Statistical Methods for Comparison of Forensic Glass Samples in Australia.

Oliver Lountain

November 10, 2021

Thesis submitted for the degree of
Master of Philosophy
in
Statistics
at The University of Adelaide
Faculty of Engineering, Computer and Mathematical Sciences
School of Mathematical Sciences

THE UNIVERSITY
of ADELAIDE

To the Level 5 Math Gang,
who made it easy to turn up every
day, and taught me to never
underestimate the importance of the
people who sit next to you.

Contents

Abstract

Glass is often broken when a crime is committed, whether it be a case of breaking and entering or a hit and run vehicle incident, for example. Forensic scientists may be tasked with analysing the broken glass in a number of ways. They may be asked to establish how the glass was broken, for example the type of instrument used to break the glass and whether it was broken from the inside or the outside. They may also be asked to connect a suspect with having been at the scene of the crime. In this thesis we restrict our focus to statistical methods to make comparison between two fragments of broken glass: one from the crime scene and another found on the clothing of a suspect. The chemical composition of the glass is measured by a technique known as laser ablation-inductively couple plasma mass spectrometry (LA-ICPMS).

We show that machine learning methods, decision trees in particular, provide near-perfect prediction accuracy, improving on the currently employed methods. Further, the strength of evidence can be quantified by extending these methods and by constructing score-based likelihood ratios – a benefit otherwise only given by the traditional likelihood ratio methods. We find that these traditional likelihood ratio-based procedures do not offer an improvement in terms of prediction accuracy, and in fact perform worse than the current methodologies in this regard.

These results demonstrate that a great deal of prediction accuracy can be gained by taking full advantage of the multivariate structure of the LA-ICPMS data. While glass evidence only constitutes a single component of a legal case, it is important that the methods used to evaluate the data are high in accuracy. In particular, in correspondence with the philosophy of "innocent until proven guilty", our models perform well in minimising the rate at which samples are incorrectly classified as matching.

Signed Statement

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give permission for the digital version of my thesis to be made available on the web, via the Universitys digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Signed: .. .. Date: 10/11/21

Acknowledgements

The completion of this thesis would not have been possible were it not for the help and support that I have received from a number of people.

Firstly, I am very grateful to the Australia New Zealand Policing Advisory Agency National Institute of Forensic Science for providing the funding for this project to go ahead.

I would also like to thank the team at Forensic Science SA: Hayley Brown, Kahlee Redman and Sharon Wilczek for providing the data, and helping me to understand the forensic science side of the project.

Further, I would like to express my appreciation to my supervisors: Melissa Humphries, Jono Tuke, and Andrew Metcalfe. Your guidance, support, and reassurance of my abilities were key to the completion of this thesis.

Finally, I wish to give my sincere thanks to my family and friends, for all of the emotional support that you have provided during the more difficult moments of my studies. In particular, Sofija, I can't thank you enough for talking to me through the various crises and panicked phone calls, and for just listening to me complain.

Part I

Introduction and Background

Chapter 1

Introduction

Those committing a crime may not be aware that after breaking glass, they may be carrying crucial evidence on their clothes that can place them at the scene of the crime. This is the nature of forensic glass evidence. One of the key tasks given to forensic practitioners is identifying whether glass samples found on the clothing of a suspect match the glass object that was broken when the crime was committed. In this thesis we explore the use of a number of statistical methodologies to make this comparison, and compare the results of new techniques with the current best practice approach.

Our work is conducted in collaboration with Forensic Science South Australia (FSSA), who have provided data from South Australian casework from 2017-2020 which has been measured by measured by a technique called laser ablation-inductively couple plasma mass spectrometry (LA-ICPMS).

1.1 Forensic Glass Evidence

1.1.1 Transfer of Glass

When a person breaks a glass object – such as a window – minuscule fragments of the broken glass are transferred to the person's clothing. This process of fragmentation of the broken glass back towards the object which broke it was observed by Nelson and Revell (1967) and is now referred to as backscatter fragmentation. Any person within "a few feet of the window" which is broken will likely have a number of minute fragments of glass transferred to their clothing. When a suspect is apprehended, these fragments can be found on their clothing and compared to the glass which was broken, potentially placing the suspect close to the object at the crime scene

when it was broken. This analysis requires a careful treatment as one of course wishes to avoid wrongfully convicting an innocent person. As such, the development of measurement techniques with a high level of precision to discriminate between, say, two panes of glass manufactured at the same factory, on the same day, but installed as windows in two distinct buildings is of the utmost importance. Likewise, we require a statistical technique which allows for a clear quantification of the level of difference between two glass samples, and a threshold of permissible difference to classify samples as coming from the same piece of glass. In this section we provide an overview of the methodology used to measure glass fragments.

1.1.2 Refractive Index Analysis

Refraction is the phenomenon of light bending, or change of direction of light as it moves from one medium to another. As light passes from a vacuum to a transparent medium such as glass, it is slowed, resulting in this change of direction. This idea is quantified by refractive index (RI) – a property had by every medium through which light can travel. It gives a measure of how fast light travels through a given material, defined by the ratio of the speed of light in a vacuum, to the speed of light in the medium, that is,

$$n = \frac{c}{v},$$

where n is the refractive index of the medium, c is the speed of light in a vacuum, and v is the speed of light in the medium. Being a property of individual pieces of glass, refractive index has historically been used to discriminate between glass samples. Now, it is often used as a first step screening process before using LA-ICPMS.

Curran *et al.* (2000) note that in the last half a century, the process of glass manufacturing has been greatly refined – particularly in the area of quality control. New methods of automated production have allowed for a great deal of control over the physical characteristics of the glass, namely the thickness of the glass, and its refractive index. Further to this, the production methodology has become increasingly similar between manufacturers, leading to an observed decrease in the variation of RI between panes of glass produced in a given factory. They note also, however, that the increased globalisation of the market – that glass is often imported from a number of different companies – has offset this issue somewhat, but that using RI alone may not be sufficient to discriminate between samples in forensic casework.

1.1.3 Elemental Analysis

Analysis of the elemental composition of glass offers some benefits over the comparison of refractive index. It is possible for two entirely distinct fragments of glass, that is, fragments originating from different sources, to have the same RI, but have observable differences in concentrations of several elements. The primary benefit of elemental composition measurements is simply the fact that the data is multivariate. The use of several measurements to compare glass samples allows for a robust comparison, both by comparing multiple elements individually, and by taking the correlations between the variables into account.

A number of different techniques have been employed for the measurement and analysis of elemental composition of glass, largely because the equipment required can be very expensive, and forensic laboratories have often been put in the position of having to adapt the equipment they have available to suit a number of different purposes (Curran *et al.*, 2000). In Australia, LA-ICPMS is conducted in some forensic laboratories and so data collected in this way will be the focus of this thesis.

1.1.4 Presenting Evidence in Court

The analysis of forensic glass evidence, along with other types of evidence such as DNA, fingerprints and handwriting, is presented in court to support either the prosecution or defence. While glass evidence only constitutes a single component of a case, it is important that the methods used to evaluate the data are high in accuracy. In particular, in correspondence with the philosophy of "innocent until proven guilty", we aim to minimise the false positive prediction rate. In other words, we aim to minimise the rate at which a statistical model suggests that the defendant was at the scene of the crime, when in fact they were not.

1.2 Analysis of LA-ICPMS Data

The current best practice techniques employed for forensic glass analysis in South Australia involve making separate univariate comparisons between the measurements of each individual element in the control and recovered glass samples. This will be explored in detail in Chapter 3. Research has been conducted into the use of techniques which take full advantage of the multivariate nature of the data through its individual elements as well as

covariance structure. We aim to apply these methods, as well as some new methods which we introduce, to the data and compare their performance against the current practice methodology.

1.2.1 Multivariate Data

The motivation behind the research presented in this thesis stems from the question of whether the currently employed techniques can be improved upon, particularly with regards to maximising overall prediction accuracy, and minimising false positive predictions. Specifically, the research is motivated by improving the methodology to take full advantage of the multivariate LA-ICPMS data, and investigating the importance of establishing a location-specific background database of glass samples. Further to this, we aim to establish how the methods perform on different data sets. We consider two data sets: a diverse set of observations originating from South Australian casework from 2017 to 2020, and a much more homogeneous set of observations taken from two glass factories in the USA. These two data sets will be described in detail in Section 2.1.

1.2.2 Location-Specific Glass Profile

The variety of glass found in a given location, whether that be a city, region or country, may be quite specific, depending on how the glass used in that location is manufactured. As such, one can establish a glass profile for a given location, which captures the distribution of glass measurements found at that location. Such a profile can be determined by collating a database of glass measurements, which can be used by forensic examiners in their statistical analysis. Firstly, such a database can be used to quantify the level of spread observed in a location. For example, if all of the window glass in a certain city originates from the same factory, it may be the case that the elemental composition measurements for windows show little variation. In this case, a tighter constraint should be placed on what is deemed "similar enough" to be classified as matching, as compared to a city in which the windows originate from many different factories, and show greatly varying measurements.

Such a profile can also be utilised to help inform the strength of evidence for or against a match. They can be used in this way when constructing likelihood ratios, which will be discussed in detail in Part II. For this purpose, the distribution of measurements from the background database is used to inform how likely a random piece of glass is to have a certain measurement. To give an example, again suppose that a large proportion of window glass in

a given city were to originate from a single factory, and that there was little variation in the elemental composition of glass from this factory. In this case, the elemental measurements of this glass would be very common, meaning that it is likely that a random piece of glass would have these measurements. As such, finding a piece of glass which matches the crime scene, and matches the glass from this commonly used factory, would be considered as weak evidence, given that much of the glass in that city has that composition.

1.3 Thesis Summary

Having introduced the thesis in this chapter, in Chapter 2 we move on to detail the necessary background information about the data on which we will test our methods, as well as the statistical classification methodology which we employ throughout. Next, in Chapter 3, we give a review of currently used methods to analyse glass data, in particular the best practice approach used by Forensic Science SA, and apply this approach to our data. This allows us to establish a baseline level of performance to which other methods can be compared. In Chapters 4 and 5 we then discuss the use of likelihood ratio (LR) based techniques incorporating a background database of evidence. These methods allow for the added benefit of quantifying the strength of evidence. Chapters 6 and 7 move on to the final method of comparison: machine learning classification. In these chapters we detail the background theory of decision tree and random forest models, and then apply these as well as logistic regression to our data sets. Finally, we bring together the ideas explored in the likelihood ratio and machine learning sections in Chapter 8, with the introduction of score-based likelihood ratios. These take the results from machine learning methods (as well as some other approaches) and construct likelihood ratios from the distributions of results. In Chapter 9 we summarise the benefits and shortcomings of each of the models that we have explored, before concluding the thesis in Chapter 10.

Chapter 2

Analysis Background

In this thesis, we aim to investigate whether different statistical methodologies can improve upon the current best practice techniques for establishing whether or not two glass samples originate from the same source. To do this, we will first establish a baseline level of performance that is achieved when the current methodology is applied to our data sets. This current practice is the use of what is known as the standard 4σ criterion, and is a univariate approach which is conducted individually for each element. We will then consider some alternative methods which take into account the multivariate structure of the data. These methods fall into two broad categories: likelihood ratio-based methods, and machine learning classifiers. Each set of methods will be presented in their own chapter or part of the thesis. In each case, we will provide the necessary mathematical theory to understand and apply the techniques; describe how the methods were implemented; and then apply the procedures to two data sets.

2.1 Data

As mentioned, we consider two data sets in this thesis, which serve as examples of data collected under two very different circumstances. The first data set, the USA ribbon data set, was collected under laboratory-like conditions, and provides an example of data with very little variability. The second set, on the other hand, consists of samples from real forensic casework collected in South Australia. This data set contains a great deal more variability, and is a much more realistic example of a database of forensic samples. By performing our analysis on both sets of data, we test the robustness of the techniques presented in this thesis.

2.1.1 USA Ribbon Data

This data set was commissioned to serve as a database of chemical composition measurements by Park and Carriquiry (2019). The data is comprised only of measurements of float glass manufactured by two companies in the United States of America. These companies will be referred to as Company A and Company B. A sample of 31 panes of float glass was taken from Company A, labelled $AA, AB, \ldots, AAR$, and 17 manufactured by Company B labelled $BA, BB, \ldots, BR$. The panes sampled from Company A were manufactured between the third and 24th of January, 2017, and those produced by Company B were sampled a little earlier from the fifth to the 16th of December, 2016 (Park *et al.*, 2020).

Glass is manufactured in long continuous sheets known as ribbons, which are then cut into panes. At both of these factories, a large number of samples were taken from each ribbon of glass in order to develop an understanding of the level of variability within a source. To achieve this, on almost all days within the sampling periods, for both manufacturers, two glass panes were collected – one from each side of the ribbon. A sample of 24 fragments were then taken from each pane, and from 21 of these fragments, five replicate measurement were taken. For the other three, 20 replicate measurements were made, resulting in a total of 165 measurements per pane of glass.

In this study, the choice of elements to measure was made following Weis *et al.* (2011), who recommended that only 18 elements are used. Three major elements: calcium, sodium and magnesium; three minor elements: aluminium, potassium and iron; and 12 trace elements: lithium, titanium, manganese, rubidium, strontium, zirconium, barium, lanthanum, cerium, neodymium, hafnium, and lead.

2.1.2 Australian Casework Data

The second data set comprises glass collected during casework in South Australia which has been measured and analysed by FSSA. The analysis by FSSA follow the standard $4 - \sigma$ criterion recommended by the guidelines ASTM-E2330-12 (2012) and ASTM-E2927-16 (2016) as will be described in Section 3.1.1. The measurements in this data set span from 2016 to 2020 at the time of writing, and as the data is from real police casework, there is some variability in the number of measurements obtained. The data also come in two distinct forms: control samples and recovered samples. Control samples are taken from a known source at the scene of the crime, and typically at least nine individual fragments are measured, and between one

and three replicate measurements are made of each fragment, depending on what is possible given the size of the fragment. Recovered samples are typically those taken from suspects' clothing as well as other sources such as the interior of a car. These samples are much less numerous, with between one and three fragments measured, and again between one and three replicate measurements for each fragment. Given that these observations are far less numerous and do not have a known source, only control fragments are considered for the analysis throughout in order to maintain a certain level of consistency in the observations used.

In this dataset, the concentrations of 19 chemical isotopes were measured: Lithium 7, Magnesium 24, Aluminium 27, Potassium 39, Calcium 42, Calcium 43, Titanium 47, Manganese 55, Iron 57, Rubidium 85, Strontium 88, Zirconium 90, Tin 118, Barium 137, Lanthanum 139, Cerium 140, Neodymium 146, Hafnium 178, and Lead 208. However, the two calcium isotopes were almost perfectly correlated, and so only Calcium 42 was used for modelling, and Tin was found to be very unreliable and contain some extraneous measurements, and so only the 17 remaining isotopes were used.

Measurements of two sides of a glass laminate

It is worth noting also, that there is significant variety in the sources of measurements in the Australian casework data, in that they can be a number of different types of glass, and in some cases, both sides of a given glass laminate are measured separately, and recorded as different objects. In order to avoid any potential issues, or misleading conclusions being drawn, only one side of such fragments have been included in the data used for analysis, typically the side labelled side A, or side 1.

2.2 Exploratory Analysis

Here we present a brief exploratory analysis of the data to give an insight into the nature of this data set. Figure 2.1 shows the distribution of four of the chemical isotopes for each case within the data. Although this shows only four of the elements in the data set, it gives an idea of the level of variability present in this data. We can see that the level of spread varies from sample to sample, and that there is a high degree of variability between some samples as well.

Preliminary analyses also indicated that there was a significant level of correlation present between some of the variables. In Figure 2.2 we get some

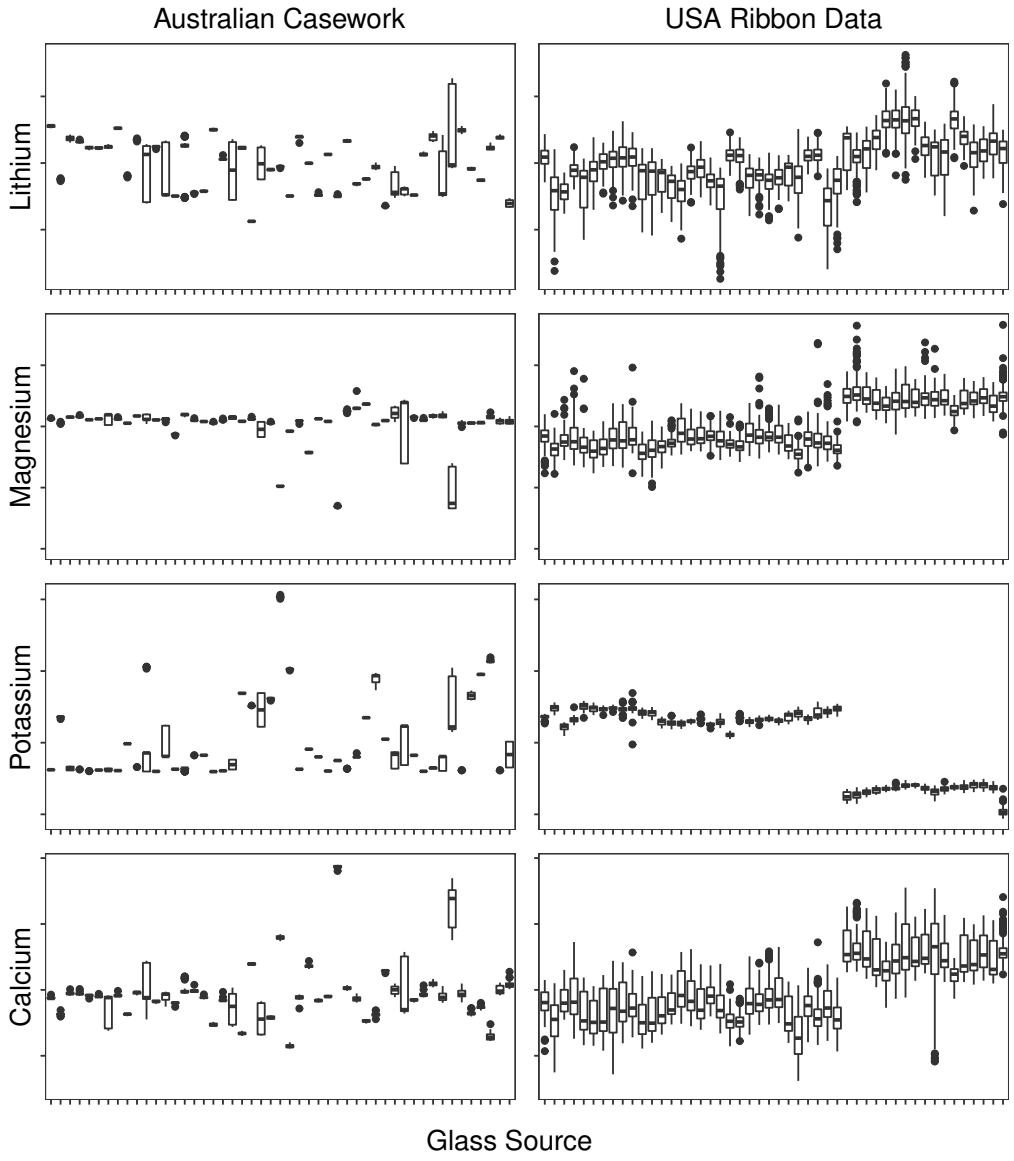


Figure 2.1: Boxplot of parts per million of Lithium 7, Magnesium 24, Potassium 39 and Calcium 42, separated by the source from which the samples originate. In the right column displaying the USA ribbon data, the measurements from both factories are included. In the Potassium plot the distinction between the two factories is most noticeable, with the rightmost observations originating from Factory B. It appears to be the case that the between source variation is more pronounced in the Australian data.

insight into the nature of the relationship between a few of the variables. We see that there is very strong positive correlation between the two calcium isotopes, but also between lanthanum and neodymium, and between zirconium and hafnium. Given just how closely related the two calcium measurements (Ca 42 and Ca 43) were, Ca 43 was removed from the Australian data set, and only Ca 42 was measured in the USA data. The other related measurements were left in the data since there were a number of observations which deviated from this observed co-linearity. In the other comparison plots in this figure we note that the pairwise distributions between elements show little in the way of clear structure.

2.3 Evaluation of Glass Evidence as a Classification Problem

Having provided a brief exploration of the structure of the data, we move on to describe how to frame our analysis as a statistical classification problem.

2.3.1 Classification Models

We are considering data for which each pair of observations originate from either the same source or two different sources. The forensic practitioner cannot determine for certain whether samples share a common source, so instead we are interested in whether pairs of observations can be considered to be "matching". Statistical classification is the problem of identifying whether pairs of glass samples match, based on some training data for which the classes are known. To this end, a statistical classifier, or classification model is a statistical model that predicts to which class a new observation belongs.

