σ^2 in the null model are computed using weighted least squares provided values of a_0 and ρ, as in the basic model. In the alternative model, the inequality constraint on α_0 and α_1 requires the use of constrained least squares fit to maximize μ, α_0 and α_1. To fit the null model, let $\mathbf{d_0} = (d(a_0, a_1), d(a_0, a_2), ..., d(a_0, a_n))'$ and $\mathbf{d_1} = (d(a_1, a_2), d(a_1, a_3), ..., d(a_{n-1}, a_n))'$. Then

the MLEs can be computed as

$$\hat{\boldsymbol{\beta}}|a_0, \rho = (\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \tag{2.12}$$

$$\hat{\sigma}^2|\hat{\boldsymbol{\beta}}, a_0, \rho = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/N \tag{2.13}$$

where $\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha \end{bmatrix}$ and $\mathbf{X} = \begin{bmatrix} \mathbf{1}_N & (\mathbf{d_0}', \mathbf{d_1}')' \end{bmatrix}$. To fit the alternative model for fixed ρ and a_0, we first use `pcls()` to obtain maximized estimates for $\boldsymbol{\beta} = (\mu, \alpha_0, \alpha_1)'$ by supplying the data matrix $\mathbf{X} = \begin{bmatrix} \mathbf{1}_n & \mathbf{d_0} & \mathbf{0}_n \\ \mathbf{1}_{N-n} & \mathbf{0}_{N-n} & \mathbf{d_1} \end{bmatrix}$. We then use (2.13) to estimate the variance component. Likelihood values are computed for parameters computed this way over a grid of ρ and a_0.

As with the basic model, the null distribution of the likelihood ratio statistic from comparing these angle models will follow a chi-bar distribution with half of the density on a point mass at 0 and half of the density on a chi-squared distribution with 1 degree of freedom. As a result, when doing many tests, if the null hypothesis is true, we would expect 50% of the p-values to be approximately uniform between 0 and 0.5, and the other half to have a point mass at 1. If the alternative hypothesis is true, we would still expect the p-values to be systematically smaller.

2.5 Results

To test the modified models, we used all the data we had available: four tool marks made at each of the possible tool angles of 30°, 45°, 60°, 75° and 85° for all six screwdrivers. Thus we have a total of $5(angles) \times 4(marks) \times 6(tools) = 120$ tool marks. We will consider two scenarios, one where the field mark has been chosen out of the available marks made by the same tool which reflects a situation where the analysis should indicate a match and a second scenario where the field mark was chosen from

the marks made by a different tool. The matching results are discussed in Section 2.5.1 and the non-matching results are discussed in Section 2.5.2.

2.5.1 Results for Matches

Data sets for matches were compiled using all 20 tool marks for a given tool: four marks from each of the five available angles. Each tool mark within a set was chosen one-at-a-time to be the field mark leaving the remaining $n = 19$ lab marks for comparison, three of which are made at the same tool angle as the field mark. This process was repeated for all six tools and the likelihood ratio test described in Section 2.4.2 was performed on each for a total of 120 tests.

Since we know all the tool marks used in each analysis are made by the same tool, we would expect the field-lab comparisons and lab-lab comparisons to result in similar data values as long as the field angle has been estimated correctly. Thus only one regression slope for the similarity measure, $d(a_i, a_j)$, would be needed for an adequate model fit so the null model described in (2.11) should fit the data as well, or nearly as well, as the alternative model (9). Based on this assumption, we would expect that half of the distribution of p-values from these likelihood ratio tests should be approximately uniform between 0 and 0.5 with the other half of the p-values to have a point mass at 1. A hypothesis test was performed to test whether or not the proportion of p-values with a density of 1 was equal to 0.5. The resulting p-value was 0.2012, so we conclude the assumption is valid. Furthermore, a Kolmogorov-Smirnov test was performed on the p-values less than 0.5 comparing them to a Uniform(0, 0.5) distribution. This test also failed to reject the null hypothesis which states the distributions are the same with a p-value of 0.2801. Thus, the histogram of p-values for all 120 LRTs, shown in Figure 2.12, is consistent with the asymptotic distribution.

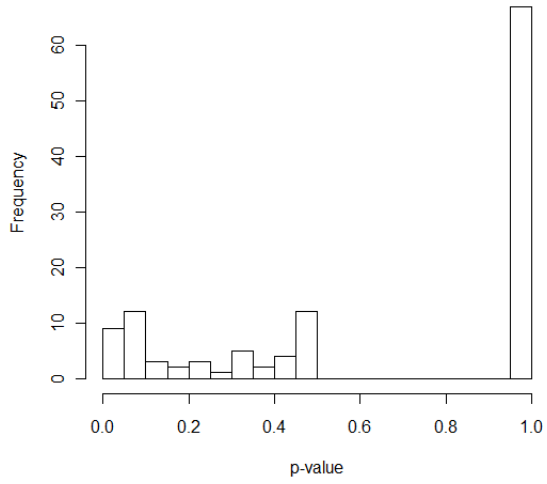
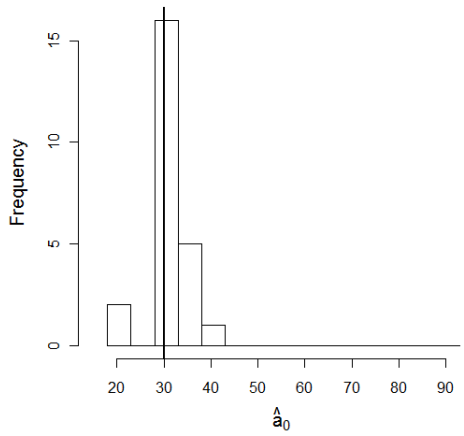
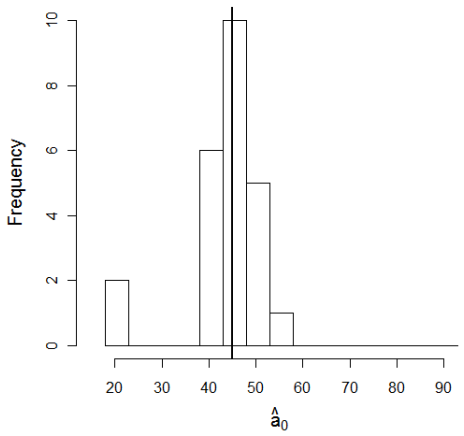


Figure 2.12: Histogram of p-values for all matching data.

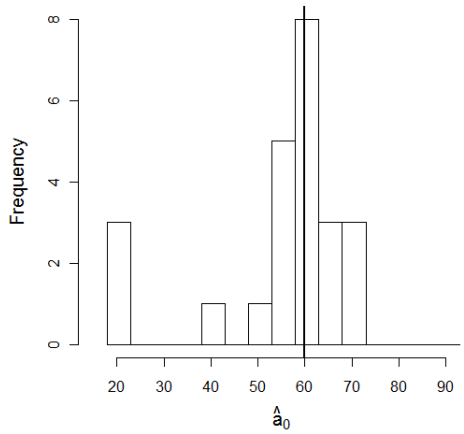
In addition to checking that the LRT is returning the results we would expect for matching data, it is also necessary to check that the angle estimation is performing correctly. We can check this by plotting the estimates for a_0 grouped by the actual value of the field angle. The five plots, one for each of the tool angles, are shown in Figure 2.13 with a vertical line representing the actual tool angle. Recall that the grid search used to estimate the field angle only considered angles at 5° increment between 20° and 90°.



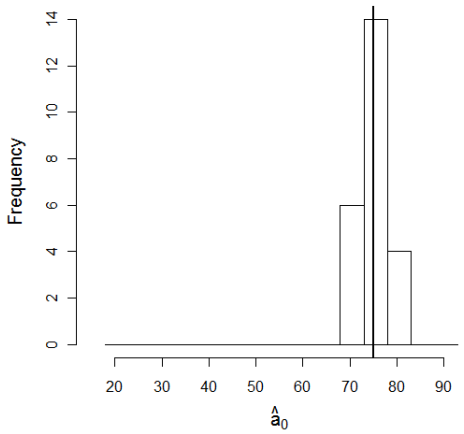
(a) Estimates $\hat{a}_0$ when $a_0 = 30°$



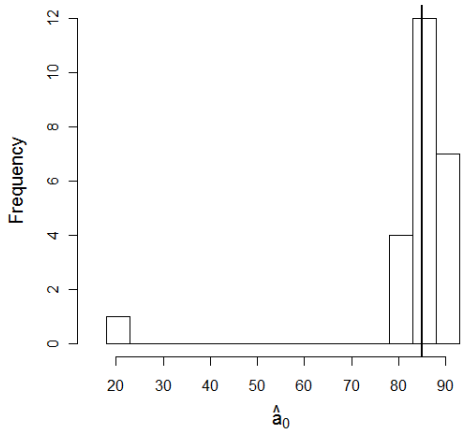
(b) Estimates $\hat{a}_0$ when $a_0 = 45°$



(c) Estimates $\hat{a}_0$ when $a_0 = 60°$



(d) Estimates $\hat{a}_0$ when $a_0 = 75°$



(e) Estimates $\hat{a}_0$ when $a_0 = 85°$

Figure 2.13: Estimated values of a_0 grouped by the true value of a_0 for matching data.

With the exception of the 60° tool marks, the estimation appears to be very accurate and rather precise. The angles are consistently estimated within five degrees of the true angle the majority of the time. Within the data available to us, a few marks were flawed so as to be dissimilar to others made by the same tool at the same angle. This occurred more in 60° tool marks than in any other angle. For this reason it is not surprising that those estimates are not as precise as the other angles. Overall, methodology based on the new model leads to expected test results and informative angle estimates when the tool marks are indeed matches.

2.5.2 Results for Non-Matches

To create non-matching data sets, we used all $n = 20$ tool marks made from the same tool as before and considered these the lab marks. The field marks were chosen out of the remaining five tools, one mark from each of the five angles available for each tool. The data sets were assembled by one-at-a-time comparing the field mark with each of the lab marks and comparing the lab marks pairwise with one another. This process was repeated for all six screwdrivers, resulting in a total of 150 non-match data sets.

For non-matching data, all of the field-lab comparisons should be close to zero regardless of the tool angles since the marks were made by different tools. However, the lab-lab comparisons that were made at the same angle should result in larger comparison values since they are true matches. This discrepancy should show up in the models through the α_0 and α_1 values. Thus we would expect the alternative model to be a better fit to these data and so the p-values from the likelihood ratio tests should be small, i.e., the distribution of p-values should be skewed with greater frequencies associated with smaller p-values. The p-values that resulted from these 150 tests are shown in the histogram in Figure 2.14. As expected, the p-values are mostly small and have an overall right skewed shape. This supports the alternative hypothesis, and is interpreted as evidence that the lab and field marks were made with different tools.

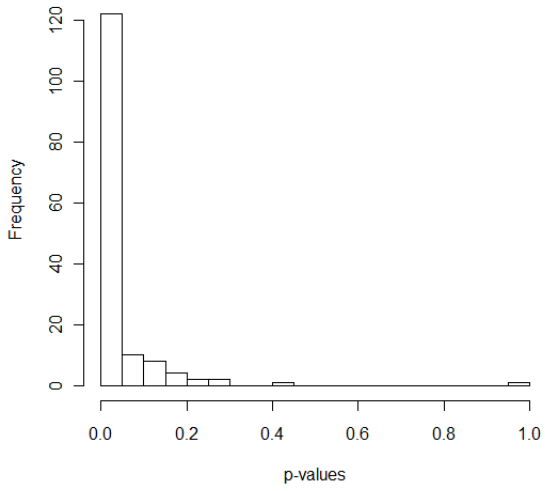


Figure 2.14: Histogram of p-values for non-matching data.

2.6 Conclusion

There has been work done by several researchers to show the difference in tool marks quantitatively. The Chumbley algorithm is one example that is effective qualitatively at distinguishing between a match and non-match in tool marks. However, the null distribution of the standardized U-statistic varies depending on many factors. The analysis presented in this paper overcomes this problem by simultaneously modeling Chumbley indices from comparisons of a field sample with multiple lab samples. This approach might also be used with other indices that have approximately normal distributions.

We have shown that the angle at which a tool is used can have a significant effect on the similarity of the tool marks. Since it is these similarities that are crucial to the tool mark matching process, it is necessary to account for these effects in the models and analyses used to examine tool marks. Including the tool angle in the model as a function of the difference in angles of marks being compared does seem to yield positive results, both in the ability of the likelihood ratio tests to choose an appropriate model, reflecting a match or non-match, and in estimating the unknown field angle when the field and lab tool marks are actually matches. Our research indicates that as long as

lab tool marks are made at angles within 10° of the field angle, the estimation process is reasonably accurate when the tool marks match, and the likelihood ratio test performs appropriately for both matching and non-matching cases.

Due to the constraints of available resources, we had access to tool marks made at only a few angles, those being 30°, 45°, 60°, 75°, and 85°. As a result, we chose a parameter value for θ in the similarity function of the modified models. Having tool marks made at more angles would allow us to include θ in the estimation process and might also improve the precision with which a_0 could be estimated when lab and field marks are made with the same tool. A case can be made for simply setting α_0 to zero in the expression of the alternative hypothesis. Here, we've left this parameter value unspecified so that the hypotheses are nested, and so simplifying the likelihood ratio test.

The research reported here, while promising, should be viewed as preliminary. A broad and impressive collection of forensic laboratory techniques have been developed for a number of important evidence matching settings, including DNA, fingerprint, and material composition, as well as the tool impression setting described here and the closely related application of ballistic evidence. In many of these areas, the development of appropriate statistical methodology is still needed. While our work is motivated by real aspects of the tool mark comparison problem (the many-to-one nature of available lab and field marks, and the demonstrated importance of tool angle), further refinement is clearly needed before application can be made to forensic practice, for example:

- Are the parametric forms we've chosen in for our models adequate for this physical setting, or would other forms be more appropriate?

- Our empirical work was limited to tool marks made in the laboratory, under controlled conditions. Are further model features needed to account for the fact that the field mark is never really produced this way, and may often (or always) be

subject to additional sources of noise?

- While we've focused on data produced by relatively mature profilometry techniques, newer non-contact measurement processes based on confocal microscopy offer the potential for 3-dimensional mapping of a tool surface and computer generated "virtual tool marks" for almost any conceivable collection of physical conditions, including tool angle, but also applied force, Ekstrand et al. (2013) . This technology is new and not yet in wide-spread use in forensic laboratories, but it likely will be available soon. How can the methodology we've outlined here be extended to make use of the much larger and more diverse sets of synthetic lab marks that will soon be available?

Finally, we need to point out that our development leads to a statistical inference which is structurally different from what would be most appropriate in forensic analysis. Tool mark examiners properly think of their evaluations as leading to a declaration of "match" or "not enough evidence to classify as a match". Our development aligns the definitive statistical statement "reject the null hypothesis" with the non-definitive forensic conclusion "not enough evidence to classify as a match". Conversely, the non-definitive "failure to reject H_0" aligns with the definitive "match". Despite this structural difficulty, our approach is a first logical step statistically because "match" corresponds to a simpler statistical model, while the alternative requires the more complex model. Our current research focuses on a recasting of this problem to reverse the roles of the hypotheses so that the more important forensic error - false declaration of a "match" - corresponds to the Type I error.

CHAPTER 3. EXAMINING THE EFFECTS OF TOOL MARK QUALITY

A paper to be submitted the *Journal of Forensic Science*

Amy B. Hoeksema [1] [2] and Max D. Morris[3] [4]

Abstract

Suppose a crime is committed such that a tool mark is left at the crime scene, and a suspect tool is identified. Forensic examiners assess the strength of evidence that the suspect tool was used in the crime by comparing the crime scene mark to marks made in the laboratory with that tool. In current practice, for stronger support of a match or non-match, multiple marks are made in the lab under the same conditions by the suspect tool. However, through flaws in the mark surface or error in the mark making process, it is possible to make tool marks under the same conditions using the same tool that do not resemble one another. Thus it is necessary to incorporate a quality control step in the tool mark matching process. Toward this end, we describe a method that could be used to verify that all the lab marks made do in fact match each other well enough to be considered reliable for comparing to a field tool mark, or to identify those that should be eliminated.

[1]Graduate student, Department of Statistics, Iowa State University
[2]Primary researcher and author
[3]Department of Statistics, Iowa State University
[4]Department of Industrial and Manufacturing Systems Engineering, Iowa State University

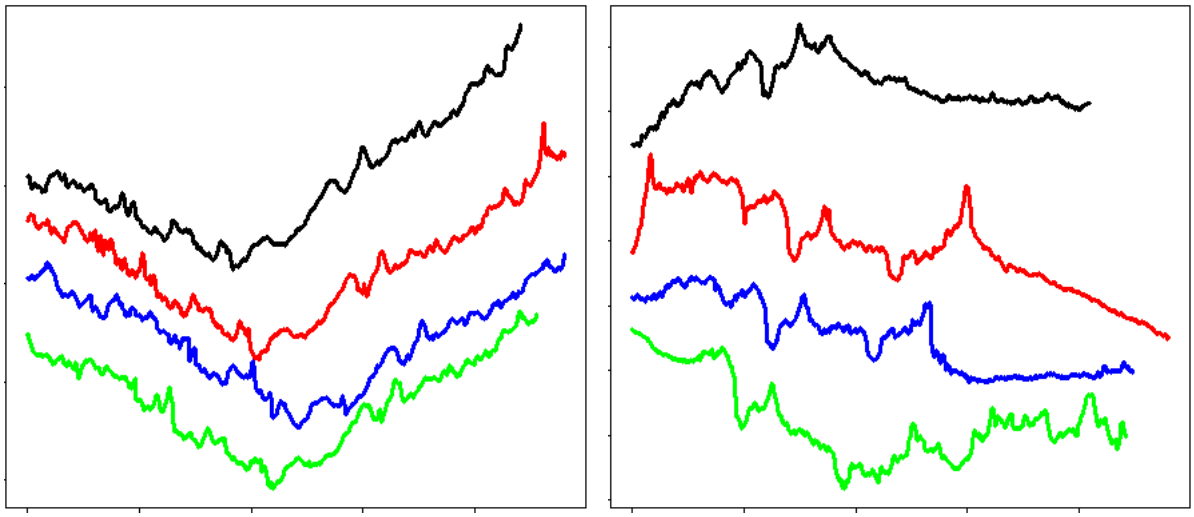
3.1 Introduction

When a tool is used in the commission of a crime, the tool mark that may be left at the scene can be valuable forensic evidence. Although tool mark comparison practices apply to marks made by any striated tool, for the purposes of this paper, we will focus on tool marks that have been made by screwdrivers; for example, a screwdriver may be used to pry open a window with a metal frame and an impression mark may be left behind. If a suspect tool is identified, forensic examiners can make tool marks in the laboratory using the suspect tool and compare them to the one left in the field. Since the comparison of a lab tool mark to a crime scene mark only provides a single data value, it is common for examiners to make multiple marks in the lab that are each compared to the field tool mark (National Institute of Justice).

Although lab tool marks are made under controlled conditions, flaws in the surface on which the marks are made and errors in the mark making process lead to variability in the resulting tool marks. Small variations in tool marks are to be expected, however, we have seen that occasionally a single "bad" tool mark can result that does not resemble the others made under the same conditions. When forensic examiners compare tool marks, they look at both the overall shape and pattern of the tool mark, which they call the *class* or *subclass characteristics*, as well as the small scale noise or *individual characteristics*. Different tools of the same type are apt to have the same or similar class and subclass characteristics, but it is assumed that individual characteristics are unique to each tool. Although we expect there to be some small differences in the individual characteristics of matching tool marks, they should at least match in class and subclass characteristics.

Figure 3.1 contains two sets of profilometer traces, each of multiple tool marks made in the lab with a common tool under the same conditions. For each screwdriver, four separate tools marks were made and then scanned using a stylus profilometer. In Figure

3.1(a), all four tool marks are relatively parallel and display similar features. We can see from these marks that all four display the same class characteristics, or large peaks and valleys, and are also very similar in the smaller perturbations. Although there are some differences in the noise, the primary signal or main features of the marks are the same. In Figure 3.1(b), we can see that the red, green and blue tool marks are all relatively similar, as in the previous panel. However, the tool mark shown in black seems to match the other three over only half the length of the tool mark, and then tapers off into noise without signal. (This likely indicates that the tool mark was incomplete, and that the profilometer stylus reached the end of the tool mark and was scanning the lead plate on which the mark was made.) As a result, the black tool mark would appear to be made by a different tool, although it was not.



(a) Example of well matching lab tool marks.

(b) Example of poorly matching lab tool marks.

Figure 3.1: Examples of tool marks made in the laboratory under identical conditions.

If a numerical matching algorithm, such as that proposed by Chumbley et al. (2010), is applied to pairs of tool marks depicted in Figure 3.1(a), the tool marks are positively identified to be made by the same tool. However, when the same algorithm is applied

to the tool marks in Figure 3.1(b), the indication is that these marks were not all made by the same tool. Although all marks are known to be made by the same tool, we get a false indication of non-match because of the poor quality of one lab tool mark.

The collection of marks displayed in Figure 3.1(b) is just one example of how "bad" marks can be made in the lab. If a lab mark is produced that does not accurately reflect the suspect tool, comparisons of that mark to the evidence mark can easily lead to a "non-match" conclusion when the correct conclusion is "match." With multiple lab marks, this danger is lessened because a single "non-match" comparison carries less weight. But if such eliminations are made subjectively, without well-grounded rules, this weakens the credibility of the forensic examination process. It is therefore important to develop objective processes to eliminate potentially flawed lab marks before comparison to the evidence mark begins.

Based on these results, we can see the importance of incorporating a quality control step in the current matching process. The purpose of this quality control component is to determine whether any of the lab marks should be eliminated as inconsistent with the rest, to avoid possible misleading comparisons to the evidence mark. The method we propose is designed to operate on data that represents each comparison of two tool marks as a single numerical index. For the purposes of this paper, we discuss using this method with data produced with Chumbley's algorithm. The details of the algorithm can be seen in Chumbley et al. (2010) with a further introduction into the use of likelihood ratio tests on tool mark comparisons in Hoeksema and Morris (2013). We provide a brief summary of those papers supplying only the pertinent notation and details of the process in Section 3.2. We then describe a model and analysis that can be used for quality control in Section 3.3, with numerous examples in Section 3.4, and conclusions and suggestions for future work in Section 4.5.

3.2 Basic Model

For consistency, we will use the same notation as Hoeksema and Morris (2013); the parts necessary for this paper are provided. Each physical tool mark is represented by a digitized profilometer scan of the depths of the grooves of the striae plotted against pixel location (e.g. as displayed in Figure 3.1). Since we are only interested in the lab tool marks for quality control, we focus on notation for lab tool marks and do not include field tool marks. Let x_i, $i = 1, ..., n$, represent the ith digitized lab tool mark. Then $y_{ij}, i, j = 1, 2, ..., n$ and $i < j$, represents the index of comparison of tool mark i to tool mark j, as described by Chumbley et al. (2010). For our purposes, a complete data set contains $N = \binom{n}{2}$ comparisons.

Computed comparison values are scaled Mann Whitney U-Statistics, and we assume that the y_{ij} are approximately normally distributed. Let $E(y_{ij}) = \mu$ and assume the y_{ij} have a common variance, σ^2. Some pairs of comparisons share a common tool mark, and the model contains a non-zero correlation between such comparisons:

$$Corr(y_{ij}, y_{kl}) = \begin{cases} 0 \text{ if } i \neq k, i \neq l, j \neq k \text{ and } j \neq l \\ \rho \text{ if } i = k \text{ or } i = l \text{ or } j = k \text{ or } j = l, \text{ but not } (i, j) = (j, k) \\ 1 \text{ if } i = k \text{ and } j = l. \end{cases} \tag{3.1}$$

Let $\mathbf{y} = (y_{12}, y_{13}, ..., y_{n-1,n})'$ be the N-vector of all data values. Finally, we further assume that the joint distribution of all pairwise comparisons of lab tool marks is multivariate normal, specifically $\mathbf{y} \sim N(\mu \mathbf{1}'_{\mathbf{N}}, \sigma^{\mathbf{2}}\mathbf{R})$, where the elements of $\mathbf{R}$ are defined as in equation (3.1).

In Hoeksema and Morris (2013), the focus was on determining a match between a field tool mark and a suspect tool based on whether the lab-lab comparisons have the same mean as the lab-field comparisons. However, if the lab tool marks do not match each other well, the distribution of lab-lab comparisons will not reflect that of a true match. That is, any y_{ij} that are comparisons that include a "bad" lab tool mark will

tend to be smaller than the other lab-lab comparisons, leading to a smaller fitted overall mean for that sample than is appropriate. Thus is it necessary to incorporate a quality control step in the matching process to assure the lab tool marks match each other well before comparing them to the field tool mark.

3.3 Quality Model and Analysis

Our proposed quality control check is to compare all the lab tool marks pairwise, then one-at-a-time isolate the comparisons involving a particular tool mark from the rest to see if it is an outlier relative to the remaining marks. To achieve this, we propose adding a "penalty" to the basic model for the mean of comparisons involving the selected tool mark which could increase as the quality of tool mark decreases. If the addition of this penalty to the model improves the likelihood significantly, this indicates that the identified tool mark is not enough like the others, and should be removed.

To implement this approach, we introduce more notation and modify the basic model for y_{ij}. Let γ_k represent the positive penalty for comparisons involving tool mark x_k, $k = 1, ..., n$. For each value of k, we will fit the model $E(y_{ij}) = \mu - \gamma_k 1_{i=k} - \gamma_k 1_{j=k}$, where $\gamma_k > 0$. Once tool mark k has been chosen for the penalty, any comparison that involves that tool mark will have a mean of $\mu - \gamma_k$, but the rest will have a mean μ. Since it is required that the penalty be non-negative, inequality constrained regression is used to fit the model with the constraint that $\gamma_k \geq 0$. Parameters μ, γ_k and σ^2 are estimated using maximum likelihood with inequality constrained regression using the `R` function `pcls()` within the `MGCV` package (Wood 2012). The correlation coefficient ρ is estimated using a grid search within $[0, .5)$. The normal likelihood is computed for each tool mark held out as the penalized mark and the log likelihoods are compared.

For a set of lab marks, we compare the largest likelihood to the baseline likelihood (i.e. that computed under the basic model with no penalty) to determine if the pe-

nalized mark is significantly different from the remaining marks, indicating that the penalty is necessary. Since constrained likelihood is used, the standard likelihood ratio test (LRT) statistic, $-2\ln(\lambda)$ where $\lambda = \dfrac{\ell(\hat{\mu}, 0, \hat{\sigma}^2, \hat{\rho})}{\ell(\hat{\mu}, \hat{\gamma}_k, \hat{\sigma}^2, \hat{\rho})}$ and $\ell(\cdot)$ denotes the normal likelihood function for the indicated set of parameter estimates, has a chi-bar(0,1) asymptotic distribution (Chernoff 1954). That is, the distribution of the LRT, under the null hypothesis of $\gamma_k = 0$, is a mixture with half of the density on a point mass at 0, and half on a chi-squared distribution with 1 degree of freedom. Since multiple likelihoods (for $k = 1, 2, ..., n$) are being compared simultaneously, we also use a Bonferroni correction to control the overall error rate of the procedure.