2.3.2 Data Preparation

In order for classification techniques to be applied, the number of categories into which the observations can be classified cannot be too large. As such, we consider each pair of observations in the dataset, and a new categorical variable is added identifying the pair as matching or non-matching based simply on whether the measurements are of the same or different panes of glass. The categories are labelled KM (known mate) and KNM (known non-mate). To define this formally, let $\boldsymbol{x}_{ij}$ be the set of all data where i denotes the pane and j denotes the fragment. We consider pairs of observations, to

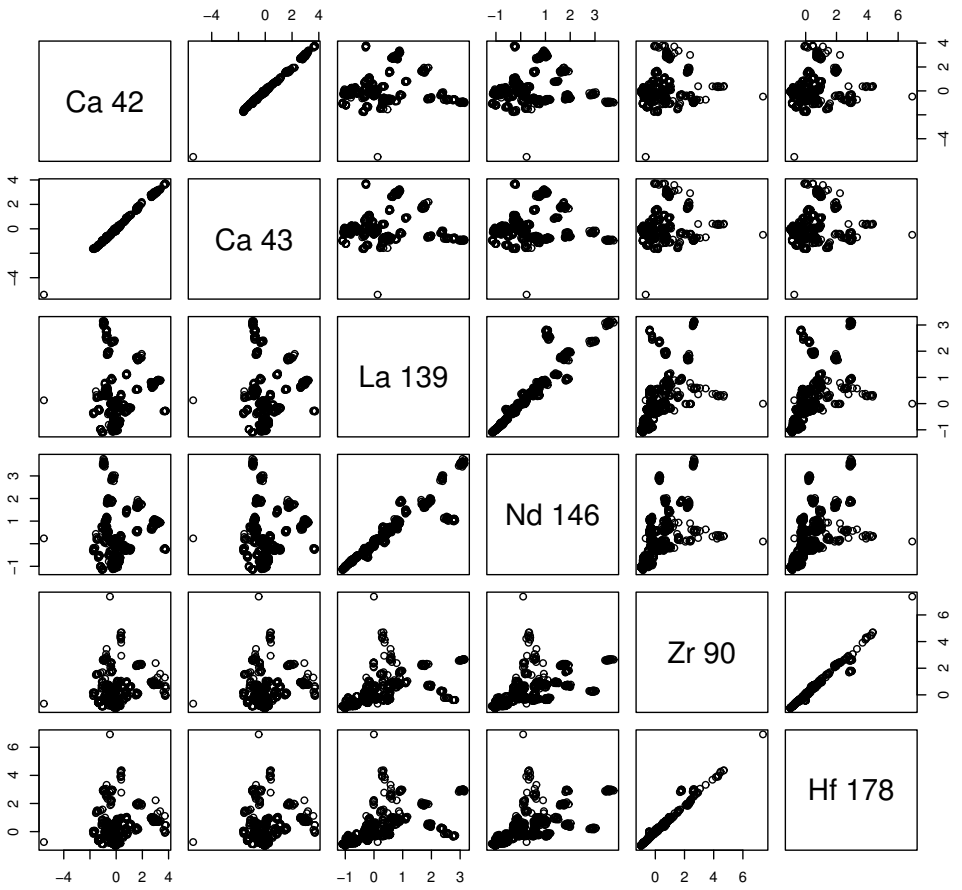


Figure 2.2: Pairwise distributions of calcium 42, calcium 43, lanthanum 139, neodymium 146, zirconium 90 and hafnium 178. We observe very strong positive correlation between the two calcium isotopes, between lanthanum and neodymium, and between zirconium and hafnium.

label as matching or not matching. In this case, the data points are given by

$$\{(\boldsymbol{x}_{i,j}, \boldsymbol{x}_{i',j'}) : i \neq i' \text{ or } j \neq j'\}.$$

That is, pairs of observations $\boldsymbol{x}_{i,j}$ and $\boldsymbol{x}_{i',j'}$ where either the glass sources (first index) are different, or if the sources are the same, the samples (second index) are different. This is the case as we do not need to consider whether there is a match between a single fragment of glass and itself.

2.3.3 Assessing the fit of a model

We use five metrics to compare the models when applied to the training set: accuracy, Cohen's Kappa coefficient, sensitivity, specificity, and the area under the receiver operating characteristic curve (ROC AUC). Accuracy is the most simple and easy to interpret of these metrics, as it is simply the proportion of correct predictions that the model makes when applied to the testing set. Generally speaking, for any classification model one wishes to maximise the accuracy of the model, that is, maximise the proportion of correct predictions made. However, accuracy may not always be the highest priority. In the case of forensic evidence, one is often more concerned with minimising the false positive rate (FPR), than the false negative rate (FNR). The false positive rate being the proportion of times that the model incorrectly classifies samples as the same, and the false negative rate the proportion of times that the model incorrectly classifies samples as different. The reason being that we wish to keep to a minimum the probability of suggesting that a suspect is guilty – *i.e.* that the glass found on their clothing matches that at the crime scene – when in reality they are innocent. It is to this end that we consider sensitivity and specificity.

Sensitivity and Specificity

Sensitivity and specificity are complementary to the false negative and false positive rates respectively. Formally, they are defined as follows

$$\text{Sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}.$$

$$\text{Specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}.$$

Sensitivity is sometimes also referred to as the true negative rate (TNR) and specificity the true positive rate (TPR). As such,

$$\text{Sensitivity} = \text{TPR} = 1 - \text{FNR},$$

and

$$\text{Specificity} = \text{TNR} = 1 - \text{FPR}.$$

Hence we find ourselves with the aim of maximising the specificity of our classification model. These quantities can also be compared intuitively using what is known as a confusion matrix. A confusion matrix is a 2×2 table displaying the proportions of true positives and negatives which have been predicted as positives and negatives. The general structure of a confusion matrix is shown in Table 2.1.

		Truth	
		Same Source	Different Source
Prediction	Match	TPR	FPR
	Non Match	FNR	TNR

Table 2.1: Example of a confusion matrix.

Alternatively, a confusion matrix may contain the number of true and false positive and negatives rather than the proportions, though it is more easily interpreted when containing the proportions.

Cohen's Kappa Coefficient

Cohen's kappa coefficient, κ, can be thought of as another measure of accuracy, corrected for data in which each category has different counts (Cohen, 1960). Most generally, the kappa coefficient compares the reliability between two raters, *i.e.* two people who are classifying observations into groups, or two models. In the context as a metric for evaluating a machine learning model, kappa compares the classification model against one which would randomly allocate new observations into classes, based on the proportion of observations in the training and testing data which fall into those classes. To properly define the kappa coefficient, we must first consider the two quantities used in its calculation: observed accuracy and expected accuracy. Observed accuracy, p_0, is simply the same as accuracy discussed above: the raw proportion of correct classifications when the model is applied to the testing data. Expected accuracy, p_e, on the other hand, is the expected value of accuracy that an entirely random classifier would have, given the confusion matrix of

the model applied to the testing data. For each class, what is known as the marginal frequency of the truth is multiplied by the marginal frequency of the prediction. That is, the true count of observations in that class, multiplied by the number of predicted observations in that class. The expected accuracy is obtained by summing these values, and dividing by the total number of observations in the testing data squared (Cohen, 1960). More formally, letting T_i and P_i be the number of observations which truly belong to class i and are predicted to belong to class i respectively, and N the total number of observations in the data, the expected accuracy is given by

$$p_e = \frac{\sum_i T_i P_i}{N^2}. \tag{2.1}$$

Cohen's kappa coefficient is then defined as

$$\kappa = \frac{p_0 - p_e}{1 - p_e}. \tag{2.2}$$

The kappa coefficient is most important when there is imbalance between the classes in the data set, which is very much the case in this analysis, as will be explained in Section 6.2 in some detail. Its value is best interpreted in comparison with other models, as other than some fairly arbitrary magnitudes which have been suggested, such as those by Landis and Koch (1977), and those by Fleiss (1973), there are no clear cut guidelines as to what is a "good" value of kappa. With that said, kappa takes values between -1 and 1, with values less than zero implying that the classifier is worse than a uniform classifier. Overall, the larger its value, the more accurate the classification model is compared to a uniform classifier.

Receiver Operating Characteristic (ROC)

Finally, another useful assessment of the reliability of a binary classification model is given by the receiver operating characteristic (ROC) curve. The ROC curve quantifies the compromise that one must often make between sensitivity and specificity. It is a plot of sensitivity (true positive rate) against $1-$specificity (false positive rate) as the classification threshold of the model is varied between its maximum and minimum values. By classification threshold, we mean a cut-off value by which a classification model makes a prediction. The details of specific thresholds will be discussed in more detail in the coming chapters. In the case of the interval-criteria in Chapter 3, the threshold can take any positive real value, for likelihood ratios in Part II, the threshold could be any real number. For now, to simplify the explanation,

consider a classification model that predicts the probability that two fragments of glass match, as per the machine learning models in Part III. This prediction must take values between zero and one, and so the same must be true for the classification threshold. Given that this classification model predicts match probabilities, this threshold is the lower bound on this probability for which two samples are predicted to be a match. When the threshold for classification is one, almost no pairs of samples will be predicted to be matching, and so both the false positive and true positive rates will be zero. As such, sensitivity will be zero, and specificity will be one, so 1−specificity will be zero as well. Then, as the threshold is decreased from one, some pairs will be classified as matching, and so the false positive and true positive rates will increase. When the classification threshold reaches zero, there will be no negative predictions, and so the false negative and true negative rates will both be zero. As such, sensitivity and 1−specificity will both be zero. For perfect predictors, the area under the ROC curve (ROC AUC) will be equal to one and as such, the ROC AUC gives another measure for the predictive performance of a classification model.

With the necessary background out of the way, we now move on to presenting the analysis. In the next chapter we discuss the current best practice methodology employed by forensic practitioners in South Australia, and apply these methods to our data.

Chapter 3

Current Practice for Analysis of Glass Evidence

In this chapter we provide a review the standard best practice analysis, and apply it to both the homogenous USA data and the diverse Australian casework data. We explore the imitations associated with these methods and propose a multivariate extension to the univariate method. This current best practice approach treats each element individually, while the method we present generalises the approach to incorporate the covariance between the elements.

3.1 Match Criteria

In this section we introduce some criteria by which pairs of glass samples are classified as matching. By this, we mean that the samples cannot be distinguished from one another by the given statistical procedure. On the other hand, samples which can be distinguished from one another, are classifies as non-matching.

3.1.1 Interval-Based Approach

The standard practice for comparing glass samples with LA-ICPMS elemental analysis is via some interval-based match criteria (Park and Carriquiry, 2019). In this setting, we consider a known (K) and unknown, or questioned (Q), sample of glass. The known and questioned labels are entirely analogous with the control and recovered labels respectively which were introduced in Chapter 2. Authors make use of both of these sets of terminology and we

will use the two interchangeably. To simplify the notation, we will denote the known and questioned samples by $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ respectively. For each sample, the concentrations of p elements are measured, for n_l fragments of a given sample of broken glass (where $l = 1, 2$). In some cases there will also be replicate measurements taken of each fragment. In order to avoid over-complicated notation, we assume that replicate measurements are contained in the index spanning the number of fragments. That is, we take n_l to span the fragments and replicates for sample l. In full, the control and recovered measurements are denoted

$$\boldsymbol{y} = \{y_{ljk} \,|\, l = 1, 2, j = 1, \ldots, n_l, k = 1, \ldots, p\}\,.$$

From this, we have that each measurement is a vector of the form

$$\boldsymbol{y}_{lj} = (y_{lj1}, \ldots, y_{ljp}).$$

The mean vector for each element in sample l is then denoted

$$\bar{\boldsymbol{y}}_l = \frac{1}{n_l} \sum_{j=1}^{n_l} \boldsymbol{y}_{lj},$$

and similarly the vector of standard deviations is

$$\boldsymbol{\sigma}_l = \frac{1}{n_l - 1} \sum_{j=1}^{n_l} (\bar{\boldsymbol{y}} - \boldsymbol{y}_{lj})^2.$$

That is,

$$\bar{\boldsymbol{y}}_l = (\bar{y}_{l1}, \ldots, \bar{y}_{lp}),$$

and

$$\boldsymbol{\sigma}_l = (\sigma_{l1}, \ldots, \sigma_{lp}),$$

Several criteria have been proposed and are used for interval-based comparisons. The two most commonly used are the standard 4σ criterion and modified 4σ criterion. These methods are described in the ASTM-E2330-12 (2012) and ASTM-E2927-16 (2016) guidelines, and both involve element-wise comparisons of glass samples.

Standard 4σ criterion

We begin with what is known as the standard 4σ interval criterion (Almirall and Trejos, 2006, Weis *et al.*, 2011, Trejos *et al.*, 2013a,b, Almirall and Trejos, 2015, ASTM-E2330-12, 2012, ASTM-E2927-16, 2016). For this standard

criterion, we consider two glass samples $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, where $\boldsymbol{y}_1$ is from a known source and $\boldsymbol{y}_2$ is the sample in question. The guidelines suggest that at least nine measurements be taken from the known source via three replicates of three fragments and that "as many measurements as are practical" be taken of the sample in question (ASTM-E2927-16, 2016). Comparison intervals are then constructed for each element individually. The k-th comparison interval is computed as the mean of the control sample plus or minus four times its standard deviation. However, a lower bound of 3% of the mean is placed on the standard deviation (ASTM-E2330-12, 2012). As such, the k-th comparison interval is defined as

$$\bar{y}_{1k} \pm 4 \times \max\left\{\sigma_{lk}, 0.03 \times \bar{y}_{1k}\right\}. \tag{3.1}$$

In other words, no matter the number of measurements obtained for the known fragment, the variability in this data can never be less than 3% of the control mean. The concentrations of each of the p elements in the questioned sample are then compared to the intervals calculated in Equation 3.1. If the concentrations lie within the interval for all elements, then $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are said to be chemically indistinguishable. In the event that one or more concentration lies outside its respective interval for any of the measurements, the samples are said to be chemically distinguishable. Equivalently, the distinction between two samples is quantified with a score determined by rearranging Equation 3.1. That is, we take the absolute difference between the concentrations of a the k-th element for the two samples and scale by the standard deviation of $\boldsymbol{y}_1$. We denote this by $S_{ASTM,k}$. The overall comparison score between the samples, denoted S_{ASTM}, is then taken to be the maximum across the p elements.

$$S_{ASTM,k} = \left| \frac{\bar{y}_{1k} - \bar{y}_{2k}}{\max\left\{\sigma_{1k}, 0.03 \times \bar{y}_{1k}\right\}} \right|. \tag{3.2}$$

$$S_{ASTM} = \max_{1 \leq k \leq p} S_{ASTM,k}.$$

This score exceeding four is equivalent to the questioned sample lying outside of the comparison interval and so the samples are declared distinguishable in this case (ASTM-E2330-12, 2012, ASTM-E2927-16, 2016).

Modified 4σ criterion

Weis *et al.* (2011) recommend a modified 4σ as an alternative to the standard criterion. In fact they proposed a more general $r\sigma$ criterion, but found that $r = 4$ gave the most desirable combination of specificity and sensitivity.

Recall from Section 2.3 that sensitivity and specificity are metrics to assess a binary classification model. Sensitivity is the ratio of the number of true positive predictions to the total number of positive predictions. Specificity is the ratio of the number of true negative predictions to the total number of negative predictions. To construct the comparison intervals in this case, one computes a fixed relative standard deviation (FRSD). That is, the standard deviation of the control sample divided the corresponding mean, for each of the p elements in a sample. Similarly to the standard criterion, when the standard deviation is below 3% of the mean the FRSD is set equal to 0.03. Using the FRSD, Weis *et al.* (2011) suggest constructing the following interval:

$$\left(\frac{\bar{\boldsymbol{y}}_{1k}}{1 + 4 \times \text{FRSD}_k}, 1 + 4 \times \text{FRSD}_k \right).$$

As for the standard criterion, if the measurements of all elements in $\boldsymbol{y}_2$ fall within the interval constructed for $\boldsymbol{y}_1$, then the two samples are said to be chemically indistinguishable. Otherwise, they are chemically distinguishable. Weis *et al.* (2011) also transform this interval to be represented with a score $S_{BKA,k}$. As before, the overall comparison score is taken to be the maximum score across the p elements.

$$S_{BKA,k} = \frac{\exp(|\log \bar{\boldsymbol{y}}_{1k} - \log \bar{\boldsymbol{y}}_{2k}|) - 1}{\text{FRSD}_k}.$$

$$S_{BKA} = \max_{1 \leq k \leq p} S_{BKA,k}.$$

Multivariate Region of 4σ

The interval-based criteria are most intuitively thought of as p individual univariate tests. However, geometrically, this is equivalent to constructing a hyper-rectangle around the observations whose dimension is the number of elements measured. This hyper-rectangle would be located at the centroid of the data, and the length in each dimension would be 8σ. Then, a recovered sample is declared as matching if it lies within this hyper-rectangle, and non-matching if it lies outside. The construction of this hyper-rectangle assumes no relationship between any of the elements. Therefore, the most natural extension of the standard interval criterion to take into account correlations between the elements would be to consider a p-dimensional ellipsoid centred about the mean vector. We would then wish for this ellipsoid to be rotated such that its principal axes align with the directions of greatest variation, and that the surface is always four standard deviations from the centre. As far as the author is aware, there is no mention of this in the forensic glass

examination literature, though it can be constructed quite easily by employing the Mahalanobis distance. The Mahalanobis distance is an example of what is known as a statistical distance, and provides a well-defined notion of distance between observations. To define it, let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two random vectors from the same distribution with covariance matrix Σ. Then the Mahalanobis distance $d_M(\boldsymbol{X}, \boldsymbol{Y})$ between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is given by

$$d_M(\boldsymbol{X}, \boldsymbol{Y}) = \sqrt{(\boldsymbol{X} - \boldsymbol{Y})^T \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{Y})}.$$

The Mahalanobis distance is a generalisation of Euclidean distance, scaled by the standard deviation in the direction between the two points. As a result, the numerical value of the Mahalanobis distance is precisely the number of standard deviations between the two points. It is worth noting also that in one dimension, the Mahalanobis distance simplifies precisely to the standard 4σ score as expressed in Equation 3.2. Further, by considering a centroid $\boldsymbol{\mu}$ and an observation $\mathbf{x}$, both in $\mathbb{R}^p$, a p-dimensional ellipsoid centred at $\boldsymbol{\mu}$, whose distance from the edge to the centre is equal to r standard deviations is given by the equation

$$\left(d_M(\boldsymbol{\mu}, \mathbf{x})\right)^2 = (\boldsymbol{\mu} - \mathbf{x})^T \Sigma^{-1} (\boldsymbol{\mu} - \mathbf{x}) = r^2. \tag{3.3}$$

This is referred to as the standard deviational ellipsoid and any point which lies within this hyper-ellipsoid is within r standard deviations of the mean vector, $\boldsymbol{\mu}$. Thus, by classifying matches for comparisons where the Mahalanobis distance is less than four, and non-matches where it is greater than four, this provides a natural generalisation of the standard 4σ interval method. For control and recovered samples $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, the comparison score in this case is therefore given by

$$S_{\text{ellipsoid}} = d_M(\boldsymbol{y}_1, \boldsymbol{y}_2).$$

To understand why this works, first consider that the general equation for an ellipsoid centred at $\boldsymbol{w} \in \mathbb{R}^p$ is

$$(\boldsymbol{w} - \boldsymbol{u})^T A (\boldsymbol{w} - \boldsymbol{u}) = 1,$$

where $\boldsymbol{u}, \boldsymbol{w} \in \mathbb{R}^p$ and A is a $p \times p$ positive definite matrix. Here, A defines the scale and rotation of the ellipsoid. Let $\lambda_1, \ldots, \lambda_p$ be the eigenvalues of A. Then the semi-axes of the ellipsoid are given by $\lambda_1^{-2}, \ldots, \lambda_p^{-2}$. The eigenvectors of A then define the principal axes of the ellipsoid, that is, its rotation in $\mathbb{R}^p$. Now, in the context of data, we note that the eigenvectors of Σ define the principal components of the data, that is, the orthogonal vectors of

greatest variance. The corresponding eigenvalues then provide the size of the deviation in each respective direction. Let $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_p$ be the eigenvectors and $\varsigma_1^2, \ldots, \varsigma_p^2$ the eigenvalues of Σ. Here, ς_i^2 is the variance in the i-th principal component. Then, since Σ is symmetric and invertible, $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_p$ are the eigenvectors of Σ^{-1} also, and $\varsigma_1^{-2}, \ldots, \varsigma_p^{-2}$ are its eigenvalues. Thus, since Σ^{-1} is also positive definite, we have that the equation

$$(\boldsymbol{\mu} - \mathbf{x})^T \Sigma^{-1} (\boldsymbol{\mu} - \mathbf{x}) = 1$$

defines a p-dimensional ellipsoid centred at $\boldsymbol{\mu}$, with principal axes $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_p$, and semi-axes $\varsigma_1, \ldots, \varsigma_p$. Finally, we can rewrite Equation 3.3 as

$$(\boldsymbol{\mu} - \mathbf{x})^T r^{-2} \Sigma^{-1} (\boldsymbol{\mu} - \mathbf{x}) = 1.$$

The matrix $r^{-2}\Sigma^{-1}$ has the same eigenvectors as Σ, and has eigenvalues $r^{-2}\varsigma_1^{-2}, \ldots, r^{-2}\varsigma_p^{-2}$. Therefore, we have that Equation 3.3 defines a p-dimensional ellipsoid centred at $\boldsymbol{\mu}$, with principal axes $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_p$, and semi-axes $r\varsigma_1, \ldots, r\varsigma_p$. That is, each semi-axis has length r standard deviations.

To help visualise the distinction between the standard and ellipsoid 4σ criteria consider Figure 3.1. The data used in this figure has been simulated to clearly demonstrate the potential distinction. To keep the visualisation simple, only two variables have been used. Variables 1 and 2 have been simulated from a bivariate normal distribution, with covariance matrix given by

$$\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}.$$

The black points show the simulated data, the straight black lines show the boundaries for the standard criterion in each variable, and the black ellipse is that for a Mahalanobis distance of four. The red cross shows a new observation which would be declared as matching by the standard criterion, but rejected by the ellipsoid method, while the opposite is true for the blue cross.

Covariance Matrix Shrinkage Estimator

Often, in practice, the number of elements measured, p, is greater than the number of observations in each sample (number of fragments multiplied by number of replicate measurements). As such, the sample covariance matrix is not guaranteed to be well-conditioned, meaning that the inverse cannot always be computed accurately. This poses issues in calculating the Mahalanobis distance. In an effort to combat this, Campbell and Curran (2009)

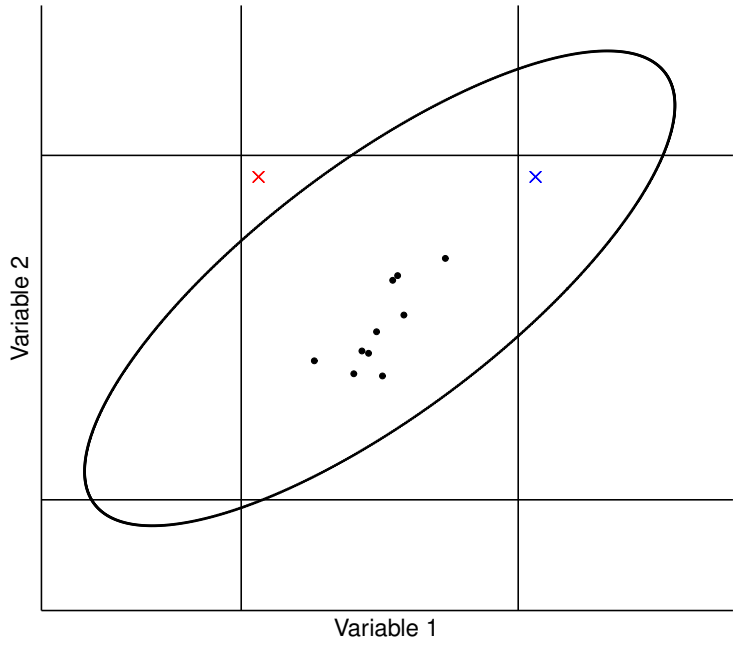


Figure 3.1: Comparison of standard and ellipsoid 4σ match criteria in two variables. Variables 1 and 2 have been simulated from a bivariate normal distribution, each with variance 1, and the covariance is 0.5. The black points show the simulated data, the straight black lines show the boundaries for the standard criterion in each variable, and the black ellipse is that for a Mahalanobis distance of four. The red cross shows a new observation which would be declared as matching by the standard criterion, but rejected by the ellipsoid method, while the opposite is true for the blue cross.

recommend using a shrinkage estimator for the covariance matrix proposed by Ledoit and Wolf (2004). This method seeks to strike a balance between the unbiased sample covariance matrix, and a much more highly structured estimator. This structured estimator is referred to as the target matrix, denoted F. The standard sample covariance is replaced by $\hat{\Sigma}_s$ (Campbell and Curran, 2009, Schfer and Strimmer, 2005), given by the convex linear combination:

$$\hat{\Sigma}_s = \hat{\delta}^* F + (1 - \hat{\delta}^*)\Sigma.$$

Here, $\hat{\delta}^* \in [0, 1]$ is an optimised shrinkage constant which minimises the distance between Σ and F. Ledoit and Wolf (2004) suggest a constant correlation model for F which achieves a good compromise between performance and ease of implementation. To construct this target, the average of all of the sample correlations is used to estimate the constant correlation. The matrix is then constructed as follows, using this average, and the vector of sample

variances. Continuing with our pooled sample covariance matrix $\Sigma = [\sigma_{ij}]$ (Equation 3.4), the sample correlations are given by

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}.$$

The mean sample correlation is then given by

$$\bar{\rho} = \frac{2}{p(p-1)} \sum_{i=1}^{p-1} \sum_{j=1}^{p} \rho_{ij}.$$

Finally, the target matrix $F = [f_{ij}]$ is now defined as

$$f_{ij} = \begin{cases} \sigma_{ii} & \text{if } i = j, \\ \bar{\rho}\sqrt{\sigma_{ii}\sigma_{jj}} & \text{if } i \neq j. \end{cases}$$

3.1.2 Hypothesis Tests

As an alternative to match criteria, hypothesis tests have been proposed to classify samples as matching or not. We do not include these methods in our analysis and comparison, as they offer little distinction from the interval-based criteria, but present them here in the interests of including a comprehensive review of currently used techniques in the field.

Multiple t-tests

The most basic approach, which is somewhat parallel to the univariate intervals, is to simply conduct p independent t-tests (Aitken and Lucy, 2004). While this method is not exactly the same as the standard 4σ interval criterion, it differs only slightly. By choosing an appropriate significance level, the confidence interval of the t-test can be considered to be almost equivalent to the 4σ interval. The difference is only in that if a two sample t-test were used, it would account for the variability in the recovered sample as well as the control sample, using a pooled standard deviation rather than just that of the control sample.

Hotelling T^2 Test Statistic

As another alternative to the interval-based methods, Campbell and Curran (2009) suggest the use of the two sample Hotelling T^2 test statistic for the comparison of multivariate means. This approach can be considered as a

multivariate generalisation of the t-test. The Hotelling test is to the ellipsoid criterion, what the multiple t-test approach is to the standard interval criteria. In fact, the Hotelling test statistic is essentially a small modification to the squared Mahalanobis distance. This approach also assumes that the data is normally distributed, that is,

$$\boldsymbol{y}_{11}, \ldots, \boldsymbol{y}_{1n_1} \overset{\text{iid}}{\sim} N(\boldsymbol{\mu}, \boldsymbol{V})$$

and

$$\boldsymbol{y}_{21}, \ldots, \boldsymbol{y}_{2N_2} \overset{\text{iid}}{\sim} N(\boldsymbol{\mu}, \boldsymbol{V}).$$

As before, we take $\bar{\boldsymbol{y}}_1$ and $\bar{\boldsymbol{y}}_2$ to be the respective vectors of sample means, and $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ the respective sample covariance matrices. The unbiased pooled sample covariance matrix is then defined as

$$\hat{\Sigma}_{\text{pooled}} = \frac{(n_1 - 1)\hat{\Sigma}_1 + (n_2 - 1)\hat{\Sigma}_2}{n_1 + n_2 - 2}, \tag{3.4}$$

Now, the Hotelling two-sample test statistic is defined as

$$t^2 = \left(\frac{n_1 n_2}{n_1 + n_2}\right) (\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)^T \hat{\Sigma}_{\text{pooled}} (\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2).$$

Under the assumption of multivariate normality, the statistic has the Hotelling T^2 distribution. In particular,

$$t^2 \sim T^2(p, n_1 + n_2 - 2).$$

This, however, is simply a transformation of an F-distribution. Specifically,

$$\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} t^2 \sim F(p, n_1 + n_2 - 1 - p).$$

As such, this F-distribution can then be used to evaluate p-values and test the null hypothesis. As discussed in the previous section, the shrinkage estimator for the covariance matrix generally needs to be used to compute the Hotelling test statistic. While we do not apply and compare the Hotelling T^2 test here, the T^2 test statistic will be of importance in the construction of some likelihood ratios in Chapter 4.

3.2 Results

In this section we apply the interval-based criteria to the Australian casework and USA ribbon data sets to establish a performance benchmark against

which we can compare the more advanced methods discussed in Parts II and III. Since the hypothesis testing approaches – multiple t-tests and the Hotelling T^2 test – are effectively equivalent to the interval based methods, only these two methods are considered here.

3.2.1 USA Ribbon Data

We begin by applying the two methods to the USA ribbon data. We note in Table 3.1 that the standard 4σ criterion was only approximately 60% accurate, and that its low value of kappa suggests little difference between it and a uniform random classifier. It received a perfect score for sensitivity, meaning that it never incorrectly classified same-source pairs as non-matching, but received quite a low score for specificity, meaning a large number of false positive predictions. The ellipsoid 4σ criterion appears to have significantly improved upon the standard method, achieving a raw accuracy of 0.950. It also scored above 0.9 for both sensitivity and specificity, with specificity in particular scoring above 0.95, meaning that false positives were minimised.

method	accuracy	kappa	sensitivity	specificity
Standard 4σ	0.616	0.109	1.000	0.600
Ellipsoid 4σ	0.950	0.575	0.917	0.951

Table 3.1: Performance metrics for standard and ellipsoid 4σ criterion applied to USA ribbon data. The standard criterion is only approximately 60% accurate while the ellipsoid criterion is 95% accurate. The standard criterion scored 0.109 for Cohen's Kappa, suggesting only a slight improvement over a uniform classifier, while the ellipsoid criterion received a score of 0.575 The standard criterion received a perfect score for sensitivity, meaning that it never incorrectly classified same-source pairs as non-matching, but received quite a low score for specificity, meaning a large number of false positive predictions. The ellipsoid criterion has high scores for both sensitivity and specificity, with both above 90%.

To provide another interpretation of sensitivity and specificity, we see that the standard 4σ criterion predicted perfectly on same source pairs, but only 60% of the time on different source pairs. The ellipsoid criterion, by contrast, predicted correctly 91.7% of the time on same source pairs, and 95% of the time on different source pairs. In other words, the ellipsoid criterion sacrifices approximately 8% accuracy on same source pairs, in order to increase different source accuracy by 35 percentage points.

3.2.2 Australian Casework Data

Moving on to the Australian casework data, Table 3.2 displays the performance metrics of these two criteria on this data set. We see that in terms of raw accuracy, both methods perform quite well, with the ellipsoid criterion performing slightly better with a near-perfect score of 0.992, an increase of 0.031 over the standard criterion. In Cohen's kappa coefficient, the difference is more pronounced, with a significant increase of 0.34 from the standard to ellipsoid criterion, suggesting that accounting for correlations has a significant impact on classification as compared to a random classifier. Looking at sensitivity and specificity, we note that the standard criterion achieves a perfect score for sensitivity, suggesting that it predicted no false negatives, that is, no pairs of fragments were classified as non-matching, when they in fact had the same source. The standard criterion traded this high sensitivity for a slightly lower specificity, suggesting that it made a small number of false positive predictions. The ellipsoid criterion, achieved a balance with close to perfect scores for both sensitivity and specificity.

method	accuracy	kappa	sensitivity	specificity
Standard 4σ	0.961	0.497	1.000	0.960
Ellipsoid 4σ	0.992	0.837	0.990	0.992

Table 3.2: Performance metrics for standard and ellipsoid 4σ criterion applied to Australian data. In terms of raw accuracy, both methods perform quite well, with the ellipsoid criterion performing slightly better with a near-perfect score of 0.992, an increase of 0.031 over the standard criterion. In Cohen's kappa coefficient, the difference is more pronounced, with a significant increase of 0.34 from the standard to ellipsoid criterion, suggesting that accounting for correlations has a significant impact on classification as compared to a random classifier. We note also that the standard criterion achieves a perfect score for sensitivity, suggesting that it predicted no false negatives.

The score of 1.000 for sensitivity means that the standard 4σ criterion performed perfectly on same source pairs, and the specificity shows that it correctly predicted 96% of the time on different source pairs. By contrast, the ellipsoid 4σ criterion made correct predictions 99% of the time on same source pairs, and 99.2% of the time on different source pairs.

3.3 Summary

In this chapter we have provided a review of the currently practiced methodologies for the comparison of forensic glass evidence measured by LA-ICPMS in South Australia. We have applied the currently employed interval-based technique (standard 4σ criterion), and suggested a simple multivariate extension to this method (ellipsoid 4σ criterion), accounting for the correlation structure in the data.

Overall, we note that both the standard and ellipsoid criteria perform quite well on the diverse Australian casework data set, with the ellipsoid method offering substantial improvement on different source classifications. The standard criterion performed quite poorly on the homogeneous USA ribbon data, being able to classify same sources pairs as matching, but struggling to correctly classify different source pairs. The ellipsoid method, taking into account the correlation structure between the variables, closed this gap and offered substantial improvement on different source classification, while maintaining good performance on same source comparisons.

Park and Carriquiry (2019) make note of some clear weaknesses in the element-wise univariate interval-based approaches. Variability in the data, whether it be by uncertainty in measurement, or inherent variation in the elemental composition of samples, leads to widening of the intervals. This in turn, has the counterintuitive effect of it being less likely that the hypothesis that the samples originate from the same source is accepted. Also, any correlations between elemental concentrations are not taken into account by the element-wise comparisons. The ellipsoid criterion and the Hotelling T^2 test aim to remedy this by combining all elements into a single score, including the covariance matrix in this calculation.

In none of the methods presented in this chapter, however, is the notion of a coincidental match entertained. The probability that two fragments may be indistinguishable but also come from different sources is not calculated, which, depending on the variation of glass in a certain location, may not be negligible. The methods discussed in Parts II and III address this by considering a background database of samples from known sources to inform the level of variability in a given population of glass samples. In particular, in Part II, the likelihood ratio methodology aims to quantify the strength of evidence for or against a match.

Part II

Likelihood Ratio Approach

Part II Glossary

Terminology

Term	Meaning
Block	A partition of a set of values $(a_1, \ldots, a_n)$. e.g. $(\{a_1, \ldots, a_{m_1}\}, \{a_{m_1+1}, \ldots, a_{m_2}\}, \ldots, \{a_{m_r+1}, \ldots, a_n\})$.
Entropy	In information theory, the entropy of a random variable is the average level of uncertainty contained in the variables potential outcomes. The entropy H of a discrete random variable $\mathbf{X}$ with outcomes $\mathbf{x}_i$ is given by $H(\mathbf{X}) = -\sum_i P(x_i) \log_2 P(x_i)$.

Abbreviations

Abbreviation	Meaning
C_{llr}	Cost Log-likelihood Ratio
ECE	Empirical Cross-Entropy
LR	Likelihood Ratio
LLR	Log-likelihood Ratio
MVK	Multivariate Kernel
MVN	Multivariate Normal
PAV	Pool Adjacent Violators

Chapter 4

Likelihood Ratio Methodology

Having established a baseline level of performance given by the standard and ellipsoid 4σ criteria, in this chapter, we describe the likelihood ratio (LR) methods used to make comparisons between glass samples. We begin by introducing and motivating likelihood ratios as a tool to compare glass samples and describe how the method can be used to add more information by quantifying the strength of the association between samples, rather than a simple black and white declaration of match given by the established criteria. We also discuss how likelihood ratios can be interpreted in a Bayesian setting in conjunction with prior and posterior odds. We next present the three methods for constructing LRs discussed by Aitken and Lucy (2004). We then explain how likelihood ratios can be interpreted, and discuss how one can assess whether a system of calculated LRs is valid. In particular, we discuss the notion of measuring the calibration of a system of LRs, in order to establish whether the numerical outputs can reasonably be interpreted in the correct manner. We suggest two post-hoc, invertible transformations which can be applied to a system of LRs to recalibrate the system without fundamentally changing the information it contains.

4.1 The Likelihood Ratio

The approaches described in Chapter 3 all define some criterion for classifying pairs of glass as matching or not. A clear improvement to approaches such as these would be one which allows for a probabilistic measure of how good a match is, or how significant the difference is when samples are said to originate from different sources. To achieve this, we consider the notion of a likelihood ratio (LR) of the samples originating from the same source.

This can be constructed in a number of different ways, but at the most fundamental level, we consider two quantities: the probability of observing the recovered sample given that it originates from the same source as the control sample, and the probability of observing the recovered sample given that it originates from a different source. These two quantities can be expressed mathematically as $P(E \mid H_p)$ and $P(E \mid H_d)$ respectively, where E is the event that the evidence is observed, *i.e.* obtaining the observed measurements of the recovered glass; H_p is the prosecutor's (same source) hypothesis; and H_d is the defence's (different source) hypothesis. In other words, the numerator is the probability of the evidence given that the samples have the same origin, and the denominator is the probability of the evidence given that they have different origins. The likelihood ratio of matching can then be computed as

$$LR = \frac{P(E \mid H_p)}{P(E \mid H_d)}.$$

If the LR is greater than 1, it supports the same source hypothesis, and if it is less than 1, it supports the different source hypothesis. The strength of the association (or disassociation) is then quantified by how large or small the LR is.

Given that the measurements of chemical composition are continuous and can each theoretically take any value on the positive real line, the exact probability $P(E \mid H_p)$ (and likewise $P(E \mid H_d)$) is equal to zero. To resolve this, the likelihood ratio is calculated using probability densities for the numerator and denominator. That is,

$$LR = \frac{f(E \mid H_p)}{f(E \mid H_d)},$$

where f is the probability density function in each case. The densities can be estimated based on background databases, which we will describe in detail in Section 4.2.

4.1.1 Bayes' Theorem and the Odds Ratio

In forensic science, likelihood ratios are often also considered within the Bayesian framework. In short, Bayesian statistics is founded on combining prior knowledge about an event with observed data, to establish what can be referred to as posterior knowledge. Mathematically, a prior probability, or probability distribution, is multiplied by the likelihood of the observed

data (and then rescaled) to obtain a posterior probability or probability distribution. This is presented in Bayes' theorem:

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{\int p(x \mid \theta)p(\theta)\mathrm{d}\theta},$$

where $p(\theta \mid x)$ is the posterior distribution of θ given the observed data, x; $p(x \mid \theta)$ is the likelihood function of the data, given θ; and $p(\theta)$ is the prior distribution of θ. Then, the integral on the denominator is taken over all possible values of θ, acting as a normalising constant.

Now we can consider Bayes' theorem applied to likelihood and odds ratios. Let $f(H_i)$ be the prior probability density of hypothesis i. Now let $f(H_i \mid E)$ be the posterior probability density of hypothesis i, given the evidence E. That is, the probability that the samples originate from the same or different sources, given the observed measurements. The relationship between these values and the likelihood ratio is given by Bayes' theorem:

$$\frac{f(H_p \mid E)}{f(H_d \mid E)} = \frac{f(H_p)}{f(H_d)} \times \frac{f(E \mid H_p)}{f(E \mid H_d)}.$$

In other words, the posterior odds is proportional to the likelihood ratio, and the constant of proportionality is given by the prior odds. The prior odds represent any prior knowledge which one may have before performing inference, via some non-scientific evidence, for example.

At this stage, it is important to observe the distinction between the likelihood ratio and the posterior odds. The likelihood ratio quantifies how much more or less likely one is to observe the evidence, given the same source hypotheses versus given the different source hypothesis. The posterior odds, however, quantifies, given the evidence, how much more or less likely the same source hypothesis is than the different source hypothesis.

4.2 Methods to Calculate Likelihood Ratios

Having motivated the use of likelihood ratios as a tool to compare glass samples, we must now determine how they can be calculated. In this part of the thesis, we aim to test and validate the accepted methods to calculate likelihood ratios present in the literature. The three methods which we present are those described by Aitken and Lucy (2004). First, using the value of the Hotelling T^2 test statistic to determine the numerator probability, and a univariate kernel density estimate for the denominator. The second uses multivariate normal densities for both the numerator and denominator, while

the third builds on this approach by replacing the denominator density with a multivariate kernel density estimate.

Definitions

For the remainder of this section and onward, we consider the following quantities calculated from the background data, as introduced by Aitken and Lucy (2004). Let N be the total number of samples in the given database, and m the number of groups, that is, the number of individual glass sources. Each group i contains n_i measurements and so $N = \sum_{i=1}^{m} n_i$. Recall also that for each sample we measure the concentrations of p elements. The background data is then denoted

$$\boldsymbol{x} = \{x_{ijk} \,|\, i = 1, \ldots, m, j = 1, \ldots, n_i, k = 1, \ldots, p\}.$$

For each group i, the vector of means for each element is denoted

$$\bar{\boldsymbol{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}.$$

Let $\boldsymbol{\mu}$ be the mean over all groups, estimated by

$$\bar{\boldsymbol{x}} = \frac{1}{m} \sum_{i=1}^{m} \bar{\boldsymbol{x}}_i.$$

Now let U be the within group covariance matrix, estimated by

$$\hat{U} = \frac{S_w}{N - m},$$

where,

$$S_w = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)^T.$$

Let C be the between-group covariance matrix, estimated by

$$\hat{C} = \frac{S^*}{m - 1} - \frac{S_w}{n(N - m)},$$

where,

$$S^* = \sum_{i=1}^{m} (\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})^T.$$

Finally, the pooled, within-group standard deviation is denoted $\boldsymbol{s}$, and is given by

$$s_k = \sqrt{\hat{u}_{kk}},$$

where $\hat{U} = [\hat{u}_{ij}]$.