3.4 Examples

To test the proposed quality control method, 30 sets of lab tool marks were examined with the analysis described in Section 3.3. In each set, there were four lab tool marks made by the same tool under the same set of laboratory conditions. All pairs of tool marks in a set were compared to one another, resulting in six comparison values for a single dataset. The quality model was fitted to these six values using each tool mark as the penalized mark and the resulting log likelihoods were ordered from smallest to largest. The baseline model was compared to the each fitted model using a chi-bar critical value with an α level of 0.01. Incorporating the Bonferroni correction, the $\alpha/4 = 0.0025$ chi-bar critical value is 7.88, so any model with a LRT statistic larger than this would indicate the penalized mark is significantly different from the remaining three marks.

Five examples are described below which depict the range of results we observed. For each, we present a figure showing the four lab tool marks that were compared to one another as digitized tool marks. The resulting comparison values are provided in a table. Finally, a table containing the results of the model fitting ordered by LRT statistic is shown indicating which tool mark is being penalized, the fitted parameters and the LRT

statistic. If the largest likelihood is significantly different from the baseline likelihood, a dashed line is included in the table separating that model.

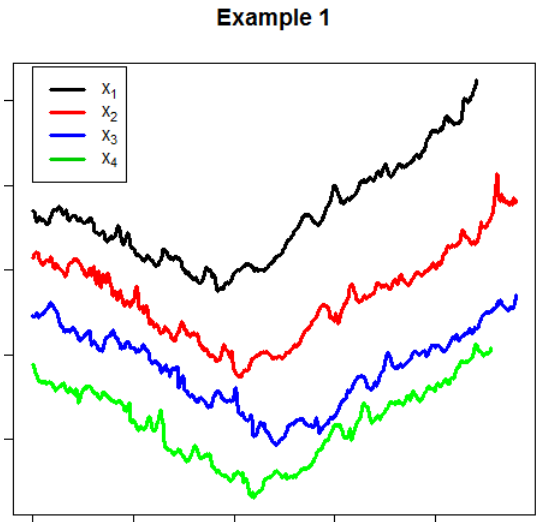


Figure 3.2: Example 1 lab marks.

y_{ij}	x_2	x_3	x_4
x_1	4.105	4.187	3.098
x_2		4.325	4.587
x_3			3.345

Table 3.1: Summary statistics from Example 1.

Penalized x_k	$\hat{\mu}$	$\hat{\gamma}_k$	$\hat{\sigma}^2$	$\hat{\rho}$	$-2\ln(\lambda)$
2	3.94	0.00	0.29	0.00	0.00
3	3.94	0.00	0.29	0.00	0.00
1	4.09	0.29	0.27	0.00	0.46
4	4.21	0.53	0.22	0.00	1.68

Table 3.2: Model estimates from Example 1.

The tool marks used in Example 1, shown in Figure 3.2 and repeated from Figure 3.1(a), appear to be well matching marks. The large comparison values in Table ?? confirm this impression since the smallest comparison value is 3.098 which indicates a strong match. In Table 3.2, we see the results of fitting the quality model with each of the four tools being given a penalty. For the first two marks, 2 and 3, the penalized models are equivalent to the baseline model since $\hat{\gamma}_2 = \hat{\gamma}_3 = 0$. Thus $\hat{\mu}$ in these two models represents the overall mean comparison value for the data. The model with the

largest log likelihood places a penalty on x_4, which we can confirm in Table ?? has the smallest comparison values. Since the largest LRT statistic is 1.66, which is much smaller than the chi-bar critical value of 7.88, we conclude that none of the tool marks differ significantly from one another. Thus, all four lab marks match each other well and could be further used to compare to a field mark.

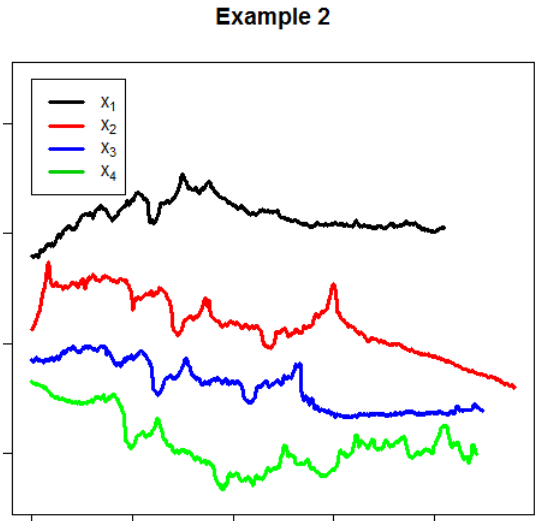


Figure 3.3: Example 2 lab marks.

y_{ij}	x_2	x_3	x_4
x_1	0.631	1.391	0.894
x_2		3.115	2.054
x_3			2.352

Table 3.3: Summary statistics from Example 2.

Penalized x_k	$\hat{\mu}$	$\hat{\gamma}_k$	$\hat{\sigma}^2$	$\hat{\rho}$	$-2\ln(\lambda)$
2	1.74	-0.00	1.13	0.45	0.00
3	1.74	0.00	1.13	0.45	0.00
4	1.74	0.00	1.13	0.45	0.00
1	2.51	1.53	0.15	0.00	8.58

Table 3.4: Model estimates from Example 2.

Example 2 shows four poorly matching tool marks. In particular, x_2 (red) and x_3 (blue), shown in Figure 3.3 (repeated from Figure 3.1(b)), visually appear to match each other well across the entire length of the tool mark. A third mark, x_4 shown in green, appears to match x_2 and x_3 over part of the trace but differs at the right end. The

fourth mark, x_1, however is significantly different from the other three since it seems to only have transfered part of the tool mark onto the surface. These visual observations are numerically confirmed in Table ?? since comparisons involving x_1 result in smaller values than the comparisons only invovling the other three tool marks.

Table 3.4 shows the results of fitting the quality model using all four marks. We can see that penalizing marks 2, 3 and 4 results in the baseline model since $\hat{\gamma}_2 = \hat{\gamma}_3 = \hat{\gamma}_4 = 0$, but penalizing mark 1 significantly increases the log likelihood. The LRT statistic from penalizing x_1 is 8.58 which is larger than the critical value, so this provides evidence that x_1 is significantly different from the other three tool marks. In this situation, the quality model successfully identifies that these lab marks do not match each other well and that x_1 should be removed from the set before comparing to the field tool mark.

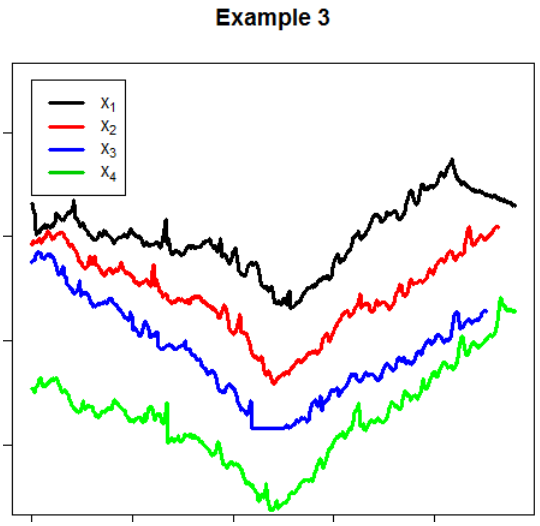


Figure 3.4: Example 3 lab marks.

y_{ij}	x_2	x_3	x_4
x_1	2.906	-0.126	2.628
x_2		1.141	3.443
x_3			0.503

Table 3.5: Summary statistics from Example 3.

Penalized x_k	$\hat{\mu}$	$\hat{\gamma}_k$	$\hat{\sigma}^2$	$\hat{\rho}$	$-2\ln(\lambda)$
1	1.75	0.00	1.82	0.45	0.00
2	1.75	0.00	1.82	0.45	0.00
4	1.75	0.00	1.82	0.45	0.00
3	2.99	2.49	0.28	0.45	11.32

Table 3.6: Model estimates from Example 3.

Example 3 shows a different type of error that can occur on lab tool marks which results in a false non-match classification. From Figure 3.4, we can see that marks 1, 2 and 4, shown in black, red and green, visually match each other well. Mark 3, shown in blue, also seems to match the other three well except in the middle of the tool mark. During the making of this tool mark, the mark was made too deep on the lead surface and as a result, when the stylus profilometer was recording the depths of the mark it reached its minimum recording value during this stretch of mark which recorded as the flatline shown.

Table ?? shows the summary statistics that confirm our observations since comparisons involving x_3 have the smallest comparison values. Table 3.6 indicates the quality model successfully determines mark 3 is significantly different from the other three marks and should be removed from the set of lab marks. Each of the other three tool marks result in a baseline model since $\hat{\gamma}_2 = \hat{\gamma}_3 = \hat{\gamma}_4 = 0$. When the penalty is applied to x_3, the LRT statistic is 11.32 which is larger than the Bonferroni corrected chi-bar critical value of 7.88.

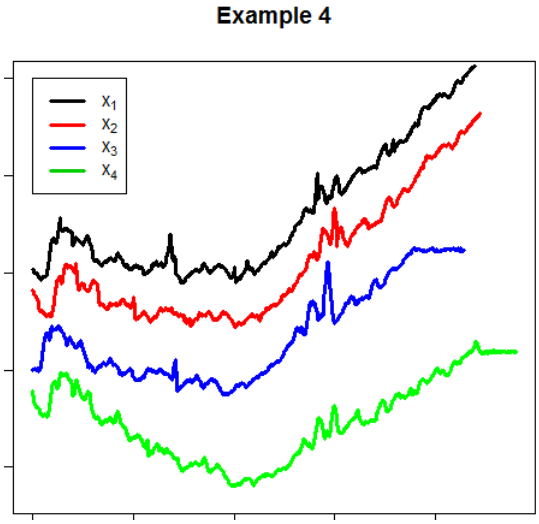


Figure 3.5: Example 4 lab marks.

y_{ij}	x_2	x_3	x_4
x_1	2.605	2.823	2.817
x_2		3.340	3.556
x_3			3.120

Table 3.7: Summary statistics from Example 4.

Penalized x_k	$\hat{\mu}$	$\hat{\gamma}_k$	$\hat{\sigma}^2$	$\hat{\rho}$	$-2\ln(\lambda)$
3	3.04	0.00	0.11	0.00	0.00
4	3.04	0.00	0.11	0.00	0.00
2	3.04	0.00	0.11	0.00	0.00
1	3.34	0.59	0.02	0.00	9.84

Table 3.8: Model estimates from Example 4.

Figure 3.5 shows four tool marks that appear to be very similar. The table of comparison values, Table ??, confirms that all four tool marks match each other very well since the smallest y_{ij} is 2.605, suggesting a strong match. However, when we examine the results of fitting the quality model to each tool mark in Table 3.8, three marks result in LRT statistics of zero, but applying a penalty to x_1 has a LRT statistic of 9.84 which is larger than the critical value. Based on these results, we would conclude that x_1 does not match the other tool marks.

Although the quality model suggests one of the tool marks is an outlier, the summary

statistics and visual comparison of the tool marks do not seem consistent with this result. Examining the parameter estimates of the significant model show that when x_1 is penalized, $\hat{\sigma}^2 = 0.02$. This very small variance estimate, combined with the small sample size of this dataset, make small differences in tool marks easily identifiable.

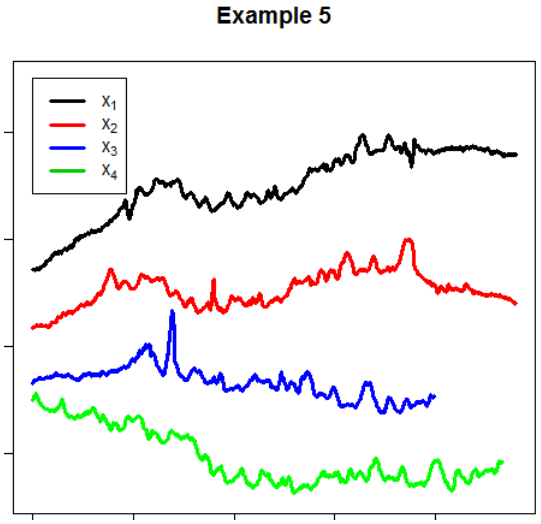


Figure 3.6: Example 5 lab marks.

y_{ij}	x_2	x_3	x_4
x_1	3.095	0.308	2.443
x_2		2.338	2.504
x_3			3.656

Table 3.9: Summary statistics from Example 5.

Penalized x_k	$\hat{\mu}$	$\hat{\gamma}_k$	$\hat{\sigma}^2$	$\hat{\rho}$	$-2\ln(\lambda)$
2	2.39	0.00	1.08	0.00	0.00
4	2.39	0.00	1.08	0.00	0.00
3	2.68	0.58	0.99	0.00	0.50
1	2.83	0.88	0.88	0.00	1.20

Table 3.10: Model estimates from Example 5.