Next we consider the measured samples which are to be compared. We consider control and recovered measurements as described at the beginning of the chapter. Recall that the control and recovered measurements are denoted

$$\boldsymbol{y} = \{y_{ljk} \,|\, l = 1, 2, j = 1, \ldots, n_l, k = 1, \ldots, p\}.$$

As for the background data, the vector of means for each element is denoted

$$\bar{\boldsymbol{y}}_l = \frac{1}{n_l} \sum_{j=1}^{n_l} \boldsymbol{y}_{lj}.$$

4.2.1 Using Hotelling's T^2 Statistic

The first method makes use of the two sample Hotelling's T^2 test statistic in the calculation of the numerator density, $f(E \,|\, H_p)$. This LR is described by Curran *et al.* (1997) and takes the numerator to be the probability density of the T^2 statistic. The density is calculated under the assumption that the statistic follows the Hotelling T^2 distribution, which, as mentioned in Section 3.1.2, is a transformation of an F distribution. More specifically,

$$t^2 \sim \frac{(N-m-2)p}{N-m-p-1} F_{(p,N-m-p-1)}.$$

This assumption is true only if the original data, that is the elemental measurements are normally distributed. This approach uses a within-group, *i.e.* within glass source, covariance matrix $\hat{U}$ which is estimated from the background population. As per Curran *et al.* (1997), we define

$$t_q^2 = \frac{(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)^T (\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)}{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \hat{\boldsymbol{q}}^T \hat{U} \boldsymbol{q}}$$

where,

$$\hat{\boldsymbol{q}} = \hat{U}^{-1} (\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2).$$

Following the assumption of normally distributed data, consider the statistic t_q^2/r, where $r = (N-m-2)p/(N-m-p-1)$. Then,

$$\frac{t_q^2}{r} \sim F_{(p,N-m-p-1)}.$$

The numerator of the LR is then defined as

$$f_F\left(\frac{t_q^2}{r}\right)\frac{1}{G}, \tag{4.1}$$

where

$$G = \left(\tfrac{1}{n_1} + \tfrac{1}{n_2}\right)\hat{\boldsymbol{q}}^T\hat{U}\boldsymbol{q},$$

and f_F is the probability density function of an F distribution with p and $N - m - p - 1$ degrees of freedom.

For the denominator term, Curran *et al.* (1997) use a kernel density estimate evaluated at the point $(\hat{\boldsymbol{q}}^T\bar{\boldsymbol{y}}_2)^2$. This estimate is made from the background database, transformed to scalars $v_i = (\hat{\boldsymbol{q}}^T\bar{\boldsymbol{x}}_i)^2$, for $i = 1, \ldots, m$. We denote also $z = (\hat{\boldsymbol{q}}^T\bar{\boldsymbol{y}}_2)^2$. The kernel density estimate of the density function is then given by

$$k(z) = \frac{1}{mhs_v}\sum_{i=1}^{m}\phi\left(\frac{z - v_i}{hs_v}\right), \tag{4.2}$$

where,

$$s_v = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(v_i - \bar{v})^2}, \tag{4.3}$$

that is, the sample standard deviation of the v_i. The function ϕ is then the probability density function of the standard normal distribution, and h is a smoothing parameter optimised by

$$h = \left(\frac{4}{2p+1}\right)^{1/(p+4)} m^{-1/(p+4)}. \tag{4.4}$$

Further information about kernel density estimation can be found in Appendix A.4. The likelihood ratio is now given by the ratio of Equations 4.1 and 4.2.

While this method takes advantage of the correlation structure present in the multivariate elemental data, it involves creating a univariate projection of the data, and considers the probability of observing the given projection. The next two methods which we consider instead calculate likelihood ratios using multivariate density functions.

4.2.2 Multivariate Normal Density Approach

The multivariate probability density functions used in this technique and the next consider likelihood ratios expressed in the form

$$LR = \frac{f(\boldsymbol{y}_1, \boldsymbol{y}_2 \,|\, \boldsymbol{\mu}, C, U, H_p)}{f(\boldsymbol{y}_1, \boldsymbol{y}_2 \,|\, \boldsymbol{\mu}, C, U, H_d)}. \tag{4.5}$$

In this procedures, both the numerator and denominator densities will be assumed to be multivariate normal and as such, Aitken and Lucy (2004) refer to this method as the multivariate normal (MVN) procedure.

In the numerator probability density, the prosecutor's hypothesis leads us to the assumptions that the means of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are equal, to say $\boldsymbol{\nu}$. The numerator term can then be expressed as

$$f(\boldsymbol{y}_1, \boldsymbol{y}_2 \,|\, \boldsymbol{\mu}, C, U, H_p) = \int_{\boldsymbol{\nu}} f(\boldsymbol{y}_1 \,|\, \boldsymbol{\nu}, U) f(\boldsymbol{y}_2 \,|\, \boldsymbol{\nu}, U) f(\boldsymbol{\nu} \,|\, \boldsymbol{\mu}, C) \mathrm{d}\boldsymbol{\nu}, \tag{4.6}$$

where f is the probability density function of the corresponding multivariate normal distribution. Aitken and Lucy (2004) assert that Equation 4.6 can then be shown to be equal to

$$|2\pi U|^{-(n_1+n_2)/2} |2\pi C|^{-1/2} \left| 2\pi \left((n_1+n_2)U^{-1} + C^{-1} \right)^{-1} \right|^{1/2} \times \exp\left(-\tfrac{1}{2} \left(H_1 + H_2 + H_3 \right) \right),$$

where,

$$H_1 = \sum_{l \in \{K,Q\}} \mathrm{tr}\left(S_l U^{-1} \right), \tag{4.7}$$

$$H_2 = (\boldsymbol{y}^* - \boldsymbol{\mu})^T \left(\frac{U}{n_1+n_2} + C \right)^{-1} (\boldsymbol{y}^* - \boldsymbol{\mu}), \tag{4.8}$$

$$H_3 = \frac{n_1 n_2}{n_1+n_2} (\bar{\boldsymbol{y}}_K - \bar{\boldsymbol{y}}_Q)^T U^{-1} (\bar{\boldsymbol{y}}_K - \bar{\boldsymbol{y}}_Q), \tag{4.9}$$

$$\boldsymbol{y}^* = \frac{n_1 \bar{\boldsymbol{y}}_1 + n_2 \bar{\boldsymbol{y}}_2}{n_1+n_2},$$

$$S_l = \sum_{j=1}^{n_l} (\boldsymbol{y}_{lj} - \bar{\boldsymbol{y}}_l)(\boldsymbol{y}_{lj} - \bar{\boldsymbol{y}}_l)^T.$$

Intuitively, the three terms H_1, H_2 and H_3 can be explained as follows. H_1 (Equation 4.7) quantifies the variability within each group of glass; H_2 (Equation 4.8) accounts for the rarity the measured elemental compositions

in the background database, by measuring the distance from the mean; and H_3 (Equation 4.9) is a measure of the distance between the control and recovered samples.

Next, the denominator probability is given by

$$f(\boldsymbol{y}_1, \boldsymbol{y}_2 \,|\, \boldsymbol{\mu}, C, U, H_d) = \int_{\boldsymbol{\nu}} f(\boldsymbol{y}_1 \,|\, \boldsymbol{\nu}, U) f(\boldsymbol{\nu} \,|\, \boldsymbol{\mu}, C) \mathrm{d}\boldsymbol{\nu} \int_{\boldsymbol{\nu}} f(\boldsymbol{y}_2 \,|\, \boldsymbol{\nu}, U) f(\boldsymbol{\nu} \,|\, \boldsymbol{\mu}, C) \mathrm{d}\boldsymbol{\nu}.$$

The separation into two integrals arises from an assumption of the independence of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ which follows from the assumption that they originate from different sources. The two integrals can be shown to be equal to

$$\prod_{l=1}^{2} |2\pi U|^{n_l/2} |2\pi C|^{-1/2} \left| 2\pi \left(n_l U^{-1} + C^{-1} \right)^{-1} \right|^{1/2} \times \exp\left(-\tfrac{1}{2} \mathrm{tr}\left(S_l U^{-1} \right) - \tfrac{1}{2} \left(\bar{\boldsymbol{y}}_l - \boldsymbol{\mu} \right)^T \left(\tfrac{1}{n_l} U + C \right)^{-1} \left(\bar{\boldsymbol{y}}_l - \boldsymbol{\mu} \right) \right). \quad (4.10)$$

The likelihood ratio is then given by the ratio of Equations 4.6 and 4.10. After simplifying, this ratio is then equal to

$$\frac{\left| C \left(n_1 U^{-1} + C^{-1} \right) \left(n_2 U^{-1} + C^{-1} \right) \right|^{1/2} \exp\left(\tfrac{1}{2} \left(H_4 + H_5 \right) \right)}{\left| \left(n_1 + n_2 \right) U^{-1} + C^{-1} \right|^{1/2} \exp\left(\tfrac{1}{2} (H_2 + H_3) \right)}$$

where H_2 and H_3 are as per Equations 4.8 and 4.9 respectively, and

$$H_4 = \left(\boldsymbol{\mu} - \boldsymbol{\mu}^* \right)^T \left(\left(\tfrac{1}{n_1} U + C \right)^{-1} + \left(\tfrac{1}{n_2} U + C \right)^{-1} \right) \left(\boldsymbol{\mu} - \boldsymbol{\mu}^* \right),$$

$$H_5 = \left(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2 \right)^T \left(\left(\tfrac{1}{n_1} + \tfrac{1}{n_2} \right) U + 2C \right)^{-1} \left(\bar{\boldsymbol{y}}_1 \bar{\boldsymbol{y}}_2 \right),$$

$$\boldsymbol{\mu}^* = \left(\left(\tfrac{1}{n_1} U + C \right)^{-1} + \left(\tfrac{1}{n_2} U + C \right)^{-1} \right)^{-1} \left(\left(\tfrac{1}{n_1} U + C \right)^{-1} \bar{\boldsymbol{y}}_1 + \left(\tfrac{1}{n_2} U + C \right)^{-1} \bar{\boldsymbol{y}}_2 \right).$$

4.2.3 Multivariate Kernel Density Estimate Approach

The assumption of normality on between-group variability may not be reasonable in all cases. As a result, Aitken and Lucy (2004) suggesting relaxing this assumption by considering a multivariate kernel density estimate for the between-group distribution. The kernel density function is taken to be that of a multivariate normal distribution with with mean $\bar{\boldsymbol{x}}_i$ and variance $h^2 C$. Aitken and Lucy (2004) refer to this method as the multivariate kernel (MVK) procedure.

For each group i in the background data, it is given by

$$K(\boldsymbol{\mu} \,|\, \bar{\boldsymbol{x}}_i, C, h) = (2\pi)^{-p/2} h^{-p} |C|^{-1/2} \exp\left(-\tfrac{1}{2} h^{-2} \left(\boldsymbol{\mu} - \bar{\boldsymbol{x}}_i\right)^T C^{-1} \left(\boldsymbol{\mu} - \bar{\boldsymbol{x}}_i\right)\right).$$

The complete probability density function is then estimated as the mean of these, that is,

$$f(\boldsymbol{\mu} \,|\, \bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_m, C, h) = \frac{1}{m} \sum_{i=1}^{m} K(\boldsymbol{\mu} \,|\, \bar{\boldsymbol{x}}_i, C, h).$$

To simplify the remaining expressions, we introduce the notation $D_l = n_l^{-1} U$ for $l = 1, 2$. Now, the numerator term in the likelihood ratio can be shown to be equal to

$$\begin{aligned} f(\boldsymbol{y}_1, \boldsymbol{y}_2 \,|\, \boldsymbol{\mu}, C, U, H_p) &= (2\pi)^{-p} |D_1|^{-1/2} |D_2|^{-1/2} |C|^{-1/2} (m h^p)^{-1} \\ &\quad \times \left| D_1^{-1} + D_2^{-1} + \left(h^2 C\right)^{-1} \right|^{-1/2} \\ &\quad \times \exp\left(-\tfrac{1}{2} \left(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2\right)^T \left(D_1 + D_2\right)^{-1} \left(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2\right)\right) \\ \times \sum_{i=1}^{m} \exp\left(-\tfrac{1}{2} \left(\boldsymbol{y}^* - \bar{\boldsymbol{x}}_i\right)^T \left(\left(D_1^{-1} + D_2^{-1}\right)^{-1} + h^2 C\right)^{-1} \left(\boldsymbol{y}^* - \bar{\boldsymbol{x}}_i\right)\right), \quad (4.11) \end{aligned}$$

where $\boldsymbol{y}^* = \left(D_1^{-1} + D_2^{-1}\right)^{-1} \left(D_1^{-1} \bar{\boldsymbol{y}}_1 + D_2^{-1} \bar{\boldsymbol{y}}_2\right)$.

The denominator term, meanwhile, can be shown to be equal to

$$\begin{aligned} f(\boldsymbol{y}_1, \boldsymbol{y}_2 \,|\, \boldsymbol{\mu}, C, U, H_d) = (2\pi)^{-p} |C|^{-1} (m h^p)^{-2} \prod_{l=1}^{2} \Bigg(& |D_l|^{-1/2} \left| D_l^{-1} + (h^2 C)^{-1} \right|^{-1/2} \\ & \times \sum_{i=1}^{m} \exp\left(-\tfrac{1}{2} \left(\bar{\boldsymbol{y}}_l - \bar{\boldsymbol{x}}_i\right)^T \left(D_l + h^2 C\right)^{-1} \left(\bar{\boldsymbol{y}}_l - \bar{\boldsymbol{x}}_i\right)\right) \Bigg). \end{aligned} \quad (4.12)$$

The likelihood ratio is then given by the ratio of Equations 4.11 and 4.12.

4.2.4 Implementation

Due to the high-dimension of the measurements in the datasets, the multivariate normal and multivariate kernel approaches yielded infinite and zero likelihood ratios in a number of cases. To address this issue, the calculations were instead performed in log space resulting in log likelihood ratios which enabled greater interpretability of the results. These values can still be used for prediction, but with a critical value of zero rather than one, as is the case for regular LRs.

4.3 Interpreting a Likelihood Ratio

The use of likelihood ratios to evaluate forensics evidence is motivated by a number of key factors. In particular, LRs can be constructed using legal propositions *i.e.* the probability of observing the evidence given that the suspect is guilty or not. In particular, they have been shown to be the best way to express the strength of forensic evidence (Aitken *et al.*, 2018, Lund and Iyer, 2017). Further, LRs allow the use of background information to establish how common certain measurements are in a given population, and can quantify the strength of a match between two evidence sources. However, one can also think about a likelihood ratio in a similar way to a score: as a numeric output of a procedure which falls above or below some critical value. Similarly to how a value of four standard deviations was chosen for the standard interval criterion in Section 3.1.1, one can choose a critical value and use to this to classify samples as matching or not matching.

4.3.1 Binary Classification

Recall that a likelihood ratio is the ratio of the conditional probability of observing two fragments of glass, given that they originate from the same source, to the conditional probability, given that they do not. If the former is larger than the latter, the LR should be greater than one, and less than one if the opposite is true. As a result, a critical value of one is the natural choice to deem samples as matching or not. When we transform the likelihood ratio to log space, that is, to a log-likelihood ratio (LLR), the values which were previously between zero and one, now lie below zero, and those which were greater than 1, now lie above zero. As such, the natural choice of critical value for LLRs, would be zero.

Alternatively, given a set of likelihood ratios calculated by one of the procedures described above, one could compute the performance metrics for a range of critical values, and choose the value which gives optimal scores in some desired metrics. The range of values over which to test would likely need to be determined by trial and error, and through observation of the complete range of LRs which have been calculated. However, while this method may yield improved classification performance, given that by definition, the critical value of one describes where an LR is in favour of the same source or different source hypothesis, using an alternative critical values brings into question the validity, or at least calibration, of the procedure used to calculate the LR.

4.3.2 Strength of Evidence

After utilising background information to inform how common certain measurements are in a given population, likelihood ratios offer the benefit of quantifying the strength of a match between a pair of fragments. It is well understood that the larger a likelihood ratio is (above 1), the stronger the evidence of match, and the smaller it is (below 1), the stronger the evidence against a match. However, what remains unclear is how one can assign an understanding of how "strong" a given likelihood ratio. With this in mind, we must question how we can establish whether an LR is valid, and from this arises the idea of the calibration of a likelihood ratio. In particular, a poorly calibrated LR might suggest that a pair of fragments have a common source, when in fact they do not, or that the evidence appears to strongly favour one hypothesis, when in fact there is only weak evidence in its favour.

4.4 Validity of Likelihood Ratios

As mentioned in the previous section, when using a likelihood ratio to quantify the strength of evidence for or against a match, it is of vital importance to establish that the LR is valid and well-calibrated, to avoid scores which are misleadingly large or small (Vergeer *et al.*, 2021). In this section we discuss some key performance metrics against which to assess LR systems, and in particular, a method to test their calibration.

4.4.1 Performance Metrics

Meuwly *et al.* (2017) mention three key three performance characteristics to evaluate the validity of a likelihood ratio procedure. The first two are accuracy and discriminating power. Accuracy is of course simply the raw predictive accuracy of the method when viewed as a binary classifier, that is, the proportion of correct predictions. Discriminating power, is defined as a "performance property representing the capability of a given method to distinguish amongst forensic comparisons where different propositions are true" (Meuwly *et al.*, 2017). This can be assessed using sensitivity and specificity in combination, as these metrics quantify how well the method can predict on same source and different source pairs respectively. Thus, if a classifier performs well on both types of sample pair, it can successfully discriminate between them. These first two measures effectively assess the method's performance as a binary classifier. One can then begin to ask how well it performs in terms of quantifying the strength of the evidence. This is assessed

via the calibration of an LR procedure.

4.4.2 Assessing Calibration

To begin to interpret the strength of the evidence shown by an LR, we must first understand what it means for a likelihood ratio to be calibrated. First, as an example, if an LR procedure is to be accurate in its interpretation, an LR of 500 should mean that the probability of obtaining samples with the observed elemental composition is 500 times higher if they have the same source, than from different sources. Similarly, an LR of $\frac{1}{500} = 0.002$ should mean that the samples are 500 times more likely to originate from different sources, than the same. Meuwly *et al.* (2017) and Ramos and Gonzalez-Rodriguez (2013) provide a succinct definition of what it means for an LR system to be well-calibrated. If an LR procedure were perfectly calibrated, any LR calculated by this method would be exactly as big or small as is warranted by the data. That is, the LR can be probabilistically interpreted in the same way as the above example, comparing the strength of the evidence in favour of one hypothesis over the other. Mathematically, an LR system, LR_0, is well-calibrated if the operation of taking the LR is idempotent. That is,

$$LR(LR_0 = V) = \frac{P(LR_0 = V \mid , H_p)}{P(LR_0 = V \mid , H_d)} = V.$$

In other words, the likelihood ratio is the same as the likelihood ratio of itself (Vergeer *et al.*, 2021). A likelihood ratio procedure can be ill-calibrated in a number of different ways. Vergeer *et al.* (2021) mention four key circumstances of ill-calibration. The procedure produces LRs which are: too large, that is all LRs favour the same source hypothesis; too small, that is the LRs favour the different source hypothesis; too extreme, suggesting stronger evidence than is reasonable; or too weak, suggesting weaker evidence than is reasonable.

Vergeer *et al.* (2021) use simulated well- and ill-calibrated likelihood ratio systems to provide a comparison of four metrics to assess the calibration of an LR. These metrics comprise three well-established methods, and a new methods which the authors propose. They found that their new method, coined DevPAV, as well as the well established cost log-likelihood ratio (C_{llr}) provided the best assessment of calibration. At the time of writing, DevPAV was a very new method, while C_{llr} was well established. The two methods are build on the same foundation and Vergeer *et al.* (2021) found that both methods performed well, and so for the remainder of this chapter, we use C_{llr} as our calibration metric. The word calibration can be confusing in its

meaning in the field of reporting likelihood ratios. We wish to make distinction between measuring calibration, and performing calibration. Measuring calibration refers to assessing how well- or ill-calibrated a system of LRs is, generally by using a metric such as C_{llr} to make comparison with a system which is known to be optimally calibrated. Performing calibration, however, refers to applying some transformation or making adjustments in some way to a system of LRs, such that the result is better calibrated than the original system. To help avoid this confusion, in the context of assessing performance we will say assessing or measuring calibration. In the case of making post-hoc adjustments to a system of LRs, we will use the terminology transformation, adjustment or re-calibration.

Pool Adjacent Violators Algorithm

In the interests of assessing the calibration of a system of LRs, it is important to establish a reference point for an optimally-calibrated system. To achieve this, we use what is known as the pool adjacent violators (PAV) algorithm (Ahuja and Orlin, 2001, Zadrozny and Elkan, 2002, Brümmer and Du Preez, 2006).

The PAV algorithm applied to an LR system is performed as follows. Suppose we have a system of likelihood ratios and their corresponding ground truth classes (LR_i, C_i), for $i = 1, \ldots, n$. We begin by sorting the LRs into ascending order. For simplicity, assume that the index i corresponds to this ordering. Now, for each LR, assign a posterior probability p_i where $p_i = 1$ if $C_i =$ same source and $p_i = 0$ if $C_i =$ different source. If a system of LRs is properly calibrated, all of the different source LRs will be less than all of the same source LRs, and so the vector of posterior probabilities $\boldsymbol{p}$, will be of the form

$$\boldsymbol{p} = (0, \ldots, 0, 1, \ldots, 1). \tag{4.13}$$

If this is the case, the PAV algorithm need not be applied. Alternatively, if there are some same source LRs which are less than some different source LRs, we will have a vector of posterior probabilities of the form

$$\boldsymbol{p} = ((0, \ldots, 0), (1, \ldots, 1), (0, \ldots, 0), \ldots, (1, \ldots, 1)). \tag{4.14}$$

Note in Equation 4.14 that each sub-vector of uninterrupted ones or zeros has been contained in parentheses. We will refer to such an sub-vector as a *block*. Note that, in general, assigning blocks is simply a way to partition the set of values p_i, and need not only contain entries with the same value. We will, however, only consider blocks in which each entry has the same value. A block starting at p_q and ending at p_r for some $q \leq r$ will be denoted $[p_q, p_r]$.

A block is said to take a value θ_{qr}, where $p_q = p_{q+1} = \cdots = p_r = \theta_{qr}$. In fact, we will actually define θ_{qr} to be the mean of $p_q, \ldots, p_r$. A pair of adjacent blocks $[p_q, p_r]$ and $[p_r, p_s]$ is said to be *in order* if $\theta_{qr} \leq \theta_{rs}$, and *out of order* otherwise. Note that Equation 4.13 is an example where all blocks are in order. The PAV algorithm is now applied to the vector $\boldsymbol{p}$, partitioned into blocks. We will denote this as

$$\boldsymbol{P} = ([p_1, p_2], [p_2, p_3], \ldots, [p_{n-m}, p_n]),$$

for some m. While there exist out of order blocks, the PAV algorithm will select a pair of out of order blocks $[p_q, p_r]$ and $[p_r, p_s]$, replace them with $[p_q, p_s]$, and compute θ_{qs}. $\boldsymbol{P}$ will then be updated to now contain one less block. This process is then repeated until no out of order blocks remain. At this point, the resulting set of posterior probabilities $\boldsymbol{p}$ will be increasing. This is a simplified version of the algorithm, including only what is relevant in our setting. The full, more general algorithm is described by Zadrozny and Elkan (2002). Once the monotonic set of posterior probabilities has been obtained, using a given prior odds, they can be converted into a set of calibrated likelihood ratios. To do so, the posterior odds are calculated as

$$O_i^{\text{post}} = \frac{p_i}{1 - p_i},$$

and the prior odds, O_i^{prior}, are given by the ratio of the number of same source LRs to the number of different source LRs in the system. From this, the likelihood ratio is calculated using Bayes' theorem. Note that the original input LR values serve only to provide an ordering. The output of so-called calibrated LR's contain no information from the input LRs other than this ordering. This procedure is detailed as pseudocode in Appendix C.

As mentioned by Ramos and Gonzalez-Rodriguez (2013), it is important to make clear that the PAV algorithm is used here only to create a reference point by which to measure the calibration of an LR system. The PAV algorithm creates a set of optimally calibrated LRs, given an observed set, and it is by the comparison of these sets that we can measure the calibration. Recently, some authors have used PAV as a transformation method to adjust the LRs calculated by a given procedure, and then report these adjusted values. We have chosen not to use PAV for this purpose, as it makes changes to the information contained in the LR system that has been calculated. In Section 4.4.3 we propose an alternative invertible post-hoc transformation method which improves the calibration of an LR system, without fundamentally changing the information contained in the system.

Empirical Cross-Entropy and the Cost Log-Likelihood Ratio

Having now established an optimally calibrated system, we can now measure the calibration of a new LR system in relation to this optimal system. To do so, we first introduce the notion of empirical cross-entropy (ECE). ECE serves to address both calibration and discriminating power of an LR system, and is an example of what is known as a strictly proper scoring rule (SPSR) (Savage, 1971, DeGroot and Fienberg, 1983, Gneiting and Raftery, 2007). The SPSR methodology is based on a Bayesian framework, and relies on prior probabilities to construct posterior probabilities by which the performance is measured. In forensic science, LR values are reported, rather than posterior odds ratios as one cannot accurately determine a prior (Cook *et al.*, 1998). To address this, Ramos and Gonzalez-Rodriguez (2013) suggest that the posterior odds ratio is computed for a range of prior probabilities, and the ECE is calculated in each case. The performance of the LR system can then be assessed using this range of ECE values.

With this technicality out of the way, we can now describe how the ECE is calculated. Denote the true hypothesis as H^i in a given comparison of samples. That is, $H^i = H_p$ if the i-th pair of samples originate from the same source, and $H^i = H_d$ if they originate from different sources. The ECE is then given by

$$ECE = -\frac{P(H_p \,|\, I)}{N_p} \sum_{H^i = H_p} \log_2 P(H_p \,|\, E_i, I) - \frac{P(H_d \,|\, I)}{N_d} \sum_{H^j = H_d} \log_2 P(H_d \,|\, E_j, I),$$

where E_i denotes the evidence for the i-th pair of samples, and N_p and N_d are the number of LR values for which each hypotheses is true. For clarity, the first sum is over the set of pairs for which H_p is true, and the second sum is over the set of pairs for which H_d is true. We can also express the ECE explicitly in terms of the prior odds O^{prior} and the i-th likelihood ratio LR_i.

$$\begin{aligned} ECE = &\frac{P(H_p \,|\, I)}{N_p} \sum_{H^i = H_p} \log_2 \left(1 + \left(LR_i \times O^{\text{prior}}\right)^{-1}\right) \\ &+ \frac{P(H_d \,|\, I)}{N_d} \sum_{H^j = H_d} \log_2 \left(1 + \left(LR_j \times O^{\text{prior}}\right)\right), \end{aligned} \tag{4.15}$$

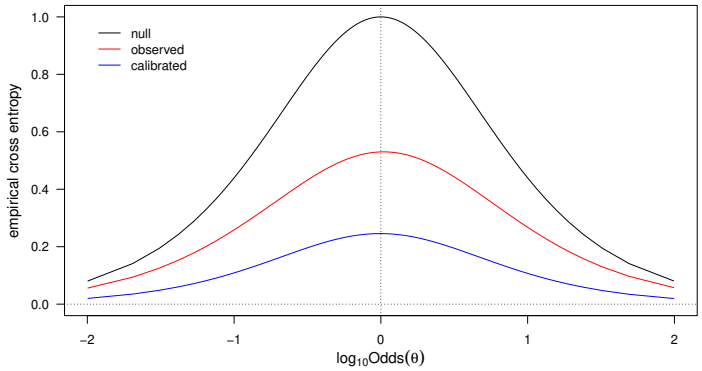
Note that each term in the ECE is weighted by the corresponding prior probability. This weighting is important in the interpretation of the ECE in information theory. A detailed explanation of this can be found in (Ramos *et al.*, 2013).

In Equation 4.15 we see that the ECE depends on the system of LR values and their corresponding known classification of same or different source, as well as the prior odds ratio O^{prior}, due to the Bayesian framework in which an SPSR is constructed. We can then view the ECE as a function of the prior odds, and Ramos and Gonzalez-Rodriguez (2013) suggest assessing LR systems via a plot of ECE against $\log_{10} O^{\text{prior}}$. The ECE of the system of LR values alone, provides a measure of the accuracy of the procedure, but in order to assess discriminating power and calibration, the so-called ECE plot is used. The ECE plot comprises three ECE curves. First, the ECE of the LR system as discussed. The lower the values in this curve, the more accurate the LR values. Second, the ECE after the PAV algorithm has been applied to the set of LR values. This shows the accuracy of optimally calibrated LR values, and thus provides a measure of discriminating power. The difference between the default and calibrated ECE curves provides a measure of calibration. An explicit numerical metric to quantify the calibration is then defined, called the cost log-likelihood ratio, denoted C_{llr}. The C_{llr} is then given by the difference between the default (or observed) and calibrated ECE curves evaluated at $\log_{10} O^{\text{prior}} = 0$ (Vergeer *et al.*, 2021). Finally, a null curve of ECE values for a system in which the LR is one for each pair of samples. This curve serves the purpose of a lower bound of performance, as no method should perform worse than one which gives a value of one for each comparison. Figure 4.1 gives three examples of ECE plots. The first shows a relatively well-calibrated LR system, in which the observed ECE curve sits below the null curve. The second and third plots show poorly calibrated systems which include misleadingly large or small LR values, which in turn, favour one hypothesis over the other. Figure 4.1b shows a system with very small different source LRs, suggesting that the system favours that prediction. Figure 4.1c shows the opposite scenario, in which the system favours same source prediction.

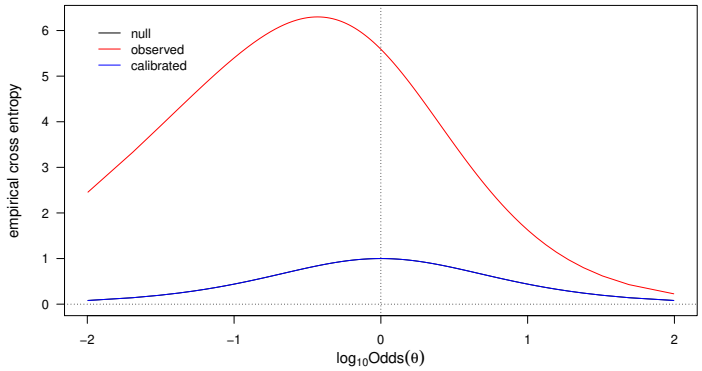
The R package *comparison* (Lucy *et al.*, 2020) was used to compute the PAV transform and create the ECE plots used throughout this chapter and the next.

4.4.3 Optimising the Critical Value

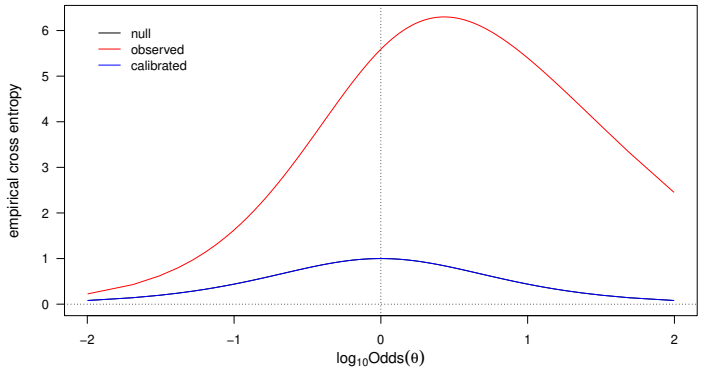
In Section 4.3 we discussed that for likelihood ratios, the natural choice of critical value is one, and for log-likelihood ratios it is zero. If it appears to be the case that a different critical value leads to optimised performance, when the LR is used as a binary clarifier, this may suggest that the LR procedure is not well-calibrated. It is common practice in forensic science to apply a post-hoc operation to recalibrate the system of LRs (Brümmer



(a) LR system in which same source and different source LRs are appropriately scaled. $C_{llr} = 0.284$



(b) LR system in which different source LRs are much smaller than same source LRs are large. $C_{llr} = 4.590$



(c) LR system in which same source LRs are much larger than different source LRs are small. $C_{llr} = 4.590$

Figure 4.1: Example ECE plots showing null, observed and calibrated ECE curves. Figure (a) shows an observed curve which is below the null curve, and so the accuracy is better than the lower bound for $LR = 1$ in all pairs. The difference between the observed and calibrated curves shows the level of calibration, and is quantified by the difference at log odds of zero. Figures (b) and (c) show examples of ill-calibrated LR systems. Figure (b) shows a system in which the different source LRs lie between 10^{-3} and 10^{-2}, while the same source LRs lie between one and ten. Figure (c) shows a system in which the different source LRs lie between 10^{-1} and one, while the same source LRs lie between 10^{2} and 10^{3}.

and Du Preez, 2006, Morrison, 2011, Ramos and Gonzalez-Rodriguez, 2013, Meuwly *et al.*, 2017, Vergeer *et al.*, 2021). As mentioned, the PAV algorithm is often used to perform this calibration, but this procedure applies a non-invertible transformation to the LR system. In particular, it retains only the ordering of the LR values, rather than the values themselves or the way that they are distributed. To address this, we consider a methodology in which an optimal critical value is chosen, and this value is used to recalibrate the calculated LRs using an injective transformation.

There are several ways to choose such a transformation, and there are several factors which must be taken into account. First, considering standard likelihood ratios, any scaling or translation of the LR scores will lead to changes in the classification of pairs of samples. The same is true for translation of log-likelihood ratios. We restrict our attention to transformations of log-likelihood ratios, because they can take any real value, and so we need not concern ourselves with a translation which could lead to a negative likelihood ratio.

Now, given a system of LLRs, suppose we were to find an optimal critical value of $\alpha \neq 0$. This could suggest two things: that the LLR system is poorly calibrated in terms of location, and that it is poorly calibrated in terms of scale. For an LLR optimal critical value which is non-zero, we cannot interpret the LR to mean that the same source hypothesis is however many times more likely than the different source hypothesis. To adjust for this, for each LLR, we could apply the transformation $\phi(s) = s - \alpha$, meaning that zero would now be the optimal critical value in our new log-likelihood ratio system. Further to this, the particular critical value could also be interpreted to indicate that the scale of the likelihood ratio is also poorly calibrated. Suppose, for example, that $\alpha \ll 0$ or $\alpha \gg 0$, that is, there is a substantial deviation from the natural critical value of zero which may suggest that the LLR system is too extreme. In this case, we propose the transformation $\varphi(s) = \frac{s-\alpha}{|\alpha|}$, which will centre and scale the LLR system.

There are potentially countless ways for one to optimise the critical value, and we suggest two criteria which optimise sensitivity and specificity. First, we seek to minimise the trade-off between sensitivity and specificity. That is, we seek a critical value which favours neither sensitivity nor specificity over the other. This is achieved by maximising the absolute difference between the two metrics. We call this the min-trade-off critical value and denote it α_{mt}. Second, we seek to maximise both sensitivity and specificity by maximising their sum, or equivalently their mean, thereby maximising performance on both same source and different source pairs. We call this the max-mean critical value, denoted α_{mm}. For the optimised critical values α_{mt} and α_{mm}

we will denote the corresponding transformations as

$$\varphi_{\mathrm{mt}}(s) = \frac{s - \alpha_{\mathrm{mt}}}{|\alpha_{\mathrm{mt}}|} \quad \text{and} \quad \varphi_{\mathrm{mm}}(s) = \frac{s - \alpha_{\mathrm{mm}}}{|\alpha_{\mathrm{mm}}|}.$$

We find that applying this adjustment substantially improved the calibration of the LR system, and we will discuss the results of this adjustment in Section 5.2 in more detail. To reiterate, unlike adjusting an LR system via the PAV algorithm, as discussed in Section 4.4.2, these transformation methods do not change the information contained in the LR system calculated by any of the procedures described in Section 4.2.

With the theory and methodology out of the way, in the next chapter, we will apply these techniques to compute likelihood ratio systems for the Australian casework and USA ribbon data sets.

Chapter 5

Likelihood Ratio Results

In this chapter we apply the methods described in Chapter 4 to calculate likelihood ratio systems for the Australian casework and USA ribbon data sets. We present only the results for the multivariate normal and multivariate kernel LR procedures, as the Hotelling T^2 statistic procedure has been superseded by the multivariate kernel techniques in the literature by Aitken and Lucy (2004).

We begin by comparing the LR procedures in terms of their performance as binary classifiers via an assessment of accuracy, Cohen's kappa coefficient, sensitivity and specificity. We then proceed to optimise the critical values and assess the calibration of the LR systems. Finding that the MVK procedure consistently outperformed the MVN, we focus on only the MVK method in this section. We begin by finding optimal critical values α_{mt} and α_{mm} which minimise the trade off between, and maximise the mean of sensitivity and specificity respectively. We then apply the corresponding transformations, and use ECE plots and the metric C_{llr} to assess the calibration of the methods before and after these transformations.

5.1 Binary Classification Results

Here we present the results of the likelihood ratio calculations applied to both data sets as a binary classifier. That is, with no transformation applied to the LR systems, and no evaluation of their calibration.

5.1.1 USA Ribbon Data

We begin by applying the multivariate normal and multivariate kernel density LR methods to the USA ribbon data. When this procedure was repeated with the data randomly distributed into each of these sets, it was found that there was some variability in the results. To address this, the procedure was repeated ten times for each model, with the background randomly sampled in each case. We note that the results were skewed in some cases, and present the median of the ten repeats for the model fit metrics.

Table 5.1 displays the confusion matrix for the multivariate normal likelihood ratio method applied to the USA ribbon data. We note that this method only correctly predicts 13.3% of the time on same source pairs, therefore making incorrect predictions more often than correct predictions. The method does, however, predict correctly on different source pairs 99.8% of the time. This suggests that the method makes little distinction between same and different source pairs.

		Truth	
		Same Source	Different Source
Prediction	Match	0.133	0.002
	Non Match	0.867	0.998

Table 5.1: Median confusion matrix of match predictions for fragments using the multivariate normal log likelihood ratio applied to USA ribbon data. We see that this criterion has 13.3% median accuracy on same source pairs, and 99.8% median accuracy on same source pairs.

Table 5.2 shows the confusion matrix for the multivariate kernel method applied to the USA ribbon data. In this case, we see improvement over the MVN procedure in terms of predicting on same source pairs, now predicting correctly 41.9% of the time, an increase of 28.6 percentage points, but still predicting correctly less often than incorrectly. We also see a very small decrease in the prediction accuracy on different source pairs, this method correctly predicting 99.1% of the time.

5.1.2 Australian Casework Data

We now move on to the more diverse Australian casework data set. For each model, the data was split into a background training set and a testing set. When this procedure was repeated with the data randomly distributed into each of these sets, it was found that the results varieried much more than was the case for the USA data. This is likely due to the highly variable

		Truth	
		Same Source	Different Source
Prediction	Match	0.419	0.009
	Non Match	0.581	0.991

Table 5.2: Median confusion matrix of match predictions for fragments using the multivariate normal log likelihood ratio applied to USA ribbon data. We see that this criterion has 41.9% median accuracy on same source pairs, and 99.1% median accuracy on same source pairs.

nature of the data set, with several separate clusters observed in several of the elements. This variability may reduce when a larger data base can be obtained to inform the background data set. To address this, the procedure was repeated 50 times for each model, rather than ten. We again note that the results were skewed in some cased, and present the median of the 50 repeats for the model fit metrics.

Table 5.3 displays the confusion matrix for the multivariate normal likelihood ratio procedure applied to the Australian casework. We note that this method predicts correctly on same source pairs 58.3% of the time, but performs much better on different source pairs, predicting correctly 100% of the time.

		Truth	
		Same Source	Different Source
Prediction	Match	0.583	0.000
	Non Match	0.417	1.000

Table 5.3: Median confusion matrix of match predictions for fragments using the multivariate normal log likelihood ratio applied to Australian casework data. We see that this criterion has 58.3% median accuracy on same source pairs, and 100% median accuracy on same source pairs.

Table 5.4 displays the confusion matrix for the multivariate kernel likelihood ratio procedure applied to the Australian casework. We note that this method improves significantly over the MVN procedure in terms of prediction on same source pairs. This method predicts correctly 83.3% of the time in this case. We note also that it performs almost identically on different source pairs, predicting correctly 100% of the time.

		Truth	
		Same Source	Different Source
Prediction	Match	0.833	0.000
	Non Match	0.167	1.000

Table 5.4: Median confusion matrix of match predictions for fragments using the multivariate kernel log likelihood ratio applied to Australian casework data. We see that this criterion has 83.3% median accuracy on same source pairs, and 100% median accuracy on same source pairs.

5.2 Calibration and Transformation Results

In the last section, we assessed how each LR system performed as a binary classifier. Now, we assess the calibration of the LR systems and in doing so evaluate the extent to which LR results can be interpreted as strength of evidence. We compare the classification performance of the systems with a variety of critical values, and find the min-trade-off and max-mean optimum values. We then apply the min-trade-off and max-mean transformations to the LR systems, and assess the calibration before and after these transformations have been applied. Having seen in Section 5.1 that the MVK procedure outperformed the MVN method, throughout this section we use only the MVK procedure.