In Example 4, we saw a set of tool marks that were so similar that a false non-match was concluded. Example 5 represents the opposite situation, where the tool marks do not appear to match each other well, but none is identified as an outlier in the analysis. Figure 3.6 shows the four lab tool marks, and Table ?? shows the comparison values. From the figure, we can see that there do not seem to be any strong class or

subclass characteristics in these tool marks. As a result, the tool marks all match each other consistently but none of them match each other strongly. In addition, Table ?? indicates that comparisons involving x_1 are much smaller than the rest. Based on these observations, we might expect that a penalty applied to comparisons involving x_1 might result in a better-fitting model. However, the quality model can not distinguish between the marks, and penalizing them individually does not improve the model significantly. The largest LRT statistic is 1.20, which is much smaller than the critical value.

As with Example 4, the reason for the seemingly false matching conclusion can be found in the parameter estimates. In this case, the estimates of σ^2 are much larger than we saw in Example 4. This indicates that there is, overall, more mark-to-mark variability in this dataset. Because the level of noise is greater, and the sample size is so small, even the mark that differs from the rest in the set is not significantly different.

3.5 Conclusions and Future Work

Hoeksema and Morris (2013) use a likelihood ratio test to successfully distinguish between known matching and known non-matching pairs of tool marks. The method relies on the use of multiple tool marks made in the laboratory by the suspect tool to strengthen the evidence by comparing samples of field-lab comparisons and lab-lab comparisons. However, we observed that the repetition of the mark-making process sometimes results in large mark-to-mark variation. If the degree of mark-to-mark variation is large enough so that the lab marks do not match each other well, this can result in a false non-match result when the lab tool marks are compared to the field tool mark. Thus we have a need to implement a quality control step in the process to assure the lab marks match well before comparing them to a field tool mark.

The model proposed in Section 3.3 imposes a penalty to comparisons involving each tool mark one-at-a-time and compares the resulting log likelihood to that of an unpe-

nalized model. If it is found that the largest log likelihood is significantly larger than the baseline, using a chi-bar distribution with a Bonferroni correction, we can conclude the mark that was penalized in that model is significantly different from the others in the set.

Five examples were given showing the range of results based on the data we had available. The penalized mean quality control step is capable of identifying outlying tool marks in the majority of situations. The exceptions are situations where the tool marks have very little mark-to-mark variation, or a large amount of mark-to-mark variation. In the first case, the small sample size makes it easier to identify small differences, so marks that we would expect to match result in a false non-match due to the small variance estimate. In the second case, again, small sample size makes it harder to identify the outlying tool mark since the variance is larger.

Future work on this model should be based around larger datasets. We saw that the small sample size makes small and large variations in tool marks lead to false non-match or false match results that were unexpected. Also, due to the small sample size, the examples we showed could not be continued recursively. With larger sample sizes, the tool mark elimination process could be repeated, removing one tool mark at a time until the remaining tool marks all match each other.

CHAPTER 4. USING SYNTHETIC TOOL MARKS IN A LIKELIHOOD RATIO TEST FOR FORENSIC COMPARISONS

A paper to be submitted to *Technometrics*

Amy B. Hoeksema [1] [2] and Max D. Morris[3] [4]

Abstract

Over the last few years, several numeric methods have been proposed for comparing a field tool mark found at a crime scene to ones made in the laboratory using a suspect tool, with the goal of determining whether the field tool mark and lab marks "match," i.e. were made with the same tool. For comparisons resulting in a single numerical index, Hoeksema and Morris (2013) proposed the use of a likelihood ratio test to analyze the difference between a sample of comparisons of lab tool marks to a field tool mark, against a sample of comparisons of two lab tool marks. In that analysis, a one-sided hypothesis test was used for which the null hypothesis states that the means of the two samples are the same, and the alternative hypothesis states that they are different and appropriately ordered. The weakness of this approach is that the hypotheses are reversed from the desired analysis; we must assume that the null hypothesis is true until we can prove

[1]Graduate student, Department of Statistics, Iowa State University
[2]Primary researcher and author
[3]Department of Statistics, Iowa State University
[4]Department of Industrial and Manufacturing Systems Engineering, Iowa State University

otherwise, which equates to assuming the tool marks were made by the same tool (i.e. the evidence supports the suspect's guilt) until we can prove otherwise. Using synthetic tool marks generated from a statistical model fitted to the lab tool marks, we propose a method for comparing marks that reverses the hypotheses to achieve the desired test.

Keywords: Outlier test, Profilometry, Striae, Synthetic data

4.1 Introduction

When a crime is committed using a striated tool, such as a screwdriver, forensic scientists sometimes rely on tool marks left at the crime scene as physical evidence. Tool marks made in the laboratory by a suspect tool are compared to the field tool mark to evaluate the likelihood of a match, i.e. that the suspect tool was also used at the crime scene. However, over the last few years the current process of determining tool mark match status through expert visual analysis has come under scrutiny after the National Research Council stated that "With the exception of nuclear DNA analysis... no forensic method has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source" (N.R.C. 2009, p. 7). As a result, several more automated methods based on digitized representations of the tool marks have been proposed over the last few years. For a more thorough background, see Hoeksema and Morris (2013). Despite the numerous proposals, "There are no standard methods for the application of probability and statistics to the analysis of tool mark evidence" (Petraco et al. 2012, p. 901)

Several numerical approaches rely on a single index of similarity to compare two tool marks; two examples are described by Bachrach et al. (2010) and Chumbley et al. (2010). Hoeksema and Morris (2013) proposed a multivariate analysis using as data any normally distributed similarity index in which two samples of comparisons, one from field tool mark with lab tool mark comparisons (field-lab) and one from lab mark with lab mark comparisons (lab-lab), are compared using a likelihood ratio test (LRT). Under this approach, it is assumed that if the tool marks were made by the same tool, both samples should also have the same mean and the LRT is constructed to test this hypothesis.

Although the method described by Hoeksema and Morris (2013) is effective, there is an inherent weakness in the structure of the hypotheses on which it is based. In

a standard hypothesis test of this sort, the null hypothesis states that the means are equal, which is interpreted as evidence that the tool marks were made by the same tool. The alternative hypothesis states that the means are not equal, which is interpreted as evidence that the tool marks were not made by the same tool. Hence the structure of the procedure leads to an assumption of suspect "guilt" unless "innocence" can be positively demonstrated. Thus it is necessary to find a way to reverse the hypotheses so that we can assume the tool marks do not match unless evidence positively indicates that they do.

In this paper, we propose such a method using synthetic tool marks generated from a model fit to the lab tool marks. If these synthetic tool marks do not match the field tool mark as well as the lab tool marks do, this provides evidence that the field tool mark was made by the suspect tool. If the field tool mark was made with a different tool, then the synthetic tool marks will still match the lab tool mark well, but the field tool mark will not match either the synthetic or lab tool marks. We will begin by providing a review of the model used in Hoeksema and Morris (2013) in Section 4.2, followed by an overview of our approach in Section 4.3. We then demonstrate the process in Section 4.4 and discuss conclusions and future research directions in Section 4.5.

4.2 Basic Model - Likelihood Ratio Test

Hoeksema and Morris (2013) described a likelihood ratio test for normally distributed index values generated by comparing pairs of tool marks, and for demonstration used data generated with the Chumbley algorithm (Chumbley et al. 2010). We will continue to use comparison values, y_{ij}, computed using this algorithm and now define some relevant notation. This algorithm can be applied to any striated tool mark, however, we will focus on data and results for screwdrivers. For our purposes, a tool mark is first reduced to a single cross-striae "scan" or depth profile along a path perpendicular to

the parallel striae of the physical tool mark. In the data available to us, these profiles consist of depth values recorded at approximately 9600 "pixel locations" along a linear path. Let x_0 represent the field tool mark, and $x_1, ..., x_n$ represent n tool marks that were made by the suspect tool in the lab. Let y_{ij}, $i < j$, represent the numerical index of similarity that results from comparing x_i to x_j using Chumbley's algorithm. A comparison between the field tool mark and a lab tool mark, y_{0j}, will be referred to as a field-lab comparison, and a comparison between two lab tool marks, y_{ij}, $1 \leq i < j \leq n$, will be referred to as a lab-lab comparison.

In the model described by Hoeksema and Morris, μ_0 is the mean for a field-lab comparison, that is $E(y_{0j}) = \mu_0$ for $j = 1, ..., n$, and μ_1 is the mean for a lab-lab comparison, so $E(y_{ij}) = \mu_1$ for $i, j = 1, ..., n$ and $i < j$. For their analyses, they propose using a LRT to compare the sample of field-lab comparisons to the sample of lab-lab comparisons in a test of

$$\mathrm{H}_0 : \mu_0 = \mu_1 \text{ vs } \mathrm{H}_A : \mu_0 < \mu_1. \tag{4.1}$$

If the means of the two samples are the same, this is evidence that the tool marks were made by the same tool. In the standard framework for comparing a simple hypothesis to a composite hypothesis, the simple hypothesis (null) is assumed to be true unless there is sufficient evidence for the alternative hypothesis. However, in this scenario, this implies we are assuming the tool marks match and trying to provide evidence that they do not match. This is not consistent with standard forensic and legal principles, thus there is a need to find a way to compare the samples of comparisons with the hypotheses reversed.

4.3 Modeling to Generate Synthetic Marks

In the courtroom, a criminal is innocent until proven guilty. Similarly, with the forensic comparison of tool marks, we should assume that the tool marks were made by

different tools until we can demonstrate substantial evidence that they match. That is, using the notation in Section 4.2 we should test

$$\mathrm{H}_0 : \mu_0 < \mu_1 \text{ vs } \mathrm{H}_A : \mu_0 = \mu_1. \tag{4.2}$$

Before describing our proposed test, we first offer a brief description of the process that forensic examiners currently use to compare tool marks visually.

All striated tool marks are thought to be made up of *class characteristics*, *subclass characteristics* and *individual characteristics*. According to the N.R.C. (2009, p.152), class characteristics are "distinctive features that are shared by many items of the same type... such as the width of the head of the screwdriver" and individual characteristics are "the fine microscopic markings and textures that are said to be unique to an individual tool or firearm. Between these two extremes are 'subclass characteristics' that may be common to a small group of firearms and that are produced by the manufacturing process, such as when a worn or dull tool is used to cut barrel rifling." Currently, a forensic examiner must confirm that the tool marks resemble one another closely enough to justify a microscopic comparison at which point, s/he can visually discount similarity due to both class and subclass characteristics. The National Institute of Justice (NIJ) states in their on-line Firearm Examiner Training module that "Examination of the tool allows the examiner to assess the level of subclass characteristics...The examiner compares the class characteristics of the two objects; if all class characteristics correspond, the examiner proceeds to compare the individual characteristics" (National Institute of Justice).

Bachrach et al. (2010) relate the different characteristics to modeling the signature and correlation components of the tool mark. Specifically, they state, "The main purpose of the signature generation component is to isolate those features that are characteristic of the specimen under consideration (individual characteristics) from those that are common to all specimens of the same type (class characteristics). Consider, for example,

the case of a group of screwdrivers of the same make and model. As these screwdrivers are manufactured to the same specifications, the overall geometric shape of the tool marks created by them is very similar. On the other hand, as no two manufactured parts are ever identical, there are microscopic variations specific to each screwdriver blade" (p. 3). They further describe the class characteristics as the "waviness" aspect of the mark (or large scale traits) and the individual characteristics as the "roughness" (or small scale traits).

To identify the match status between two tool marks, it is critical to verify that the individual characteristics of the tool marks match and that they do not just match in class characteristics. Taking this into account, we propose using the lab tool mark to create synthetic tool marks; marks that have been statistically generated as realizations of a model fit to the lab tool mark to match in class characteristics and vary only in individual characteristics. If we can show that the field tool mark matches the lab tool mark better than any of the synthetic tool marks, while the synthetic tool marks and the lab tool mark all match one another equally well, this provides evidence that the field tool mark was made by the same tool as the lab tool mark.

The rationale for our proposed test is not direct, and deserves additional elaboration. As described above, suppose that the crime lab had a suspect tool and that it is used to create a tool mark. But suppose that the lab could also procure additional tools of the same type, and perhaps with similar wear characteristics, and could make additional marks with these as well. Our proposal, then, would be that were the tool used at the crime scene *different* from the suspect tool, the field-lab comparison should not result in a value that is unusual compared to the comparison of the field tool mark to those marks made by tools known not to be the one used in the crime. Indeed, in this case, each tool mark examined would have been made by a different tool, and there should be no *a priori* reason to expect any of the comparisons to be unusual relative to any of the others. But if the suspect tool actually was used in the commission of the crime,

then the field-lab comparison produces the only data value generated from marks that (truly) match, and so this comparison value might be expected to be unusual (an outlier) compared to the others. The null hypothesis ("the suspect is innocent") is that every pair of tool marks is, in a sense, "exchangeable."

This proposal might, in fact, merit some consideration, but would likely be physically difficult to implement in practice. It might be possible to procure tools that were of the same type as the suspect tool, but it is hard to imagine how they might be found to have plausible comparable degrees of wear, so that the "exchangeable" argument could be made. Our proposal is that rather than finding physical tools that are comparable in this sense, that the suspect tool be used as the basis for modeling synthetic tool marks to take their place. In essence, we have to generate tool marks that share the class and subclass characteristics of the suspect tool mark, but differ in individual characteristics while still being "comparable." In 2012, Neumann (2012) developed a methodology that relies on simulated fingerprints to incorporate probability into fingerprint analysis. In this paper, we have proposed to generate synthetic tool marks by simply decomposing the suspect tool mark into two segments, the first of which (the smooth) represents all class and subclass characteristics, and the second of which (the residuals) represents individual characteristics. Further we propose that the residuals can be used to fit a statistical model that can be regarded as a reliable "generator" of comparable tool mark profiles - essentially that the individual characteristics of a tool can be regarded as random draws from this model.