5.2.1 USA Ribbon Data

In Section 5.1 it was found that there was only a small amount of variability in the results when the LR techniques were applied using different subsets of the USA ribbon data for background and testing. As a result, in this section we proceed only with the results from a single run of the LR procedure.

Figure 5.1 shows the sensitivity and specificity of the MVK procedure applied to the USA ribbon data at varied critical values. We see that trade-off is minimised at -74, and the mean is maximised at -99.

Figure 5.2 shows the ECE plots for the MVK procedure applied to the USA data. We see in Figure 5.2a that when no adjustment is applied, the observed ECE curve lies well above the null curve, and shows poor calibration with a C_{llr} of 20.103. The calibration is substantially improved when transformations are appleid, with both transformations bringing the observed ECE curve well below the null curve (Figures 5.2b and 5.2c). The two transformations appear to perform almost identically, with the min-trade-off achieving a C_{llr} of 0.095, and the max-mean transformation a C_{llr} of 0.089.

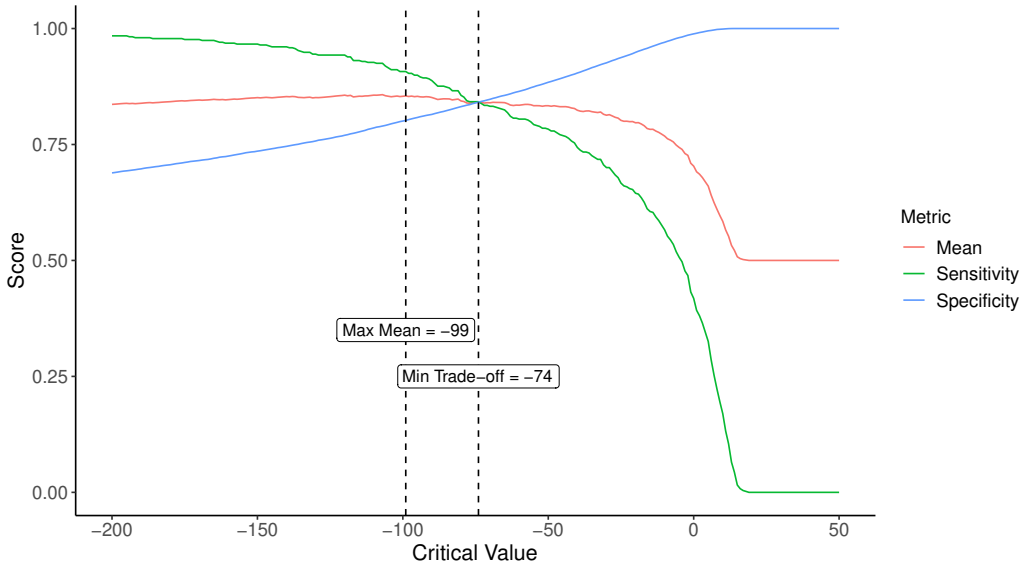
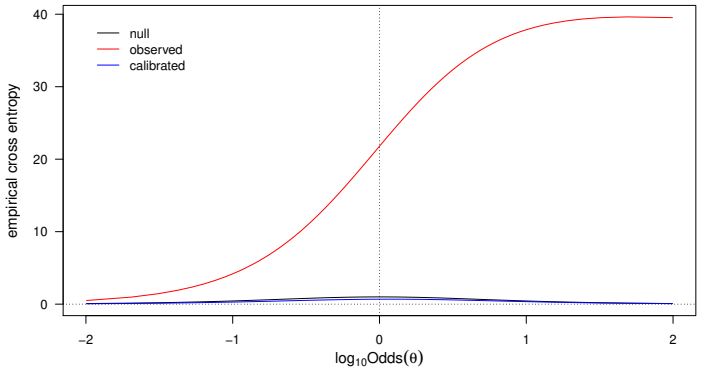


Figure 5.1: Plot of sensitivity, specificity and the mean of the two metrics for the MVK LR procedure applied to the USA ribbon data, with varying critical values. The dotted vertical lines show the critical values at which the mean of the metrics is maximised, and the trade-off between sensitivity and specificity is minimised.

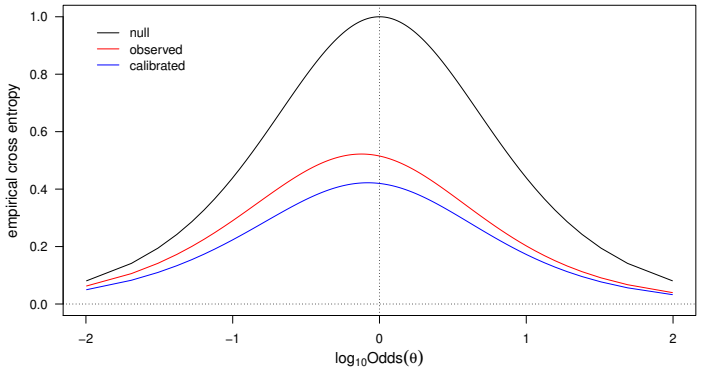
Table 5.5 shows the classification performance metrics for the MVK methods applied to USA data with and without transformations applied. We see that the transformed LR systems compromised in accuracy and Kappa in order to achieve a better balance between sensitivity and specificity. Without adjustments, the model was able to predict near perfectly on different source pairs, but correctly less than half of the time on same source pairs. In order to make sensitivity and specificity approximately equal, they were both reduced to 0.84 in the min-trade-off transformation. In the max-mean transformation however, sensitivity was favoured, with the model predicting correctly 90% of the time on same source pairs, and 80% of the time on different source pairs.

Model	Accuracy	Kappa	Sensitivity	Specificity	C_{llr}
Default	0.983	0.315	0.418	0.989	20.103
Min-trade-off	0.841	0.077	0.842	0.841	0.095
Max-mean	0.803	0.065	0.907	0.802	0.089

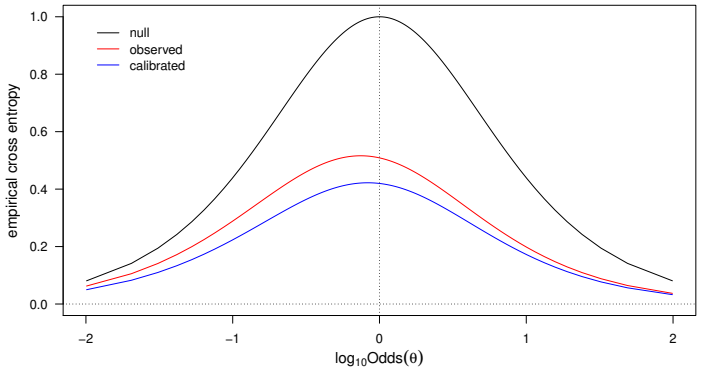
Table 5.5: Performance metrics for multivariate kernel density LR procedure applied to USA ribbon data before and after transformations.



(a) MVK LR with no transformation. $C_{llr} = 20.103$



(b) MVK LR with min trade-off transformation. $C_{llr} = 0.095$



(c) MVK LR with max mean transformation. $C_{llr} = 0.089$

Figure 5.2: ECE plots of MVK LRs for USA ribbon data with and without transformations applied. We see in (a) that the LR system with no transformation applied contain misleading evidence in favour of same source predictions. The transformed LR systems show better levels of calibration, each having C_{llr} values less than 0.8, compared to 20 for the untransformed LR system. The min trade-off transformation was calibrated best with a C_{llr} of 0.715.

5.2.2 Australian Casework Data

Recall that in Section 5.1, for the Australian casework, the LR procedures were repeated 50 times with random samples of the data set used as the background training set, and as the testing set. The performance was evaluated separately for each of these 50 repeats, and the median of the performance metrics were presented. Having evaluated this variability, in this section, in order to obtain a single, reliable estimate, we aggregate the LR calculations from each of these 50 repeats, to obtain a larger set of LR values. This process can be thought of as similar to the process of bootstrap aggregation. Throughout this section, the analysis is conducted on this aggregated set of LR values.

Figure 5.3 shows the metric scores a varying critical values for the MVK LLR procedure. In this case, we note that the curves are very flat the the performance varies only slightly at largely different critical values. We find $\alpha_{mt} = -2140$ and $\alpha_{mm} = -685$.

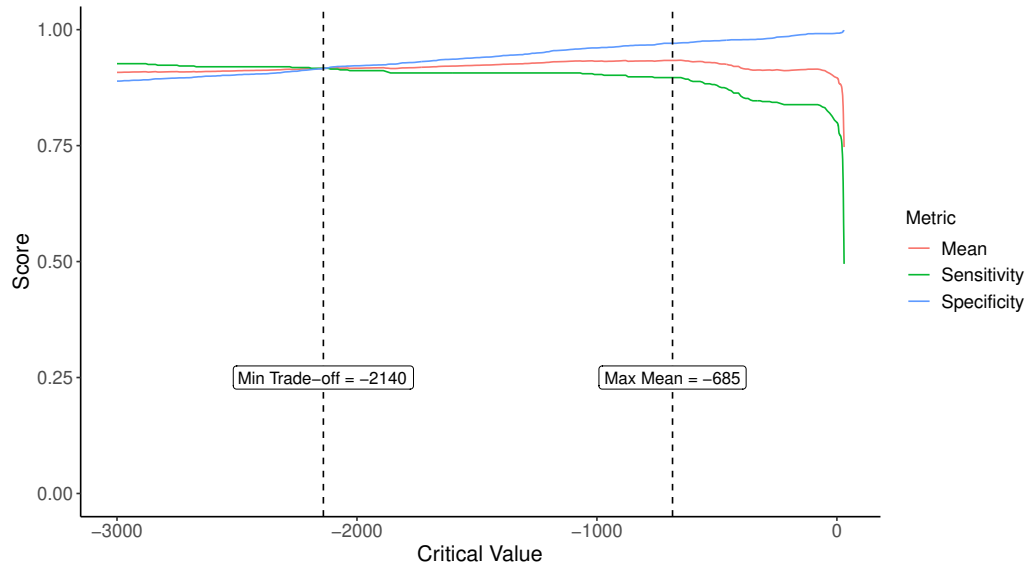


Figure 5.3: Plot of sensitivity, specificity and the mean of the two metrics for the MVK LR procedure applied to the Australian casework data, with varying critical values. The dotted vertical lines show the critical values at which the mean of the metrics is maximised, and the trade-off between sensitivity and specificity is minimised.

Figure 5.4 shows the ECE plots for the MVK procedure applied to Australian casework data with the two transformations applied. The ECE plot for the untransformed LLR system could not be produced, as infinite val-

ues were yielded in the calculation of the ECE. This, in itself, suggests poor calibration of the original system. in Figure 5.4a we note that when the min-trade-off transformation is applied, the resulting LR system has an ECE curve which sits entirely underneath the null curve, and results in a C_{llr} of 0.207. When the max-mean transformation is applied, the system is less well-calibrated, and the ECE curve is not entirely underneath the null curve. In this case, the C_{llr} is 0.504, more than double that after then min-trade-off transformation. Since a C_{llr} could not be determined for the unadjusted LLR system, it is difficult to establish how much of an improvement each of these methods give, though it is clear that the min-trade-off transformation performs best.

Table 5.6 shows the performance metrics for the MVk procedure before and after transformations. We note that in terms of overall accuracy, Cohen's kappa and specificity, the default, untransformed method performed best. The two transformations led to smaller compromises between sensitivity and specificity. In particular, the max-mean transformation reduced specificity by 0.021, allowing for an increase of nearly 0.1 in sensitivity compared to the default model.

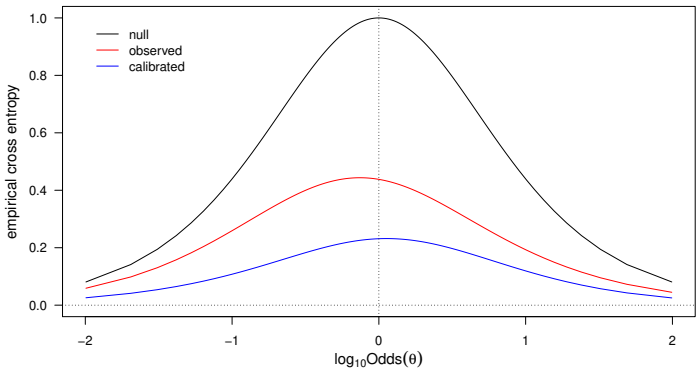
Model	Accuracy	Kappa	Sensitivity	Specificity	C_{llr}
Default	0.976	0.837	0.800	0.992	∞
Min-trade-off	0.916	0.604	0.917	0.916	0.207
Max-mean	0.964	0.788	0.897	0.971	0.504

Table 5.6: Performance metrics for multivariate kernel density LR procedure applied to Australian casework data before and after transformations.

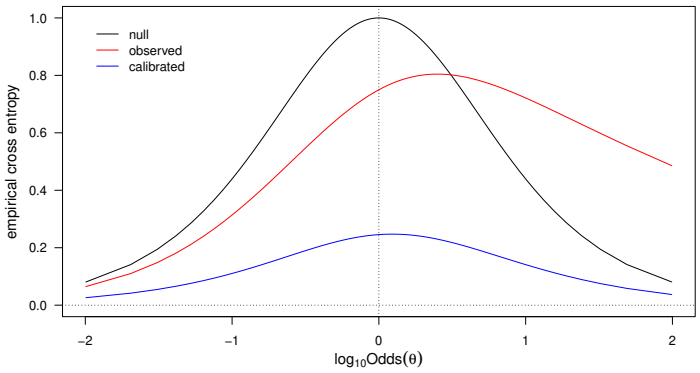
The interested reader can find a visual comparison of the distributions of the LR systems pre- and post-calibration in Appendix B

5.3 Summary

In this chapter we have evaluated the use of likelihood ratio-based techniques to make comparison between forensic glass samples. We have assessed these methods first viewed simply for the purpose of binary classification, and then as a method which allows one to quantify the strength of forensic evidence. We began by comparing two multivariate density-based methods, which are well-established in the literature, to calculate LRs from elemental composition measurements. For each method, we assessed their classification performance using zero as a critical values for log-likelihood ratios. We



(a) MVK LR with min-trade-off transformation. $C_{llr} = 0.207$



(b) MVK LR with max-mean transformation. $C_{llr} = 0.504$

Figure 5.4: ECE plots of MVK LRs for Australian casework data with and without transformations applied. We see that the min-trade-off transformation yields the best calibrated system of LLRs, while in the case of the max-mean transformation the ECE curve lies above the null curve at higher log odds.

then investigated choosing a critical value which optimises classification performance, and applying a transformation to the LLR system based on this value. The calibration of the LR systems was then also assessed before and after these transformations using empirical cross entropy curves.

The USA ribbon data provides insight into the performance of the MVK method on a homogeneous data set in which there is little variation between sources. Without transformation, the MVK procedure predicted near-perfectly on different source pairs, and correctly less than 50% of the time on same source pairs. In this case, the calibration was very poor with a C_{llr} of over 20. The C_{llr} was reduced to less than 0.1 by both transformation methods, which both also raised sensitivity to above 84%–90% in the case of the max-mean transformation. In doing so, however, the specificity was reduced to less than 85% in both cases. These results suggest that the MVK LR method was unable to provide good discrimination between the same and different source pairs in this data set, as when sensitivity and specificity were both maximised, the result was 84% prediction accuracy.

On the more diverse Australian casework data it was found that the multivariate normal and multivariate kernel procedures predicted correctly 99% of the time on different source pairs, but less effectively on same source pairs, with the MVN procedure correctly predicting same source pairs less than half of the time. The MVK procedure, meanwhile, predicted correctly 80% of the time.

In the assessment of calibration, it was found that the MVK procedure was very poorly calibrated, but that this was greatly improved when transformations were applied. The LR system was best calibrated with the min-trade-off adjustment, which led to correct predictions 91% of the time on all pairs. The max-mean transformation resulted in a system which was not calibrated as well, but had 97% correct prediction on different source pairs and 90% correct prediction on same source pairs.

With these results on mind, one must make a decision about the compromise between binary classification performance, and calibration. The max-mean transformations applied to the MVK procedure resulted in what can subjectively be described as the best compromise, as specificity was favoured, resulting in the lowest false positive rate.

In terms of prediction accuracy, these methods have performed less accurately than the ellipsoid criterion described and evaluated in Chapter 3 which achieved sensitivity and specificity of approximately 0.99 on the Australian casework data. The LR methods do, however, have the benefit of quantifying the strength of the evidence in favour of either hypothesis. In

the next chapter, we investigate the use of machine learning classifiers to take full advantage of a background database and the mutilvariate structure of the data and aim to further improve predictive accuracy.

Part III

Machine Learning Classification

Part III Glossary

Terminology

Term	Meaning
Class	Variable denoting the classification of an instance.
Edge	Line connecting nodes in a graph. See Appendix A.3
Entropy	In information theory, the entropy of a random variable is the average level of uncertainty contained in the variables potential outcomes. The entropy H of a discrete random variable $\mathbf{X}$ with outcomes $\mathbf{x}_i$ is given by $H(\mathbf{X}) = -\sum_i P(x_i) \log_2 P(x_i)$.
Feature/Attribute	Any variable within a data set.
Instance	Data point/observation in the training data.
Node	Node of a graph. See Appendix A.3
Testing data	Data set used to test the predictive performance of a machine learning model.
Training Data	Data set used to train a machine learning model. The model learns the how to make predictions based on the training data.

Abbreviations

Abbreviation	Meaning
AUC	Area under the curve
C_{llr}	Cost Log-Likelihood Ratio
DT	Decision Tree
ECE	Empirical Cross Entropy
KDE	Kernel Density Estimate/Estimation
LR	Logistic regression
ML	Machine Learning
RF	Random Forest
ROC	Receiver Operating Characteristic
SLR	Score-Based Likelihood Ratio
SMOTE	Synthetic Minority Oversampling Technique

Chapter 6

Machine Learning Methodology

We have now seen that the likelihood ratio methodologies improve upon the current practice by allowing for the strength of evidence to be quantified, but fall short in their performance as binary classifiers. In this part of the thesis, we discuss the use of some machine learning classification techniques for predicting whether pairs of glass samples match. The work is largely motivated by some recent work conducted by Park and Carriquiry (2019) who investigate the use of random forests, and Bayesian Additive Regression Trees (BART) for this purpose, and compare the results with the more traditional interval-based approaches. Following this work, we begin with logistic regression, and then increase the model complexity by applying decision tree models, and random forests.

In this chapter we introduce the theory involved in machine learning classification. We first describe mathematically how the three models are constructed, and present the algorithms which the models use to make classification. We then also discuss the preparatory steps which must be applied to the data – specifically addressing imbalance in the number of same source and different source pairs. We describe three resampling methods to do this: downsampling the majority class, upsampling the minority class with replacement, and synthetic minority oversampling examples (SMOTE).

Prior Work in This Area

Work began only recently investigating the use of machine learning algorithms for the evaluation of elemental forensic glass evidence when Park and Carriquiry (2019) investigated two classification models for this pur-

pose: random forests and Bayesian Additive Regression Trees (BART). The authors challenge the currently employed techniques of using univariate intervals around mean values of concentrations of elements to compare glass fragments in that these techniques do not account for dependencies between concentrations of different elements. In this setting, rejecting the null hypothesis of no difference between two samples becomes less likely when the variance of concentrations of elements is larger. In turn, this leads to favouring the prosecution, as the intervals for classifying a match would in general be wider. The authors suggest a score-based approach in which RF and BART methods are used to determine a similarity score between pairs of glass fragments. The similarity score is the class probability that is predicted when using a ML classification model, and a score of 0.5 or more is classified as a match. These methods make use of dependencies between certain elemental concentrations to establish estimated probabilities to give what they refer to as a degree of similarity between fragments to report to authorities in the legal setting.

Park and Carriquiry (2019) demonstrated primarily that this approach is superior to the univariate hypothesis test based approaches insofar as minimising classification error between fragments. They used a combination of three data sets to test their methodologies, one of which we have gained access to – the USA ribbon data set. We will use this data set to validate their findings, and compare the performance with the same models applied to the Australian casework data.

6.1 Machine Learning Classifiers

Recall that we are considering statistical classification models to predict whether pairs of glass fragment are matching. In this chapter, we consider machine learning classifiers which predict an empirical probability that each pair of fragments is matching or not matching. If this predicted probability is greater than 0.5 in favour of a match, the pair is classified as matching.

We consider three different classification models, each slightly more advanced and complex than the last, to predict whether glass samples match. We begin with logistic regression, and then consider a decision tree model, and finally a random forest, which is an extension of the decision tree model. The choice of these three models is in part motivated by their easy to interpret nature in comparison to the full suite of machine learning classifiers available. Many machine learning algorithms suffer from being "black boxes": while the algorithms themselves are known, it is often a mystery as to the

specific criteria by which they make classification. This can present problems when a model makes incorrect classifications, as one cannot necessarily determine why the decision was made. Even in the case of consistent and accurate predictions, many are sceptical of models which they cannot explain (Pearl, 2019, Pearl and Mackenzie, 2020). This is of particular importance in the forensic setting, when predictions are being made on evidence which is to be presented in court, and one needs to argue the validity of this evidence. The three models here are less vulnerable to this criticism, and are more easily explained. Logistic regression follows a set of clear mathematical equations and decision trees have the advantage of mimicking the human decision-making process (James *et al.*, 2017) – a point in favour of their interpretability.