4.3.1 Modeling the Lab Tool Mark

Unlike in Hoeksema and Morris (2013), for these analyses, we are only interested in a single lab tool mark. In modeling the lab tool mark, we wish to capture the class characteristics as fixed and repeatable while treating the individual characteristics as random. This is so that realizations from the fitted model may be regarded as synthetic

tool marks produced by hypothetical tools that are distinct from, but share the class characteristics of, the suspect tool. To achieve this, we fit a Loess smoother to the digitized lab tool mark using the `loess` function in `R`. The procedure requires a single tuning parameter, called the *span*, be specified. The span determines the percentage of data points in the mark that are used in fitting the smoothed value at any one location; the larger the span, the more smooth the fitted function will be. Figure 4.1 illustrates the effects of changing this parameter with the black curve representing the actual tool mark. In this case there are data values corresponding to 9600 pixel locations in each tool mark, so a span value of 0.01 specifies that the data at 96 pixel locations is used to fit the smooth at each point. As a result, the red line which was made using a span of 0.01 almost completely covers the actual tool mark and (visually) hardly smooths at all. The remaining lines, green for a span of 0.05, blue for a span of 0.10 and purple for a span of 0.20, show much higher degrees of smoothing. For the purposes of this analysis, we chose to use a span of 0.05 to capture the class characteristics. The residuals from the (span=0.05) smooth displayed in Figure 4.1 are shown in Figure 4.2.

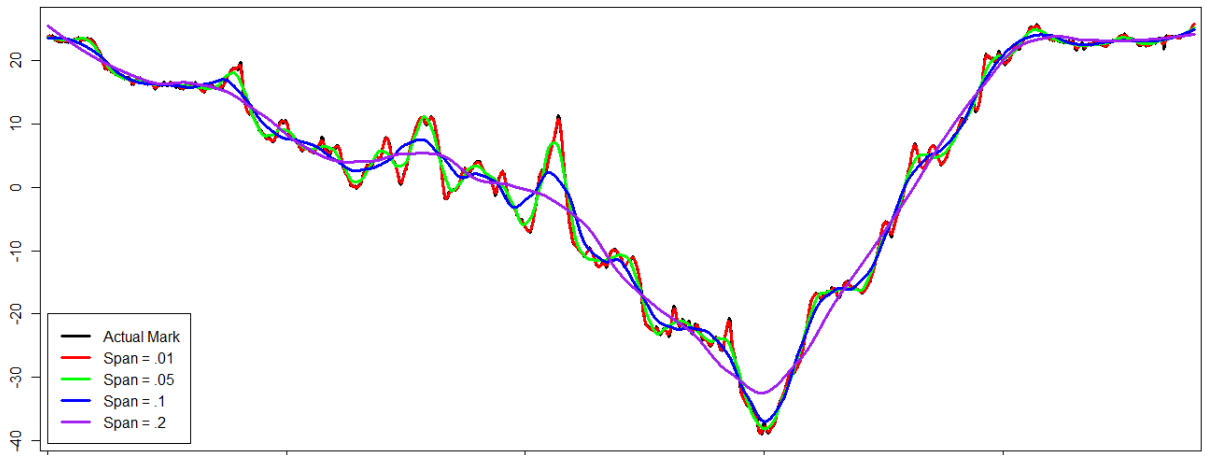


Figure 4.1: A tool mark, shown in black, with four different smoothing curves with smoothing parameters of 0.01 (red), 0.05 (green), 0.10 (blue) and 0.20 (purple).

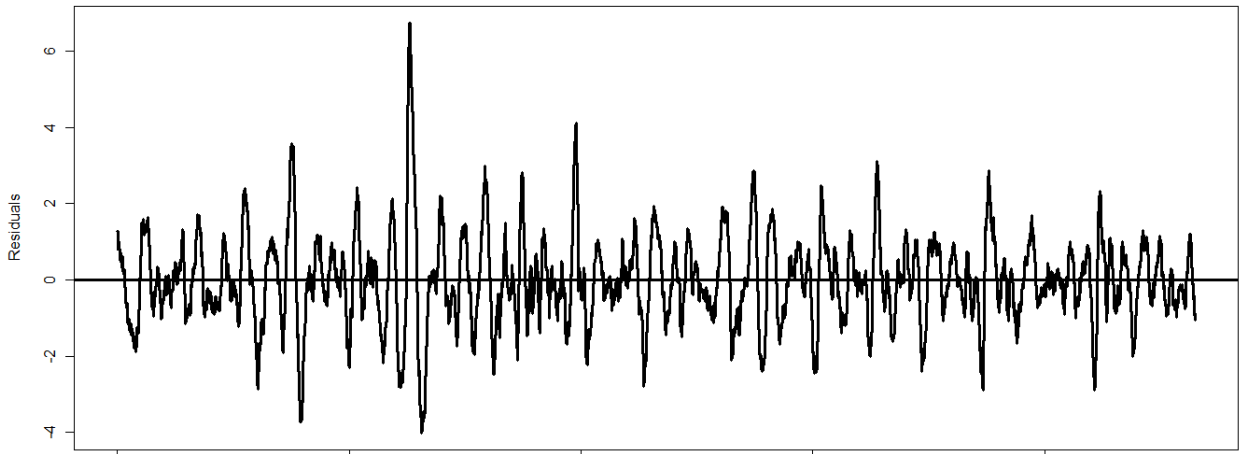


Figure 4.2: Residuals from the mark in Figure 4.1 using a Loess smoother with span = 0.05.

We regard a digitized tool mark x to be comprised of a smooth component, s, estimated by the Loess fit, and a residual component r

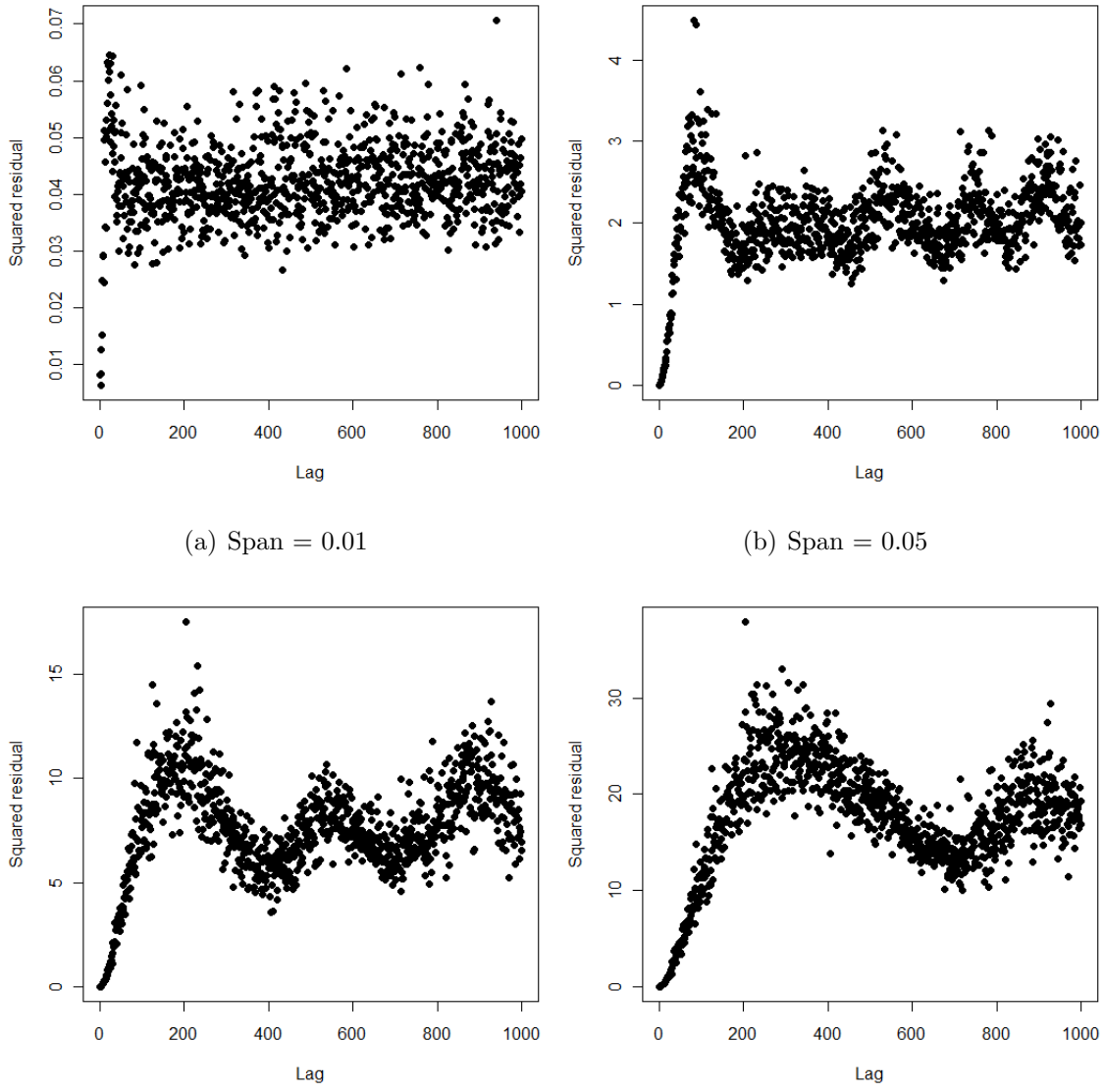
$$x(i) = s(i) + r(i) \tag{4.3}$$

where i is an index representing pixel location along the tool mark. For purposes of modeling, we regard the residuals from the Loess smooth as a realization of a stationary stochastic process. To fit a stochastic process model to the residuals from the smooth of the lab tool mark, we need to determine an appropriate covariance structure. To do this for the tool mark and smooths shown in Figure 4.1, we used variograms which are shown in Figure 4.3. Using the residuals from the Loess smoother, we plot the squared difference between residuals against the lag (absolute difference between pixel indices for those two residuals) to form variograms. Due to the large sample size in the lab tool marks, we randomly chose 100 pairs of residuals for each lag value and have plotted only the average squared difference of residuals for each lag. In Figure 4.3(b), corresponding to a span of 0.05, we can see that the variogram appears approximately linear with positive slope between lag values of 0 and about 50. For lags greater than

50, the variogram is approximately constant with minor oscillating pattern. Thus, we chose to use a covariance function that has the form

$$Cov(r(i), r(i+l)) = \begin{cases} 0 \text{ if } l > L \\ \sigma^2(1 - \frac{l}{L}) \text{ if } l \leq L. \end{cases} \tag{4.4}$$

which represents the covariance for residual r between pixel locations i and $i + l$.



(a) Span = 0.01

(b) Span = 0.05

(c) Span = 0.10

(d) Span = 0.20

Figure 4.3: Variograms showing the mean of 100 randomly chosen squared residuals at each lag point for smoothing parameters of (a) 0.01 , (b) 0.05, (c) 0.10 and (d) 0.20.

With the span parameter and covariance function chosen, we can now fit a stochastic process model to the residual series from the lab tool mark. The parameters we will estimate are $\mu = E(r)$, $\sigma^2 = Var(r)$ and L. Due to the large sample size of these tool marks, the matrix calculations involved in computing the maximum likelihood estimates are unstable. Instead, for a given value of L (which is estimated separately), method of moments (MoM) estimators for μ and σ^2 are used. In this case, since we are fitting the model to the residuals of the mark,

$$\widetilde{\mu} = \bar{r} \tag{4.5}$$

$$\widetilde{\sigma}^2 = \frac{s^2}{1 - \frac{L-1}{n-1}\left(1 - \frac{L+1}{3n}\right)} \tag{4.6}$$

where $\bar{r}$ and s^2 represent the sample mean and sample variance of the residuals, respectively.

The goal of this method is to use one lab tool mark to generate synthetic tool marks that differ from it only in small-scale details typical of "individual characteristics." For a specified value of L and the resulting values of $\widetilde{\mu}$ and $\widetilde{\sigma}^2$, we simulate a large number of synthetic tool marks. The index y is then computed for comparisons of each synthetic mark to the lab mark (lab-synthetic comparisons) and for each pair of synthetic tool marks (synthetic-synthetic comparisons). The difference between these two samples is characterized using the Kolmogorov-Smirnoff (K-S) statistic. We use the L value leading to the minimum K-S statistic as our estimate, $\widetilde{L}$.

For demonstration of this method, we used a grid of values for L, computed the MoM estimates for μ and σ^2, and generated 50 tool marks. For each value of L, Figure 4.4 displays boxplots of the two index samples with the white boxes representing the lab-synthetic comparisons and the grey boxes representing the synthetic-synthetic comparisons. The value of L is shown on the x-axis. Recall the goal is to identify the pair of boxplots that are most alike which is done objectively using a Kolmogorov-Smirnoff test

statistic;this will be the value of L that is chosen for the rest of the analysis. The test statistics and p-values from performing the K-S test on each set of samples are shown in Figure 4.5(a) and 4.5(b), respectively. The value of L is chosen for which the sample of synthetic-synthetic comparison values was most like the sample of lab-synthetic values, that is the value of L that produced samples with the smallest K-S test statistic, or largest p-value. In this particular example, $\widetilde{L}$ is 50.

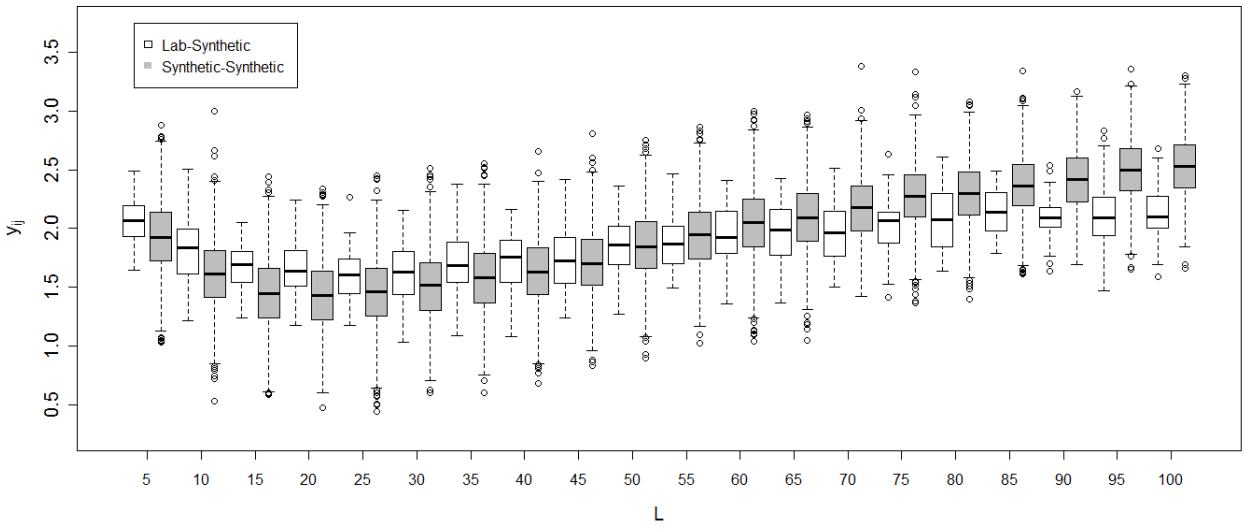
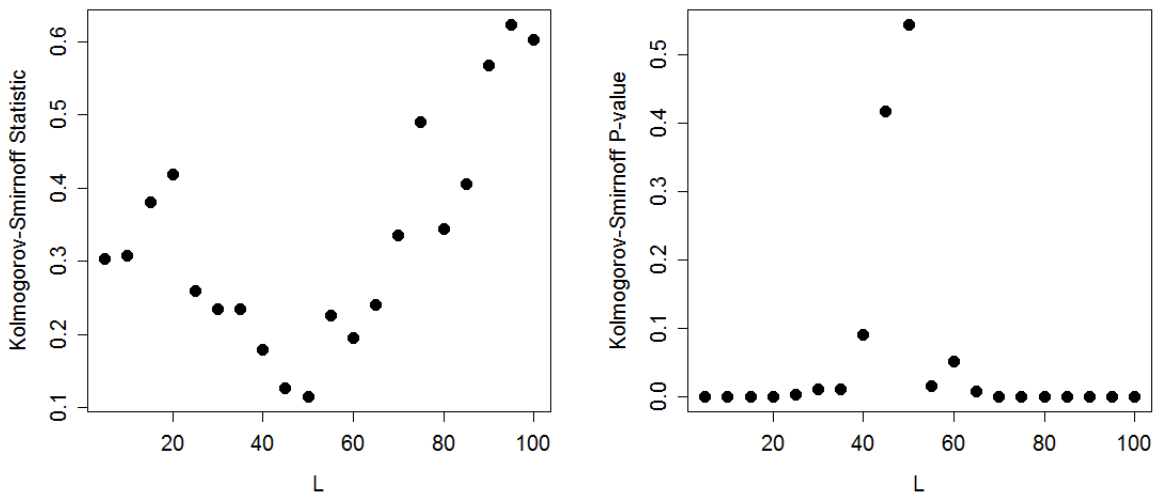


Figure 4.4: Boxplots showing the lab-synthetic comparisons (white) and synthetic-synthetic comparisons (red) for each value of L shown on the x-axis.



(a) Kolmogorov-Smirnoff statistics by value of L. (b) Kolmogorov-Smirnoff p-values by value of L.

Figure 4.5: Results of the Kolmogorov-Smirnoff test showing the test statistics (a) and the p-values (b) for each value of L that was considered.

4.3.2 Creating Synthetic Tool Marks

In order to explain how we simulate data, we must first describe part of the algorithm that is used to compute the comparison values (y's). Following by analogy the process used by expert forensic examiners, Chumbley's algorithm computes a comparison value for two tool marks by determining how similar the marks are numerically. The algorithm first finds the small subsets of each tool marks that match most closely; these are called the *best match windows*. Once the best matching windows have been found, the algorithm calculates the comparison value by evaluating similarity of the areas surrounding the best match windows to determine if those areas match well too. Since tool marks are very large datasets, even non-matching marks can have best match windows that match well, however they are less likely to match along the surrounding areas.

Section 4.3.1 describes our method of modeling the data from a single lab tool mark resulting in a smooth, s and parameter estimates for the residual data, $\widetilde{\mu}$, $\widetilde{\sigma}^2$, and $\widetilde{L}$.

Synthetic tool marks are created by adding a generated residual series (representing individual tool characteristics) to the smooth (representing class characteristics). However, to intelligently add noise, and assure that the synthetic tool marks match the lab tool mark in characteristics that are most important to the matching algorithm used, we force the synthetic marks to be identical to the lab marks within the best match window. Thus, the best match window in the lab mark is first identified when it is compared to the field mark, it is then copied into the analogous segment of each synthetic tool mark, and the remainder of the synthetic residuals are generated conditional on these fixed values.

Because we are assuming a Gaussian model for the residuals, conditional simulation of synthetic residuals is straightforward. We reorder and partition the vector of residuals from the smooth of the lab tool mark, $\mathbf{r}$, as $\mathbf{r} = \binom{\mathbf{r}_1}{\mathbf{r}_2}$, where $\mathbf{r}_1$ is the section of residuals of length n_1 outside the best match window, and $\mathbf{r}_2$ is the section of residuals of length n_2 inside the best match window. Then using the estimate of the covariance function defined in (4.4), we construct variance and covariance matrices:

$$Var(\mathbf{r}_1) = \Sigma_{11} \tag{4.7}$$

$$Var(\mathbf{r}_2) = \Sigma_{22} \tag{4.8}$$

$$Cov(\mathbf{r}_1, \mathbf{r}_2) = \Sigma_{12} \tag{4.9}$$

and simulate individual characteristics conditioned on the best matching window using the conditional multivariate normal distribution with

$$E(\mathbf{r}_1|\mathbf{r}_2) = \widetilde{\mu}\mathbf{1}_{n_2} + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{r}_2 - \widetilde{\mu}\mathbf{1}_{n_1}) \tag{4.10}$$

$$Var(\mathbf{r}_1|\mathbf{r}_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \tag{4.11}$$

The completed synthetic tool mark is then calculated as $\mathbf{s}+\mathbf{r}$, the vectors of smooth and residual components respectively. An example showing the lab tool mark (black) with 10 synthetic tool marks (red) modeled in this way is shown in Figure ??. The vertical black

lines identify the best match window, in which all the tool marks are exactly identical. Outside the best match window, the marks appear to follow the same trends or have the same class characteristics, but differ in individual characteristics or the noise of the tool mark.

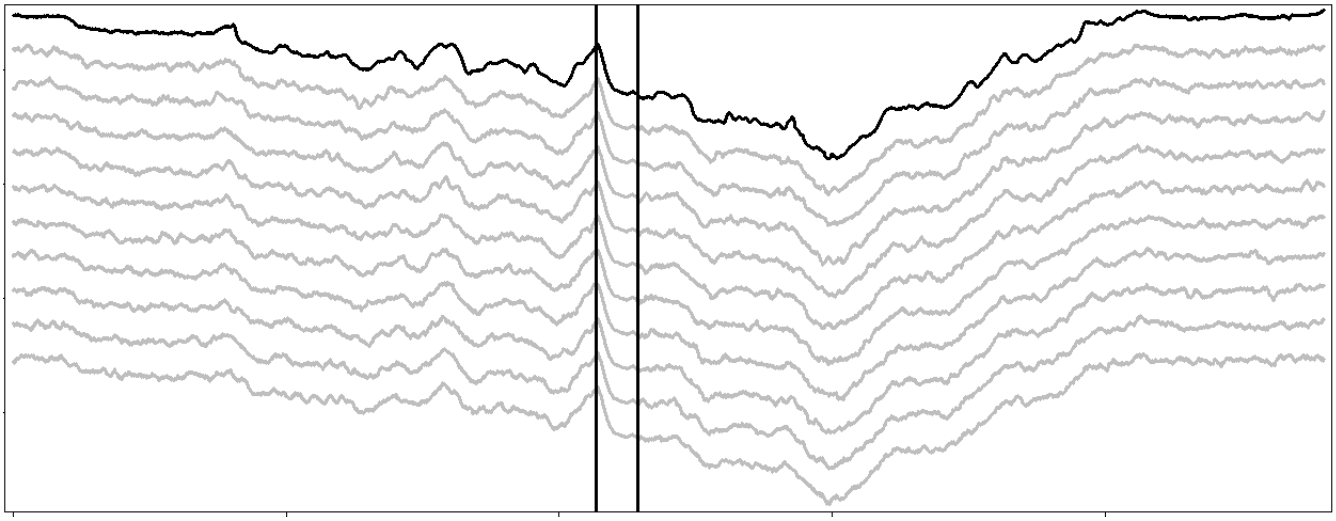


Figure 4.6: A lab tool mark (black) alongside 10 synthetic tool marks (grey) that were modeled off the lab tool mark.

4.3.3 Field Versus Lab Analysis

The last step of the analysis is to incorporate the field tool mark into the testing procedure and compare it to the lab mark as well as the synthetic marks. In the optimization step for L, 50 synthetic tool marks were created and compared to the lab mark and each other, and the two samples were compared using the Kolmogorov-Smirnoff statistic. The parameter $\widetilde{L}$ was selected by minimizing the K-S statistic; and $\widetilde{\mu}$ and $\widetilde{\sigma}^2$ are MoM estimates conditioned on $\widetilde{L}$. We already verified that the synthetic tool marks are indistinguishable from the lab tool mark using a KS statistic; now we use the same 50 synthetic tool marks and compare them to the field tool mark.

If the field tool mark was made by a tool other than that used in the lab, there is no reason to expect a better match between the lab and field marks than between the

synthetic and field marks. However, if the same tool made both lab and field marks, there is reason to believe this comparison will more strongly indicate a match than comparisons of the field mark to the synthetic marks (each of which contains "artificial" individual characteristics). That is, if the physical marks do indeed match, the class and individual characteristics should all match so that the field-lab comparison is larger than any of the comparisons of the field tool mark to the synthetic tool marks. In order to perform a test of the null hypothesis that the field and lab marks were not made by the same tool, we will use an outlier test on all the comparisons involving the field tool mark and determine where in the sample, the field-lab comparison lies. Since, under the null hypothesis, all comparisons involving the field sample come from the same distribution, we can compute a p-value as the number of field-synthetic comparisons that are greater than or equal to the field-lab comparison, and divide this value by $n = 51$, the total number of comparisons involving the field mark.

4.4 Results

To demonstrate the method described in Section 4.3, we now present example analyses based on known matching and known non-matching tool marks. For each available set of data, there are four matching tool marks, all made in the lab under the same conditions, which we will call x_1, x_2, x_3 and x_4. For the known matching data examples, x_1 was chosen to be the lab tool mark and 50 synthetic tool marks were modeled from it and analyses were completed using each for x_2, x_3 and x_4 as the field mark. For the known non-match examples, three tool marks made under the same conditions but by a different tool were chosen as the field tool marks.

Within each example, we observe the samples of lab-synthetic comparisons, synethetic-synthetic comparisons, and three samples of field-synthetic comparisons (one for each of the three field marks) along with the three field-lab comparison values. For the purposes

of these examples, we will denote the samples of comparisons as L-S for lab-synthetic and S-S for synthetic-synthetic. The three samples involving the field tool mark are divided into the sample of field-synthetic comparisons, denoted F-S, and the individual comparison value for the field-lab comparison, denoted F-L. The three samples are distinguished by color, red for x_2, blue for x_3 and green for x_4. For each example there is a figure showing boxplots of the lab-lab comparison, lab-synthetic comparisons, and each set of field-synthetic comparisons and the field-lab comparison denoted by the notations given. There is also a table giving the p-values from the Kolmogorov-Smirnoff test comparing the lab-synthetic comparisons to the synthetic-synthetic comparisons, and the p-values from each of the three outlier tests performed. Finally, to show visually how well the tool marks match one another, there is a figure showing the four digitized tool marks used.

4.4.1 Known Matches

Example 1 contains four matching tool marks, shown in Figure 4.8. The lab mark, x_1, shown in black, was used to create 50 synthetic tool marks. We can see from the boxes in Figure **??** that the synthetic-synthetic comparisons (grey) visually match the synthetic-lab comparisons (white), which is confirmed by the large p-value of 0.8609 from the K-S test shown in Table 4.1. After confirming the lab-synthetic comparisons are indistinguishable from the synthetic-synthetic comparisons, we then performed an outlier test using each of the three other matching tool marks, x_2, x_3 and x_4 (shown in red, green and blue respectively) as the field mark, comparing the field-lab comparison to the field-synthetic comparisons. The single colored dashes in Figure **??** show the field-lab comparisons and the results of the outlier tests are shown in Table 4.1. All three outlier tests have a p-value of 0.020, so we reject the null hypothesis that the data value in question (the field-lab comparison) is from the same distribution as the rest of the data (field-synthetic comparisons) and we would conclude that the lab tool mark

and all three field tool marks were made by the same tool.