6.1.1 Logistic Regression

In order to make comparison with the work done by Park and Carriquiry (2019) on random forests and Bayesian additive regression trees, we chose to start with a fully interpretable, logistic regression model. Recall that logistic regression is a generalisation of linear regression which predicts an odds ratio of observations belonging to each class, which is then transformed to a probability. Observations are then assigned to the class for which their predicted probability is highest. In the case of binary classification, the class which receives a probability greater than 0.5. It is also important to note that the logistic regression model is sensitive to scale and location of variables. As such, for this model, the data were scaled by standard deviation and centred about zero before fitting the model. A more detailed recap of logistic regression can be found in Appendix A.1.

6.1.2 Decision Trees

Decision trees are popular as they mimic the natural human decision-making process. A decision tree can be thought of as a flow chart, or network which has a tree structure. This means that it has one root node which branches off into child nodes, which can branch off further, until they terminate at the final classification. The root node, and all other internal nodes are referred to as decision nodes. The nodes at the end of the tree, from which there are no further branches, are known as leaf nodes. Decision nodes represent features within the data, and branches represent classification decisions based on these features' values. The leaf nodes then represent the classification outcomes based on the decisions that came before it. The decision tree learns to partition based on variable values in the training data set, and continues to

partition recursively. As such, one can easily visualise a decision tree and see how it replicates the decision making procedure of humans. Figure 6.1 depicts a simple example of a decision tree classifying observations two predictor variables A and B.

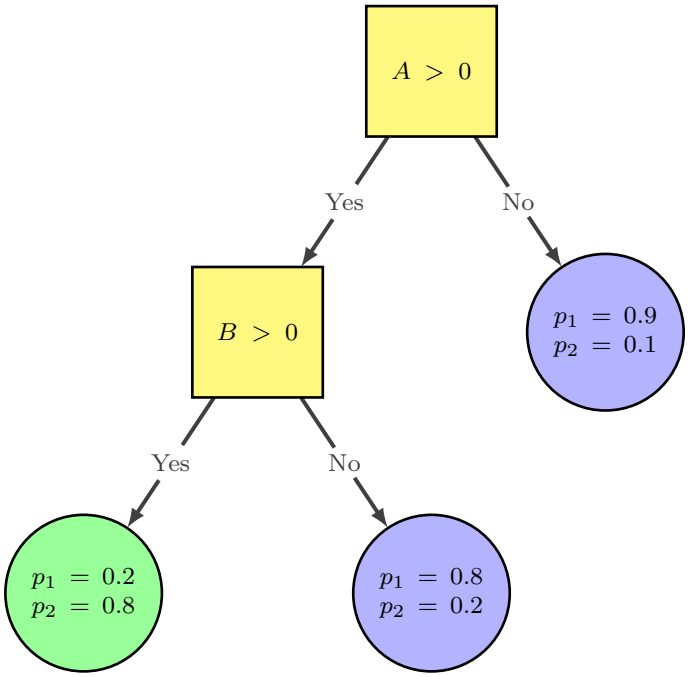


Figure 6.1: Example of a decision tree for a binary classification problem with two real-valued variables A and B. The decision nodes are represented by yellow squares, and the leaf nodes as green or blue circles. At the root node, the tree first checks whether A is positive. If not, the observation is predicted to have class 1 with probability 0.9. If A is positive, the tree then checks if B is positive and predicts class probabilities based on the result.

Decision trees also do not suffer from the property of being like a "black box", as the logic used at each branch of the decision tree is accessible to the user, and so the decision-making process can be better understood, and issues in the model can be diagnosed. A key benefit of decision trees over logistic regression, is that they do not rely on distributional assumptions of the dataset, but rather are non-parametric (Rounds, 1980) and they are able to handle high-dimensional data well, without a significant loss of accuracy.

The algorithms by which decision trees are constructed rely on attribute selection measures (ASMs). Two major ASMs used in decision tree models: Gini impurity and information gain, both of which are available in the scikit-

learn package in Python (Pedregosa *et al.*, 2011).

Gini Impurity

Gini impurity measures the probability that a new observation is incorrectly classified, if it was classified at random according to the distribution of class labels in the data. The Gini impurity G, for classification of a data set $\mathcal{D}$ into K classes, is defined as

$$G(\mathcal{D}) = \sum_{i=1}^{K} p_i(1 - p_i),$$

where p_i is the proportion of instances of class i in the training data. In a binary classification problem, we have only p_1 and p_2, where $p_1 = 1 - p_2$. Therefore, the Gini impurity is simply given by

$$G(\mathcal{D}) = p_1(1 - p_1) + p_2(1 - p_2) = 2p_1p_2.$$

To choose an attribute to split the data, we can consider the Gini impurity for a binary split by each potential attribute. Suppose an attribute A partitions $\mathcal{D}$ into $\mathcal{D}_1$ and $\mathcal{D}_2$. The Gini impurity of this split by attribute A, is then given by

$$G_A(\mathcal{D}) = \frac{|\mathcal{D}_1|}{|D|}G(\mathcal{D}_1) + \frac{|\mathcal{D}_2|}{|\mathcal{D}|}G(\mathcal{D}_2).$$

For continuous-valued attributes, as is the case for all variables in the glass datasets, the decision tree model will iterate through all pairs of adjacent values as a potential splitting point, and the point with the lowest Gini impurity is chosen. This is then repeated with all potential attributes, and again, the attribute which minimises G is chosen.

Information Gain

Next, information gain measures the decrease in entropy, or equivalently the gain in information, when a split is made in the data set. Entropy can be thought of as the level of uncertainty in the possible outcomes for a given variable. To compute information gain, we first define the information, or entropy H of the data $\mathcal{D}$, again with K classes, to be

$$H(\mathcal{D}) = -\sum_{i=1}^{K} p_i \log_2 p_i.$$

The entropy associated with a split by attribute A, $H(\mathcal{D} \mid A)$, is given by

$$H(\mathcal{D} \mid A) = -\sum_{a \in A} P(a) \sum_{i=1}^{K} P(i \mid a) \log_2 P(i \mid a),$$

where the $a \in A$ are possible values of attribute A. Now, the information gain IG, is given by

$$IG_A(\mathcal{D}) = H(\mathcal{D}) - H(\mathcal{D} \mid A).$$

Converse to Gini impurity, we seek to maximise information gain, and so the attribute which maximises it is chosen for the decision node. The algorithm for decision trees, known as C4.5, is given in Algorithm 6.1.

Algorithm 6.1: Decision Tree Algorithm (C4.5)

Data : Attribute List $\mathcal{A}$, Training data $\mathcal{D}$
Result: Decision Tree Model

1 **if** $\mathcal{D}$ *contains only one class OR* $\mathcal{A} = \emptyset$ **then**
2 | Create leaf node
3 | **return**
4 **end**
5 **foreach** $A \in \mathcal{A}$ **do**
6 | Calculate $IG_A(\mathcal{D})$
7 **end**
8 Set $A_{\text{best}} \leftarrow \text{argmax}_{A \in \mathcal{A}} IG_A(\mathcal{D})$ and create decision node
9 Set $\mathcal{A} \leftarrow \mathcal{A} \setminus \{A_{\text{best}}\}$
10 Partition $\mathcal{D}$ into $\{\mathcal{D}_1, \ldots, \mathcal{D}_n\}$ according to A_{best}
11 **foreach** $\mathcal{D}_i$, $i = 1, \ldots, n$ **do**
12 | Create child node N_i
13 | Recurse with inputs $\mathcal{A}, \mathcal{D}_i$
14 **end**

This algorithm is written using the information gain attribute selection measure. If one was to use Gini impurity, line six would be changed to

$$\text{Calculate } G_A(\mathcal{D}),$$

and line eight would become

$$\text{Set } A_{\text{best}} \leftarrow \text{argmin}_{A \in \mathcal{A}} G_A(\mathcal{D}).$$

6.1.3 Random Forests

Random forests are an example of a type of model known as an ensemble learning model. Such models combine multiple machine learning models to improve upon the results that would be obtained if the models were to be used individually. In the case of random forest, an ensemble of decision tree models is created, using bootstrap aggregation, also known as bagging. The random forest method should increase performance of decision trees by reducing variance without increasing bias, as a single decision tree might be sensitive to variance within the training data, but this should be reduced when taking the average of a number of trees, each trained on a random sample of the data.

Bootstrap aggregation is preformed as follows: consider some training data $\mathcal{D} = (X, Y)$, where $X = (x_1, \ldots, x_n)$ are the observed data and $Y = (y_1, \ldots, y_n)$ are the observed responses, or classes. The algorithm then proceeds with the following steps:

- Sample a set of observations at random from the training set., with replacement.
- Sample a random subset of the features.
- Fit a decision tree to this sample of observations, using only the sampled features.
- Repeat this process B times, on samples $\mathcal{D}_b = (X_b, Y_b)$ for $b = 1, \ldots, B$.
- Make classification by taking the class into which the majority of the trees classify a sample.

Note that each time a tree is fit, it uses only a sample of the features, rather than the whole set. This method is used to reduce the chance of significant correlation between the trees in the forest, as if a select few features carry a lot of weight in prediction, they will be selected in the majority of the trees. Where a decision tree can be thought of as analogous to a single human decision maker, a random forest model takes the majority vote of a number of decision makers. The number of attributes sampled is considered a hyperparameter of the model. It is often set to be the square root of the total number of attributes, but we chose to tune the value in our hyperparameter selection process.

6.2 Data Preparation

In order to make use of the models described in the previous section, some preparatory steps must be applied to the data. Recall from Section 6.2 that for the purposes of classification, our data are considered in pairs. That is, the observations are represented in the form:

$$\{(\boldsymbol{x}_{i,j}, \boldsymbol{x}_{i',j'}) \,:\, i \neq i' \text{ or } j \neq j'\},$$

and are classed as KM (known mate) or KNM (known non-mate) if the samples $\boldsymbol{x}_{i,j}$ $\boldsymbol{x}_{i',j'}$ are, or are not from the same source respectively. In order for a machine learning model to make sense of this data, it must be transformed into a single observation, rather than a pair. Keep in mind, the single observation is still multivariate in the sense that it has the observations of 18 different chemical isotopes. For each element k, the transformed observation is given by $x_{i,j,k} - x_{i',j',k}$. We refer to this as differencing a pair of observations, and the data in this form will be referred to as *differenced data*. So that a sufficiently large data set could be created for use with the machine learning models, this process of differencing was applied to each individual observation in the data set, rather than first taking an average over the observations in each glass source, and differencing pairs of averages.

Recall also from Section 6.1.2 that for logistic regression another step of preparation is required. This model is sensitive to scale and location of variables, and so the data were scaled by standard deviation and centred about zero. This step was performed for logistic regression only.

Finally, the data are split into training and testing groups, with 75% of the data allocation for training the models. The models are then applied to the remaining 25% of the data – the testing set – and their performance is then measured using the metrics described in Section 2.3.3.

6.2.1 Class Imbalance

When differencing data where the number of groups is large, and the number of observations is relatively small, one is left with many more data points in the non-matching group than in the matching group. The following explanation describes the exact imbalance if the entire dataset were different, but the result holds true in our case since samples were taken uniformly at random. Consider data in N groups with n_i observations in group i, for $i = 1, \ldots, N$.

Then, group i contributes $\binom{n_i}{2}$ matches, totalling to

$$\sum_{i=1}^{N} \binom{n_i}{2} = \sum_{i=1}^{N} \frac{n_i(n_i-1)}{2}.$$

matched observations in the differenced data set. Then, in Group i, each observation contributes one non-match from every observation in all of the other groups, of which there are

$$\sum_{\substack{j=1 \\ i \neq j}}^{N} n_j,$$

for each i. There are then a total of $\sum_{i=1}^{N} n_i$ observations in the original data set, and so the total number of non-matches in the differenced data is

$$\sum_{i=1}^{N} \sum_{\substack{j=1 \\ i \neq j}}^{N} n_i n_j.$$

For both of the data sets which we consider in this section, after the process of differencing, there were many more different source observations than same source. In order to address this imbalance in the representation of each class, a resampling technique should be applied to the training data.

Park and Carriquiry (2019) presented four resampling methods to pre-process the data, and the results were compared for each of these methods. They pre-processed by downsampling the majority class, upsampling the minority class with replacement, synthetic minority over-sampling technique (SMOTE), and random over-sampling examples (ROSE).

Standard Downsampling and Upsampling

Downsampling (or under-sampling) the majority class involves sampling without replacement from the majority class until the number of samples obtained is the same as the number of samples in the minority class. If samples are taken uniformly at random, this method should provide an unbiased estimate of the original data. Upsampling (or over-sampling) the minority class with replacement involves taking the original minority class, and then sampling again from this class, with replacement, until the number of samples is equal to the number of samples of the majority class.

SMOTE

Park and Carriquiry (2019) then also discussed two more advanced techniques: synthetic minority oversampling technique (SMOTE) (Chawla *et al.*, 2002) and random over-sampling examples (ROSE) (Lunardon *et al.*, 2014), both of which generate artificial samples in the minority class. Park and Carriquiry (2019) found that using SMOTE led to the best performance of the random forest and BART models, which we will seek to verify for the RF model. They also found that applying ROSE to their training data set yielded the worst model performance overall, compromising raw accuracy and specificity in favour of sensitivity. We investigated the ROSE algorithm early on, but found that the same was true when applied to the Australian casework and USA ribbon data sets and so we did not pursue the method any further.

The SMOTE algorithm makes use of the k-nearest neighbours algorithm, and generates synthetic samples for each observation in between it, and its k nearest neighbours (Chawla *et al.*, 2002). The authors motivate the algorithm by stating that oversampling the minority class, with replacement, does not improve the recognition of minority classes, but instead identifies more and more specific sets of feature values to make decisions. That is, it narrows the choice of feature values by which it classifies observations. The SMOTE algorithm starts by drawing straight lines between each observation in the minority class, and its k nearest neighbours. Synthetic samples are then generated at a random point along each of these lines. Depending on the number of samples needed, some or all of the k neighbours are chosen for this process, and if more samples are required, neighbours of neighbours are used also. (Chawla *et al.*, 2002) suggest using $k = 5$. In theory, this methods should have the opposite of the effect the authors described for oversampling, in that the SMOTE algorithm should generalise and widen the choice of feature values for making classification. Figure 6.2 provides a visualisation of the process in two variables and the pseudo-code for the algorithm is provided in Appendix C.

In the next chapter, we move on to applying these techniques first one the USA ribbon data set – in order to validate the work of Park and Carriquiry (2019) – and then the Australian casework data set as well. We will then evaluate whether these methods offer an improvement over the current practice and likelihood ratio procedures.

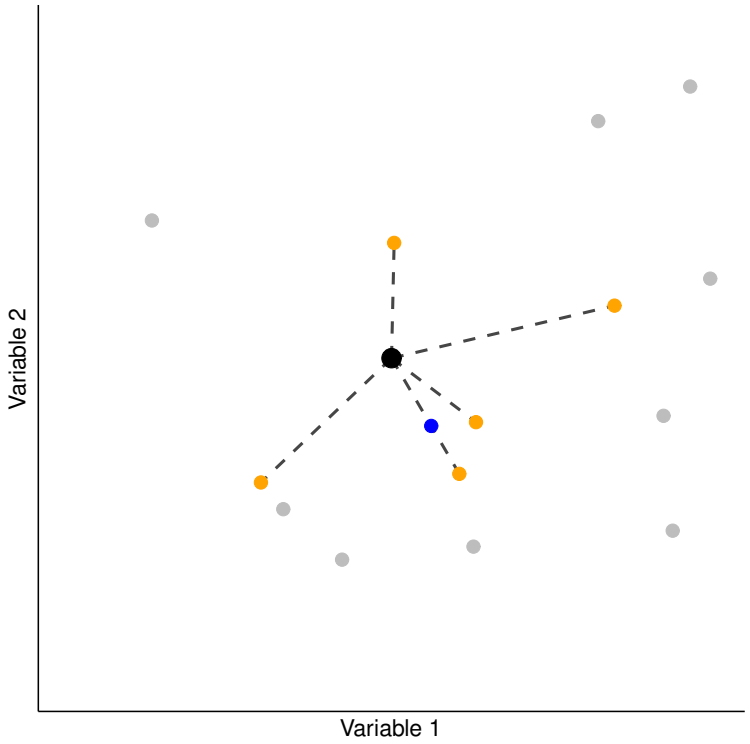


Figure 6.2: Visualisation of SMOTE sampling procedure. The black dot in the centre represents sample in original data from which to generate a new sample. The orange dots are its five nearest neighbours, and the dotted lines show the paths between the sample and its neighbours. One of the nearest neighbours is chosen at random, and the blue dot is the new sample which has been generated uniformly at random along the dotted path. The other grey dots are other observations in the data which are further from the black observation than the five nearest neighbours.

Chapter 7

Machine Learning Results

In this chapter we present the results of the machine learning models described in the previous chapter applied to USA ribbon data and the Australian casework. We aim to improve upon the classification performance offered by the ellipsoid criterion and the likelihood ratio procedures.

We begin with the USA ribbon data, in order to validate the methods employed by Park and Carriquiry (2019), before applying the procedure to the Australian casework. We first provide an assessment of the three resampling techniques: downsampling, upsampling and SMOTE, to establish which method yields the best performance. Then, we assess which of the three machine learning models: logistic regression, decision tree and random forest, performs best on each data set. For both data sets, we nominate the decision tree as the best choice of model. We find that upsampling the data and then using a random forest provides the highest scores across the classification metrics, but that the random forest offers only a small increase in performance over decision trees, at the cost of transparency in the way that the model makes decisions, as well as computational complexity.

Figure 7.1 provides a graphic visualisation of the key stages of the workflow of preprocessing the data and fitting different models. Starting with the original dataset on the left, the data is differenced, and the differenced data is resampled to balance the match and non-match classes. After this, the three classification models are trained and tested. The blue path through the graph shows the resampling method and model that were selected as performing best on the Australian casework data.

Finally, we make an assessment of how well the models trained on one data set can generalise to predicting on another, To do this, we train the decision tree and random forest models on each of the data sets, and use them to predict on the other. We find that the models perform poorly when

applied to the other data set, particularly when predicting on different source sample pairs. In this case, the decision tree models distinguished themselves from the individual decision trees, but the performance was still too poor for the models to be used in practice. This result supports the idea that relevant background population databases should be constructed to train models for classification in a given location.

The ML methods in this chapter were implemented using the scikit-learn package in Python (Pedregosa *et al.*, 2011).

7.1 Comparison of Resampling Techniques

In this section we make a comparison of the resampling techniques by assessing the model fit metrics for each of the three models described in Section 6.1. The downsampling, upsampling with replacement, and SMOTE techniques were applied to both data sets, and then the logistic regression, decision tree and random forest models were fit to each version of the data. In doing so, we are able to summarise the difference in the performance of the three models when trained on differently pre-processed data. To make these comparisons easier to understand and interpret, we have elected to present the areas under the ROC curves for each resampling method applied to each model. As discussed in Section 2.3.3, the ROC AUC gives a good overall statistic of model fit, and for these models, the curves themselves provide little extra information and the AUC itself is sufficient. We have then chosen to include the accuracy, Cohen's kappa coefficient, sensitivity and specificity for the three resampling methods when the random forest model is fit.

7.1.1 USA Ribbon Data

For the USA data, we note in Table 7.1 that upsampling and SMOTE perform best, with SMOTE only outperforming upsampling in the logistic regression case. SMOTE scores highest for logistic regression, while upsampling performs best for the decision tree, and the results are equal for the random forest. Downsampling performed worst overall, with the most notable distinction in the decision tree models, where downsampling the majority class resulted in an ROC AUC 0.106 less than the value for SMOTE.