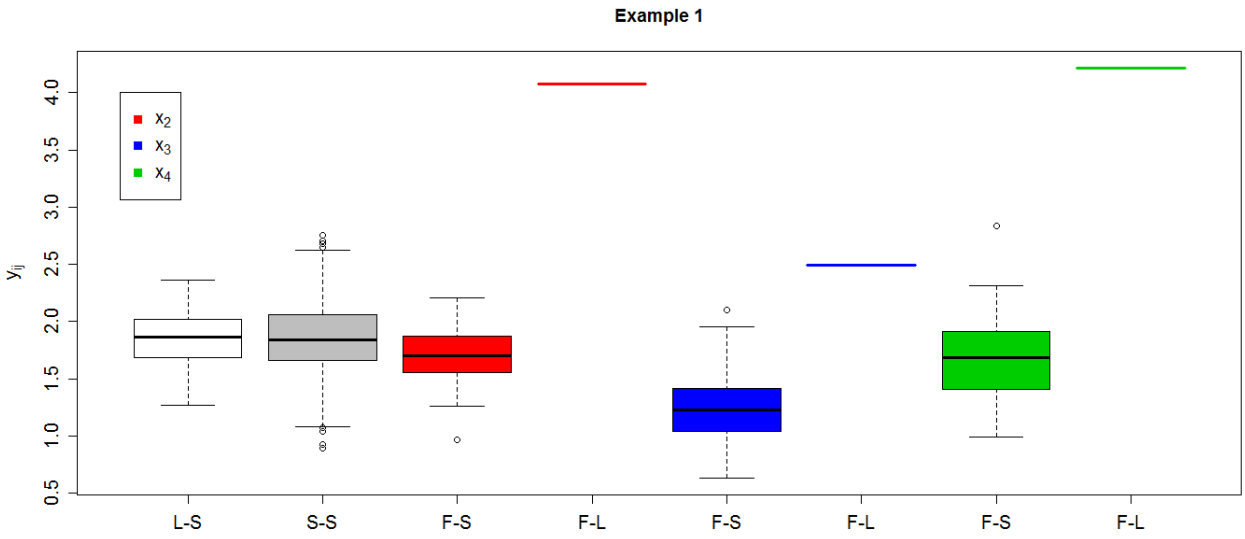


Figure 4.7: Example 1 y_{ij}s showing lab-synthetic comparisons (white), synthetic-synthetic comparisons (grey) and three sets of field-synthetic comparisons using x_2 (red), x_3 (blue), and x_4 (green). The single lines are the field-lab comparisons for the three field marks.

Example 1 Test Results

Test	p-value
K-S test: Lab Mark, x_1	0.543
Outlier test: Field 1, x_2	0.020
Outlier test: Field 2, x_3	0.020
Outlier test: Field 3, x_4	0.020

Table 4.1: Example 1 p-values.

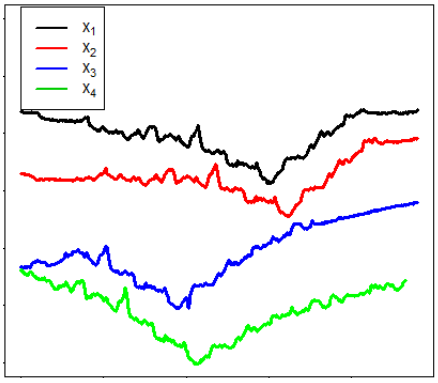


Figure 4.8: Example 1 tool marks.

Example 2 also contains four tool marks that were made by the same tool under the same conditions, however, we can see from Figure 4.10 that one of the marks, x_2 shown in red, does not resemble the other three tool marks. As before, x_1 shown in black was used to create 50 synthetic tool marks which were compared to one another and are shown as boxplots in Figure ?? in white (lab-synthetic comparisons) and grey (synthetic-synthetic

comparisons). Table 4.2 confirms that these comparisons are indistinguishable since the p-value is 0.1773. Each of the three remaining tool marks was compared to the lab and synthetic tool marks. We can see from the results of the outlier test in Table 4.2 that x_3 (blue) and x_4 (green) both have small p-values of 0.020 so we would conclude these two marks were made by the same tool as the lab tool mark. However, x_2 shown in red, fails to reject the null hypothesis in the outlier test since the p-value is 0.392 and we conclude that it was not made by the same tool as the lab tool mark. Although we know this to not be true, it is not a surprising result since we can visually confirm in Figure 4.10 that x_2 does not resemble any of the other three tool marks in this set.

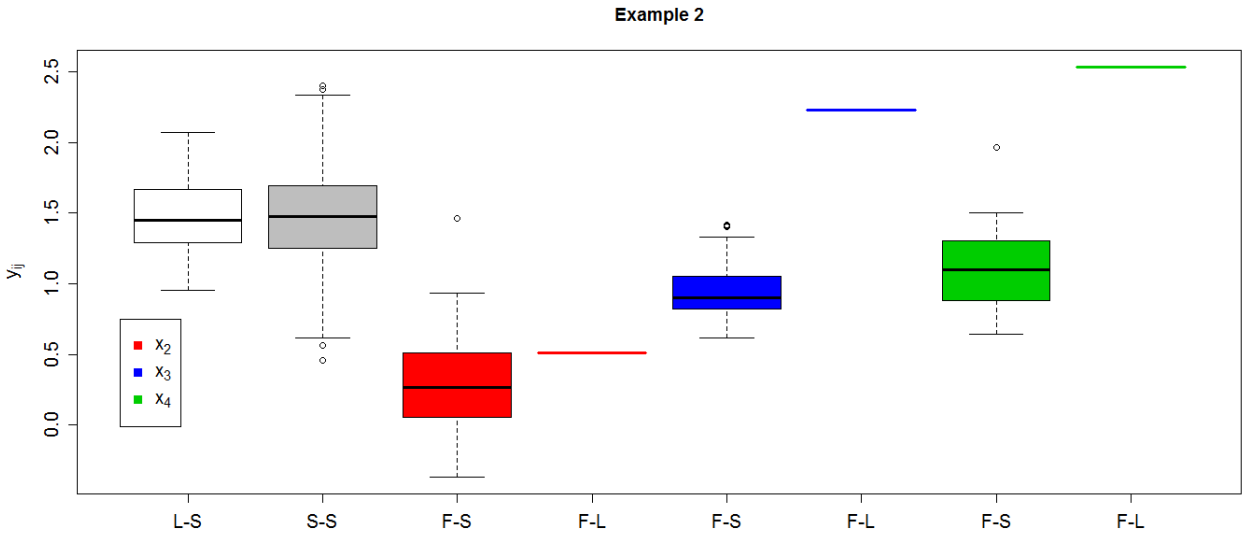


Figure 4.9: Example 2 y_{ij}s showing lab-synthetic comparisons (white), synthetic-synthetic comparisons (grey) and three sets of field-synthetic comparisons using x_2 (red), x_3 (blue), and x_4 (green). The single lines are the field-lab comparisons for the three field marks.

Example 2 Test Results

Test	p-value
K-S test: Lab Mark, x_1	0.937
Outlier test: Field 1, x_2	0.275
Outlier test: Field 2, x_3	0.020
Outlier test: Field 3, x_4	0.020

Table 4.2: Example 2 p-values.

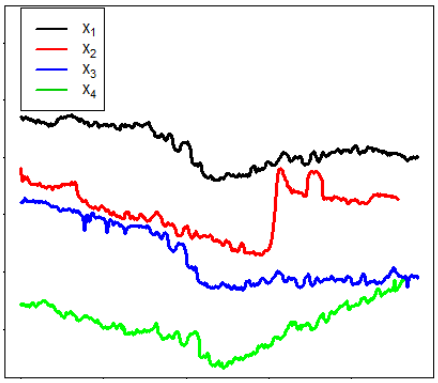


Figure 4.10: Example 2 tool marks.

4.4.2 Known Non-Matches

In the first non-match example, Example 3, we are using a lab tool mark, x_1, that is known to be made by a different tool than the three field marks, x_2, x_3 and x_4. Figure 4.12 shows the four tool marks, and we can see that the black tool mark shows none of the individual characteristics of the other three tool marks. Table 4.3 shows that the synthetic tool marks are indistinguishable from the lab tool mark, since the K-S test has a p-value of 0.6134, and we can visually see this from the lab-synthetic comparison (white) and synthetic-synthetic comparison (grey) boxes in Figure ??. However, none of the three field marks reject the null hypothesis in the outlier test, as demonstrated in the boxplots. Not only are none of the field-lab comparisons outliers, but the boxes also indicate the field and synthetic tool marks do not match since the lab-synthetic comparisons are all much larger than the field-synthetic comparisons. Since the field and lab tool marks do not match, it is not surprising that the boxes showing the field-synthetic comparisons resemble those of non-matching comparisons since the synthetic tool marks are generated from a model of the non-matching lab tool mark.

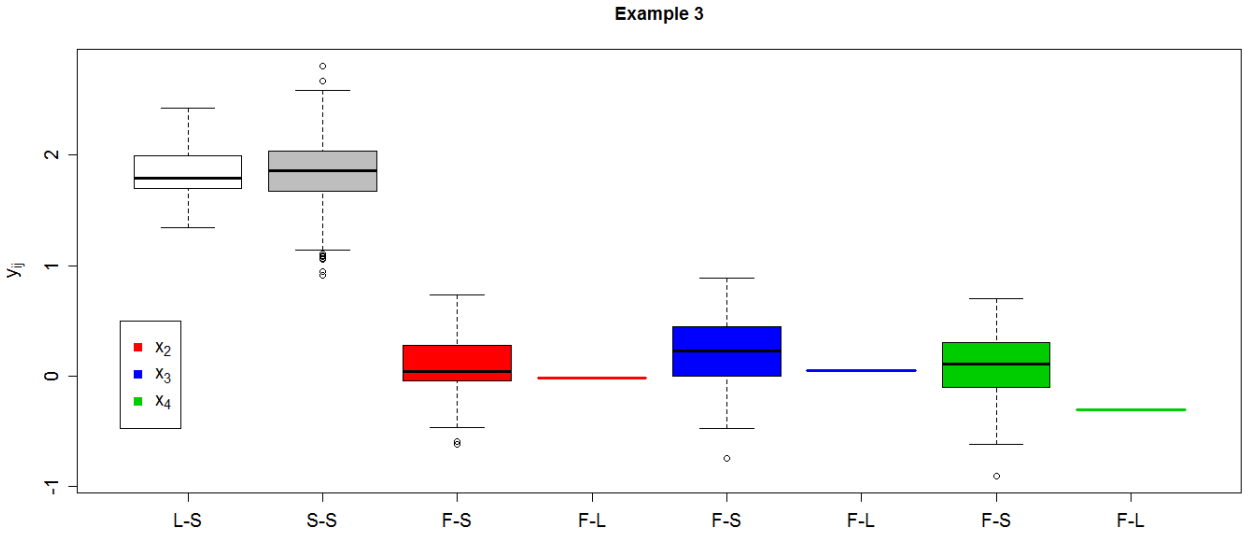


Figure 4.11: Example 3 y_{ij}s showing lab-synthetic comparisons (white), synthetic-synthetic comparisons (grey) and three sets of field-synthetic comparisons using x_2 (red), x_3 (blue), and x_4 (green). The single lines are the field-lab comparisons for the three field marks.

Example 3 Test Results

Test	p-value
K-S test: Lab Mark, x_1	0.717
Outlier test: Field 1, x_2	0.686
Outlier test: Field 2, x_3	0.726
Outlier test: Field 3, x_4	0.922

Table 4.3: Example 3 p-values.

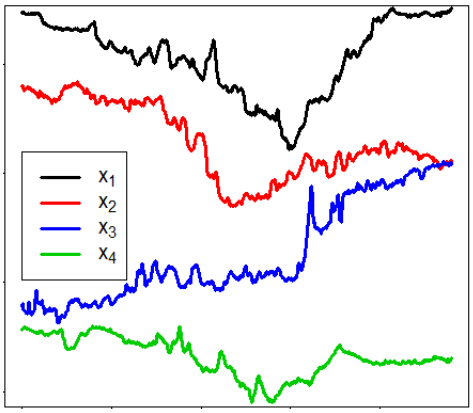


Figure 4.12: Example 3 tool marks

Example 4 is another set of four tool marks such that x_1 (black) is a tool mark made by a different tool than x_2 (red), x_3 (green), and x_4 (blue); marks are shown in Figure 4.14. The first two boxplots in Figure ?? show the lab-synthetic (white) and synthetic-synthetic (grey) comparisons from making 50 tool marks genereated from a model of the lab tool mark. Table 4.4 shows that once again, the lab tool marks are indistinguishable from the synthetic tool marks since the K-S test p-value is 0.9762. When the three

field tool marks are compared to the synthetic tool marks, we once again note that the red, green and blue field-synthetic comparisons boxplots are similar to those from non-matching comparisons. Similarly, the outlier test returns large p-values for all three field tool marks which indicates the lab and field tool marks were not made by the same tools.

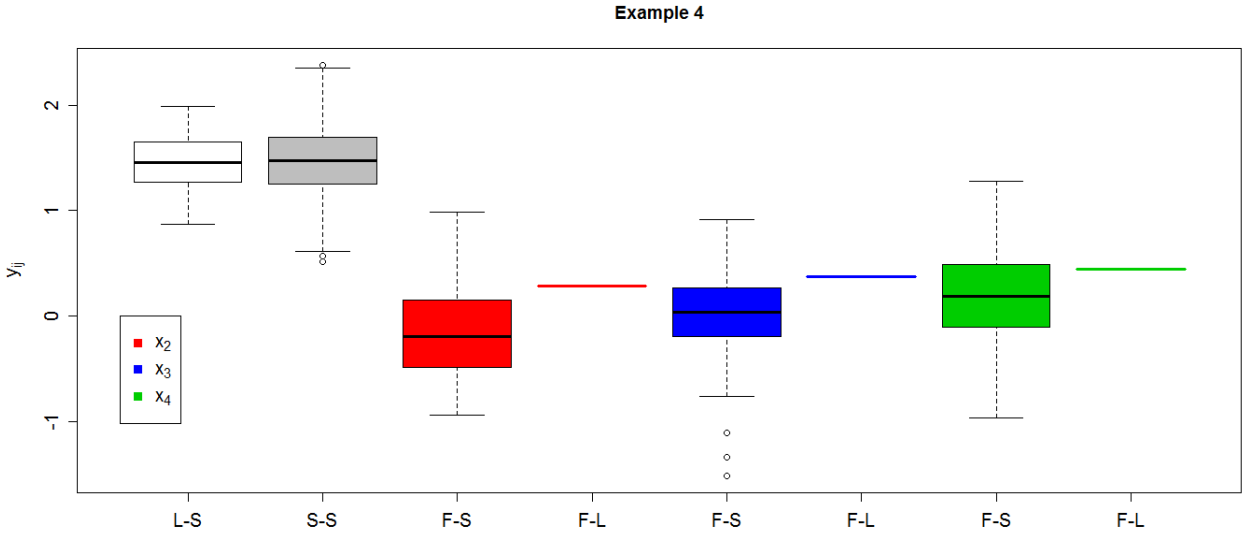


Figure 4.13: Example 4 y_{ij}s showing lab-synthetic comparisons (white), synthetic-synthetic comparisons (grey) and three sets of field-synthetic comparisons using x_2 (red), x_3 (blue), and x_4 (green). The single lines are the field-lab comparisons for the three field marks.

Example 4 Test Results

Test	p-value
K-S test: Lab Mark, x_1	0.552
Outlier test: Field 1, x_2	0.196
Outlier test: Field 2, x_3	0.216
Outlier test: Field 3, x_4	0.294

Table 4.4: Example 4 p-values.

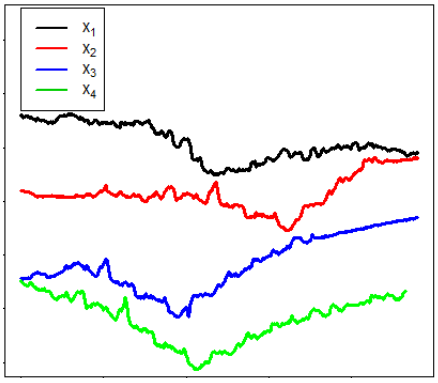


Figure 4.14: Example 4 tool marks

4.5 Conclusions and Future Work

Although the Likelihood Ratio Test procedure proposed in Hoeksema and Morris (2013) consistently fails to reject the null hypothesis of equal means when known matching tool marks are compared, the error of falsely rejecting a claim that the field and lab tool marks were not made by the same tool (the alternative hypothesis in this case) cannot be controlled. Further methods are needed that can prove tool marks match rather than proving they are not non-matching. In the methods described in Section 4.3, we propose modeling synthetic tool marks that match the lab tool mark in class and subclass characteristics and differ only in individual characteristics, similar to random noise. When we compare the lab-synthetic comparisons to the synthetic-synthetic comparisons, we confirm using a Kolmogorov-Smirnoff test that the lab tool mark is indistinguishable from the synthetic tool marks. When the field tool mark is compared to the lab and synthetic tool marks, if the lab and field tool marks were made by the same tool, the field-lab comparison is distinguishable from the field-synthetic comparisons as an outlier. When the tool marks were not made by the same tools, the field-lab comparison is not distinguishable.

The method as it has been presented, succeeds at identifying known matches and known non-matches in example calculations. Our method might be improved through better parameter estimation. For the purposes of this paper, the smoothing parameter was chosen, then the lag parameter, L, was estimated using a coarse grid search from 0 to 100 in increments of 10. For future work, both parameters could be estimated together. In addition, we used method of moments estimators for μ and σ^2 primarily due to the numerical challenge of maximum likelihood estimation. Future work could explore whether there are better ways to estimate these parameters.

We fully admit that, before practical application of this approach could be considered, substantial further justification would be needed. In particular, while it is generally

understood that the "length characteristics" of class and individual characteristics are relatively longer and shorter, respectively as was suggested by Bachrach et al. (2010), it is hard to defend the premise that selecting a choice of smoother span by simple graphical examination (as we have done here) reasonably corresponds to separation of class and individual marks. And even if so, the claim that there is enough information in one set of residuals to build a statistical model representing the population of class characteristics is clearly also debatable. While the follow-up work that would be required to fully develop a methodology of the type described here for practical application is substantial (and certainly beyond what our limited resources would support), we believe that our approach has substantial merit, and that such follow-up research would be valuable.

CHAPTER 5. SUMMARY AND CONCLUSIONS

Chapters 2, 3 and 4 present the proposals for improving the process of tool mark comparison statistically. We began in Chapter 2 by proposing a new approach involving the use of multiple lab tool marks and analyzing samples of comparisons all at once, rather than analyzing them individually. Doing so, we can apply a likelihood ratio test (LRT) to the two samples to determine the match status of the tool marks. With this method in place, we then enhanced the model by adding a component to account for the angle at which tool marks are made. We determined the angle at which a tool is held affects the appearance of the resulting tool mark, so the basic model is amended by allowing for the different angles within a comparison while also predicting the angle at which the field tool mark is made. While our resources only allowed for the use of marks made at five different angles, we found that the prediction of the field angle is accurate, but for further work, tool marks should be made at five degree increments.

In Chapter 3, we address the effect that the quality of lab tool marks has on the likelihood ratio test approach that was proposed in Chapter 2. Mark-to-mark variation has a significant effect because if the lab tool marks do not match each other well, the LRT could falsely conclude a non-match. We propose adding a quality control step to the process during which we compare the lab tool marks to one another, before comparing them to the field tool mark, and determine if one of the lab tool marks is significantly different from the others. If a lab tool mark is determined to be significantly different from the others, it can be removed before comparison to the field tool marks to remove the effects of poor tool mark quality. When the method was applied to the the tool

marks available to us, we found that the majority of lab tool marks identified a poorly matching tool mark when appropriate, and returned no significant differences when appropriate. However, in a few situations, the small sample size became a factor when the four tool marks compared either matched too well and had a very small estimated variance, or did not match well but had a large estimated variance. As a result, the method unexpectedly identified an outlying tool mark in the first case, and did not identify an outlying mark in the second case. With a larger number of lab tool marks to compare, the variance estimates will have less of an affect and the method could be applied iteratively, removing one tool mark at a time until the remaining marks all match each other well.

In Chapter 4, we address a concern from Chapter 2, in which the hypotheses from the test used in the likelihood ratio test need to be reversed. When you assume that the null hypothesis is true in Chapter 2, that is that the means of the sample of lab-lab comparisons and field-lab comparisons are the same, we are assuming the tool marks match before analysis is done. To reverse the hypotheses, we propose modeling synthetic tool marks from the lab tool marks. This is to simulate the affect of having many tools of the same make and model (same class characteristics) but with different wear and other individual characteristics. Once we verify the lab tool mark and synthetic tool marks match well using a Kolmogorov-Smirnoff statistic, we then determine a match if the comparison between the field tool mark and the lab tool mark is an outlier compared to the sample of comparisons of the field tool mark to each of the synthetic tool marks. Using the available tool marks, we show that when the lab and field tool marks match each other well, the field-lab comparison is an outlier. However, there are also examples of known matches that do not match well visually, and thus the field-lab comparison is not an outlier.

Each of these chapters provides examples that show, in most cases, the proposed methodologies are able to enhance the basic model and method from Chapter 2. How-

ever, the lack of available resources is apparent in each. For further research, more tool marks from more angles, specifically at 5 degree increments, is necessary. In the angle analysis, this would allow us to estimate all of the parameters in the model, in particular θ. We could also estimate the field angle more precisely and further check the accuracy of the model. The quality control analysis proposed in Chapter 3 could be improved with more data since it would remove the affect of the variance of the tool marks, and allow the model to be applied iteratively. Finally, in the proposed model using synthetic tool marks, we would like to further explore the methods of parameter estimation.

BIBLIOGRAPHY

Bachrach, B., Jain, A., Jung, S., and Koons, R.D. (2010), "A statistical validation of the individuality and repeatability of striated tool marks: Screwdrivers and tongue and groove pliers," *Journal of Forensic Sciences*, 55(2), 348–357.

Biasotti, A.A. (1959), "A Statistical Study of the Individual Characteristics of Fired Bullets," *Journal of Forensic Sciences*, 4(1), 34–50.

Biasotti, A.A, and Murdock, J.E. (1997), "Firearms and toolmark identification: Legal issues and scientific status," In *Modern Scientific Evidence: The Law and Science of Expert Testimony*, ed D.L. Faigman, D.H. Kay, M.J. Saks, and J. Sanders, 124 – 151, St Paul: West Publishing Co.

Burd, D.Q. and Kirk, P.L. (1942), "Tool Marks – Factors Involved in Their Comparison and Use As Evidence," *Journal of Criminal Law and Criminology*, 32(6), 679– 686.

Chernoff, Herman. (1954), "On the Distribution of the Likelihood Ratio," *Annals of Math Stat.*, 25, pp 573–578.

Chumbley, L. S., Morris, M. D., Kreiser, M. J., Fisher, C., Craft, J., and Genalo, L. J., Davis, S., Faden, D. and Kidd, J. (2010), "Validation of tool mark comparisons obtained using a quantitative, comparative, statistical algorithm," *Journal of Forensic Sciences*, 55(4), 953–961.

Hoeksema, A.B. and Morris, M.D. (2013), "Significance of Angle in the Statistical Comparison of Forensic Tool Marks," submitted to *Technometrics*.

Daubert v. Merrell Dow Pharmaceuticals, Inc. (1993), 509 U.S., 579–589.

Ekstrand, L., Zhang, S., Grieve, T., Chumbley, L. S. and Kreiser, J. (2013), Virtual tool mark generation for efficient striation analysis, *Journal of Forensic Sciences*, (accepted w/ minor revision)

Gambino, C., McLaughlin, P., Kuo, L., Kammerman, F., Shenkin, P., Diaczuk, P., Petraco, N., Hamby, J., and Petraco, N.D.K. (2011), "Forensic Surface Metrology: Tool Mark Evidence," *Scanning*, 33, 272–278.

Kowalski, J. and Tu, X. (2007), "References, in Modern Applied U-Statistics," Hoboken: John Wiley & Sons, Inc.

Neumann, Cedric. (2012), "Fingerprints at the crime-scene: Statistically certain, or probable?" *Signficance*, 9 (1), 21–25.

National Institute of Justice, "Firearm Examiner Training Module 13: Toolmark Idenfitication," [online]. Available at *http://www.nij.gov/training/firearms-training/module13/fir_m13.htm.*

National Research Council, Committee on Identifying the Needs of the Forensic Sciences Community, (2009). "Strengthening forensic science in the United States: A path forward," The National Academies Press.

Neel, M. and Wells, M. (2007). "A comprehensive statistical analysis of striated tool mark examinations part 1: Comparing known matches and known non-matches," *AFTE Journal*, 39(3), 176–198.

Nichols, Ronald G. (1997). "Firearm and Toolmark Identification Criteria: A Review of the Literature," *Journal of Forensic Science*, 42(3), 466–474.

Nichols, Ronald G. (2003). "Firearm and Toolmark Identification Criteria: A Review of the Literature, Part II," *Journal of Forensic Science*, 48(2), 318–327.

Nichols, Ronald G. (2007). "Defending the Scientific Foundations of the Firearms and Tool Mark Idenfitication Discipline: Responding to Recent Challenges," *Journal of Forensic Sciences*, 52(3), 586–594.

North Caroline State Crime Lab (2012). "Technical Procedure for Tool Mark Examination: Firearm and Tool Mark Examination," textithttp://www.ncids.com/forensic/sbi/Firearms/Technical/Tool-Mark-Examination.pdf.

Petraco, N.D.K., Shenkin, P., Speir, J., Diaczuk, P., Pizzola, P., Gambino, C., Petraco, N. (2012). "Addressing the National Academy of Sciences Challenge: A Method for Statistical Pattern Comparison of Striated Tool Marks," *Journal of Forensic Sciences*, 57(4), 900–911.

Tamasflex. "Comparison Microscope". Digital image. Wikipedia. N.p., 14 Nov. 2012. Web. 15 Nov. 2012. [online] *http://en.wikipedia.org/wiki/Comparison_microscope.*

Wevers, G., Neel, M. and Buckleton, J. (2011). "A comprehensive statistical analysis of striated tool mark examinations part 2: Comparing known matches and known non-matches using likelihood ratios," *AFTE Journal*, 43(2), 137–145.

Wood, Simon N. (2012). "MGCV." Available online at *http://cran.r-project.org/web/packages/mgcv/mgcv.pdf.*