In Table 7.2 we see that, for the random forest models, there was little difference between upsampling and SMOTE across the four metrics, particularly specificity, in which they were equal. In accuracy, Cohen's kappa and sensitivity, upsetting performed best. Again, downsampling performed worst

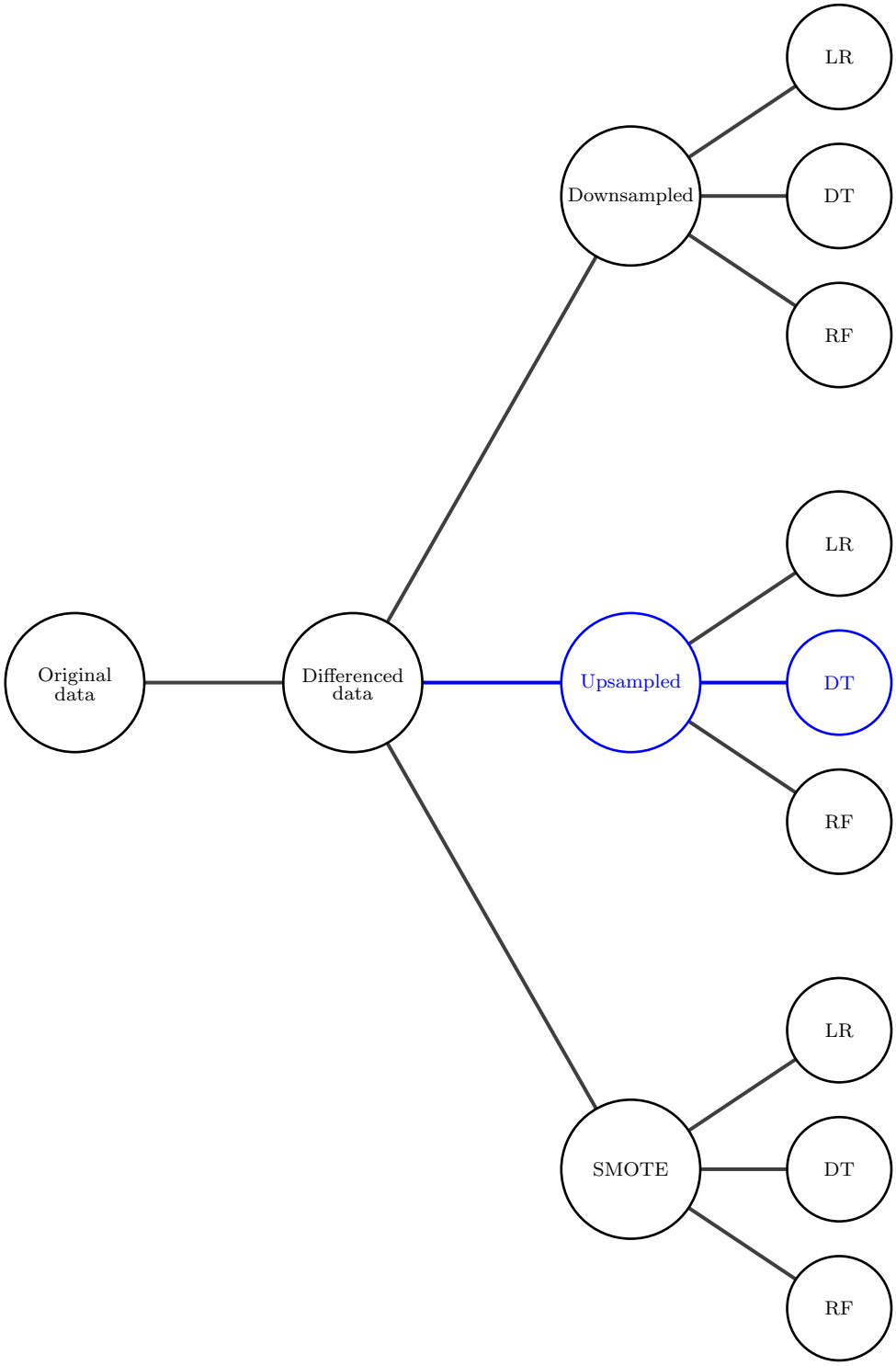


Figure 7.1: Visual Summary of machine learning workflow. The blue path shows the resampling procedure and model which were ultimately chosen for both the USA ribbon data and the Australian casework.

Resampling Method	Logistic Regression	Decision Tree	Random Forest
Upsampled	0.858	0.994	1.000
SMOTE	0.862	0.977	1.000
Downsampled	0.823	0.871	0.980

Table 7.1: ROC areas under the curve for three resampling techniques applied to the three ML models trained and evaluated on USA ribbon data. We see that, for each ML model, there is at most a difference of 0.017 between AUC for upsampled and SMOTE resampled data. SMOTE scores higher for logistic regression, upsampling higher for the decision tree, and the results are equal for the random forest. Downsampling performs worst out of the three, particularly in the logistic regression and decision tree models.

in all metrics, with the greatest difference in Cohen's Kappa, suggesting the least improvement over a uniform random classifier.

Resampling Method	Accuracy	Kappa	Sensitivity	Specificity
Upsampled	0.998	0.995	0.995	1.000
SMOTE	0.991	0.981	0.981	1.000
Downsampled	0.928	0.855	0.892	0.962

Table 7.2: Model fit metrics for random forest models applied to to USA ribbon data with three resampling methods. Across all four metrics the difference between upsampling with replacement and SMOTE is at most 0.017. Upsampling scores highest in all metrics, except specificity in which it scores equal best with SMOTE, while downsampling scores worst in all metrics.

Overall, upsampling with replacement appears to lead to the best model performance, with slightly higher scores that SMOTE in most cases. This is relatively consistent with the findings of Park and Carriquiry (2019), though they found that SMOTE resampling led to better performance than upsampling. Given than SMOTE is a more complicated method, and contains synthetic samples not contained in the actual data, we conclude that upsampling is the preferable approach to preprocess the data.

7.1.2 Australian Casework Data

For the Australian casework data, there is almost no difference between the performance of the three resampling techniques. The models trained on downsampled data perform worst, but only by a small margin. The difference between models trained on data upsampled with replacement, and SMOTE

resampling is even smaller. Table 7.3 gives the areas under the ROC curves for the three resampling methods applied to the three models. We can see that the only difference between the three methods is seen in the third decimal place.

Resampling Method	Logistic Regression	Decision Tree	Random Forest
Upsampled	0.993	0.998	1.000
SMOTE	0.994	0.996	1.000
Downsampled	0.993	0.986	0.997

Table 7.3: ROC areas under the curve for three resampling techniques applied to the three ML models trained and evaluated on Australian data. We see that, for each ML model, the only difference between the three resampling methods is seen in the third decimal place.

Resampling Method	Accuracy	Kappa	Sensitivity	Specificity
Upsampled	0.998	0.996	0.996	1.000
SMOTE	0.997	0.994	0.994	1.000
Downsampled	0.992	0.984	0.989	0.995

Table 7.4: Model fit metrics for random forest models applied to to Australian casework with three resampling methods. Across all four metrics the difference between upsampling with replacement and SMOTE is at most 0.002. In fact, except in the case of Cohen's kappa, for which downsampling scores 0.01 lower than SMOTE, there is only a difference between the resampling methods in the third decimal place.

In Table 7.4 we present a comparison of the three resampling techniques using accuracy, Cohen's kappa coefficient, sensitivity and specificity. In this comparison we include only the results for random forest models, so as not to confuse the results with a comparison of models. We again note that there is only a difference between the resampling methods in the third decimal place, except in the case of Cohen's kappa, for which downsampling scores 0.01 lower than SMOTE. In particular, we note that across all four metrics the difference between upsampling with replacement and SMOTE is at most 0.002.

Overall, it appears that upsampling seems to yield the best performance, even if only slightly better than SMOTE. By the same reasoning as we argued for the USA data, since SMOTE is a more complex method, and produces artificial samples, we conclude that upsampling is the preferable approach for preprocessing the data.

7.2 Comparison of Models

Having seen that preprocessing the data by upsampling the minority class with replacement was the optimal method, we proceed for the remainder of the chapter using this technique. We now present the results of the logistic regression, decision tree and random forest models applied to the USA ribbon and Australian casework data sets, and provide a recommendation of which model is preferable in each case.

For the decision tree and random forest models, a hyper-parameter tuning procedure was necessary to optimise their performance. The tuning process was conducted using a random grid search with cross validation. For both models, we tuned the minimum samples needed to split a node, the minimum samples required at each leaf node, the maximum tree depth, the maximum features considered at each split, and in the case of random forests, the number of trees used. A more detailed explanation of the tuning process is provided in Appendix D. In Chapter 8 we will present the final choices of model, and provide a link to their specifications on GitHub.

7.2.1 USA Ribbon Data

When applied to the USA ribbon data, the decision tree and random forest models performed best. Both had overall accuracies greater than 0.99, and the random forest performed particularly well in its adjusted accuracy, with a Cohen's kappa coefficient of 0.994. Both of these model achieved perfect specificity scores, and the random forest saw an increase of 0.009 in sensitivity over the single decision tree model. In this case, the logistic regression model performed much more poorly, with an overall accuracy of 0.827, a very low value for Cohen's kappa of 0.654 and quite a low sensitivity of 0.710. This suggests that almost 30% of pairs which matched in truth, were classified as not matching. It did however achieve a specificity of 0.945.

7.2.2 Australian Casework Data

Table 7.6 shows the model fit metrics for each of the three models fit to the Australian casework data. All three models achieved an overall accuracy of over 97%, and the decision tree and random forest models then performed almost identically. In particular, these two models achieved perfect scores for specificity (to three decimal places), suggesting that no pairs of fragments were incorrectly classified as matching, when they were not matching in truth. This is a very good result to have achieved, as this is the main source of

USA Ribbon Models

Model	Accuracy	Kappa	Sensitivity	Specificity	ROC AUC
Random Forest	0.997	0.994	0.994	1.000	0.997
Decision Tree	0.993	0.985	0.985	1.000	0.993
Logistic Regression	0.827	0.654	0.710	0.945	0.827

Table 7.5: Model fit metrics for machine learning models applied to USA ribbon data. We see that the logistic regression model performed worst in all metrics, with an accuracy 0.186 lower than the decision tree. The decision tree also scored significantly lower in sensitivity and Cohen's kappa coefficient, 0.654 compared to 0.985 and 0.994. The difference between the random forest and decision tree models was less substantial, but the random forest model performed best across the board.

error which we seek to minimise. The random forest then achieved increases of 0.001 in Cohen's kappa and sensitivity, and equivalent scores in the other three metrics. Logistic regression performed worst in all metrics, though still achieving a low false-positive rate with a specificity of 0.991. Overall, these results are comparable to the USA ribbon data results in terms of overall ranking, but the distinction between the models is much less substantial here.

Due to the similarity in performance in the random forest as compared to a decision tree, and substantial improvement in both of these compared to logistic regression, we recommend the decision tree model for the Australian casework.

Australian Casework Models

Model	Accuracy	Kappa	Sensitivity	Specificity	ROC AUC
Random Forest	0.998	0.998	0.998	1.000	0.998
Decision Tree	0.998	0.997	0.997	1.000	0.998
Logistic Regression	0.976	0.953	0.962	0.991	0.976

Table 7.6: Model fit metrics for machine learning models applied to Australian casework data. We see that all models achieved an accuracy of at least 97%, and that the two decision tree models achieved perfect scores for specificity, meaning no false positive predictions were made. Logistic regression performed worst in all metrics, with a score of 0.991 in specificity, 0.009 lower than the score of 1.000 for the other two models. The decision tree and random forest models performed almost identically.

7.3 Overfitting

We have now compared the three machine learning classifiers, and see that the decision tree and random forest models appear to offer the best performance. Next, we consider checking whether the models are overfit to the training data.

Given that a model's predictive capability is based on the features in the training data set, one is at risk of developing a model which is too heavily biased towards these features. In the most extreme case, this would be evidenced by near perfect performance when predicting on the training data, and poorer performance when predicting on testing data.

For the Australian casework and to USA ribbon data, we now apply the three models to both the training and testing data. Using this, we can compare the performance between the models applied to the training data and the testing data. In general, we expect that the model would perform better on the training data, and then aim to minimise the difference in performance, as a large difference would be a sign of overfitting.

7.3.1 USA Ribbon Data

In Table 7.7, we note that the logistic regression model actually performs slightly better when applied to the testing data as compared to the training data. This suggests that the model is definitely not overfit to the training data, and it is simply due to chance that the testing data was in some sense "easier" for the model to classify.

In Tables 7.8 and 7.9, we see perfect performance when the decision tree and random forest models are applied to the training data. The difference is largest in the case of the decision tree model applied to the USA ribbon data, with a difference of 0.015 in Cohen's kappa and sensitivity. However, given that across the board, the scores for all metrics are above 0.98, the results suggest that no model is overfit to the training data.

7.3.2 Australian Casework Data

In Table 7.10, we note that across the board, the logistic regression model performs better when applied to the training data, but that the difference is of the order of 0.01 in all metrics. Any difference is negligible, and so we can be satisfied that the model is not overfit.

In Tables 7.11 and 7.12 we note that both the decision tree and random

Logistic Regression

Data	Accuracy	Kappa	Sensitivity	Specificity	ROC AUC
Training	0.823	0.646	0.703	0.943	0.823
Testing	0.827	0.654	0.710	0.945	0.827

Table 7.7: Comparison of model fit metrics for logistic regression applied to USA Ribbon training and testing data. We note that, unexpectedly, the model performs slightly better on the testing data, with increases of 0.002 to 0.008 in the model fit metrics.

Decision Tree

Data	Accuracy	Kappa	Sensitivity	Specificity	ROC AUC
Training	1.000	1.000	1.000	1.000	1.000
Testing	0.993	0.985	0.985	1.000	0.993

Table 7.8: Comparison of model fit metrics for decision tree applied to USA Ribbon training and testing data. We see that the model performed better on the training data, with increased performance in all metrics except specificity, with the largest difference of 0.015 in Cohen's kappa and sensitivity.

Random Forest

Data	Accuracy	Kappa	Sensitivity	Specificity	ROC AUC
Training	1.000	1.000	1.000	1.000	1.000
Testing	0.997	0.994	0.994	1.000	0.997

Table 7.9: Comparison of model fit metrics for random forest applied to USA Ribbon training and testing data. We see that the model performed better on the training data, with increased performance in all metrics except specificity, with the largest difference of 0.006 in Cohen's kappa and sensitivity.

Logistic Regression

Data	Accuracy	Kappa	Sensitivity	Specificity	ROC AUC
Training	0.977	0.954	0.962	0.992	0.977
Testing	0.976	0.953	0.962	0.991	0.976

Table 7.10: Comparison of model fit metrics for logistic regression applied to Australian casework training and testing data. We see a very small decrease in performance of the model applied to the testing data compared to the training data, with decreases of 0.001 in all metrics except sensitivity, where there was no difference at all.

Decision Tree

Data	Accuracy	Kappa	Sensitivity	Specificity	ROC AUC
Training	1.000	1.000	1.000	1.000	1.000
Testing	0.998	0.997	0.997	1.000	0.998

Table 7.11: Comparison of model fit metrics for decision tree applied to Australian casework training and testing data. We see a small decrease in performance of the model applied to the testing data compared to the training data, with the most substantial decrease of 0.003 in Cohen's kappa and sensitivity.

Random Forest

Data	Accuracy	Kappa	Sensitivity	Specificity	ROC AUC
Training	1.000	1.000	1.000	1.000	1.000
Testing	0.998	0.998	0.998	1.000	0.998

Table 7.12: Comparison of model fit metrics for random forest applied to Australian casework training and testing data. We see a very small decrease in performance of the model applied to the testing data compared to the training data, with decreases of 0.002 in all metrics except specificity, where there was no difference at all.

forest models receive perfect scores across the board when applied to the training data. As for these models applied to the USA ribbon data, with near-perfect scores of over 0.99 in all metrics, the performance on the testing data provides support that these models are not overfit.

7.4 Generalisability of Models

Building on the assessment of overfitting, in this section we assess how well the models generalise when applied to different elemental glass data. This is achieved by taking the models which were trained on the USA ribbon data, and applying the model to predict on the Australian casework data, and vice versa. In doing so, we can see whether each type of model could be used effectively in a different scenario.

7.4.1 Method

For this to be achieved properly, some modifications to the data were needed. Recall from Chapter 2, that the Australian data contains tin, which the USA data does not, and that the USA data contains sodium, which the Australian data does not. As such, each of these elements were removed from the respective data sets, and the models were trained on these reduced data sets. We tested all three machine learning models on both data sets, but include only the results of the decision tree and random forest models as they performed best in both cases. In each case, the models were trained on a training set, but then applied to the entirety of the other balanced data set. That is, one model was trained on 75% of the Australian data – the training set – and then applied to the entire balanced USA data set, and vice versa.

7.4.2 Results

Tables 7.13 and 7.14 shows the model fit metrics for the decision tree and random forest models trained on USA ribbon data applied to Australian data. The first row shows the metrics for the model applied to the USA training data, which has changed only slightly in Cohen's kappa coefficient and sensitivity as compared to the data set including sodium shown in Tables 7.8 and 7.9. We note that both models perform poorly when applied to the Australian data, each with an overall accuracy of 50% and a Kappa coefficient of zero. These two metrics in combination tell us that the model performed exactly as well as a classifier which randomly guesses into which

class each observation should fall. However, upon inspection of the sensitivity and specificity, which are zero and one respectively, we note that the only way which this can occur is if all pairs of samples were classified as not matching. This is because, in order for sensitivity to be equal to one, the number of true negatives must be equal to the number of true negatives plus the number of false positives. Therefore, the number of false positives must be zero. Equivalently, for the specificity to be zero, the number of true positives must be zero. So, it must be the case that there were no positive, or in other words no "match" predictions whatsoever. This was verified by checking the class predictions for each of these models. As a result, all three of these models are effectively useless for predicting on Australian casework data. This can be attributed to the fact that the USA ribbon data comprises samples taken from the same two factories at similar times. Therefore, samples which are considered different, are often extremely similar in their chemical composition, which leads to the model requiring two samples of glass to be near identical for it to classify them as a match.

USA Decision Tree Model

Data	Accuracy	Kappa	Sensitivity	Specificity	ROC AUC
USA testing data	0.993	0.985	0.985	1.000	0.993
Aus data	0.500	0.000	1.000	0.000	0.500

Table 7.13: Model fit metrics for decision tree models trained and tested on USA ribbon data and applied to Australian casework data. We see a substantial difference in the performance on the USA ribbon and the Australian casework data. The overall accuracy and ROC AUC both decrease by 0.497 to 0.500 and Kappa and specificity decrease to zero. Sensitivity increases slightly from 0.995 to 1.000.

USA Random Forest Model

Data	Accuracy	Kappa	Sensitivity	Specificity	ROC AUC
USA testing data	0.997	0.995	0.995	1.000	0.997
Aus data	0.500	0.000	1.000	0.000	0.500

Table 7.14: Model fit metrics for random forest models trained and tested on USA ribbon data and applied to Australian casework data. We see a substantial difference in the performance on the USA ribbon and the Australian casework data. The overall accuracy and ROC AUC both decrease by 0.497 to 0.500 and Kappa and specificity decrease to zero. Sensitivity increases slightly from 0.995 to 1.000.

Similarly, Tables 7.15 and 7.16 shows the metrics for the DT and RF models trained on Australian casework applied to the USA ribbon data. In

this case, we see the opposite effect where the models were trained on data with a great deal of variability, therefore allowing a higher degree of variability in its classification of matches. Unlike the models trained on the USA ribbon data, there is now a difference in how the two models generalise to the Australian casework. The decision tree model has an accuracy of less than 50%, and subsequently a kappa value less than zero – meaning that the model performs worse than a uniform random classifier. With sensitivity, specificity and ROC AUC all less than 0.5, we see that the model underperforms on same and different source comparisons.

The random forest model, however, performs better than the American RF model applied to Australian data, with an accuracy of approximately 0.6, and a sensitivity of 0.883, meaning that it more often than not correctly predicts non-matches. Its specificity is lower than the DT model, though, at 0.342, suggesting that the model predicts a large number of false positives. Overall, this suggests that the Australian casework models generalise slightly better than the USA ribbon model, but the performance is still much lower than would be acceptable for assessing evidence.

Australian Decision Tree Model

Data	Accuracy	Kappa	Sensitivity	Specificity	ROC AUC
Aus testing Data	0.998	0.997	0.997	1.000	0.998
USA Data	0.453	-0.094	0.496	0.411	0.453

Table 7.15: Model fit metrics for decision tree model trained and tested on Australia Casework data and then applied to USA Ribbon data. We see substantial difference in the performance on the Australian casework and USA ribbon data. All metrics receive scores less than 50%, and Cohen's kappa in particular is negative, suggesting worse performance than a uniform random classifier.

Australian Random Forest Model

Data	Accuracy	Kappa	Sensitivity	Specificity	ROC AUC
Aus testing Data	0.998	0.996	0.996	1.000	0.998
USA Data	0.607	0.215	0.873	0.342	0.607

Table 7.16: Model fit metrics for random forest models trained and tested on Australia Casework data and then applied to USA Ribbon data. We see substantial difference in the performance on the Australian casework and USA ribbon data. Wee see decreases in all metrics, with accuracy decreasing by 0.391. Kappa, specificity and ROC AUC also see large decreases, while sensitivity sees a smaller decrease by only 0.123.

7.5 Discussion and Summary

In this chapter we have investigated the use of machine learning methods to make comparison between forensic glass samples by binary classification. We have compared three ML methods, increasing in complexity, starting with logistic regression, then using decision trees and finally random forests. We also compared the performance of each of these methods after preprocessing the data using three resampling methods: downsampling the majority class, upsampling the minority class with replacement, and oversampling the minority class using the SMOTE algorithm. We found that in all three models, upsampling with replacement resulted in the best performance on both data sets, but that the difference between this and SMOTE was minimal.

With regards to the three models which were compared, in line with their increasing levels of complexity, we found that for both data sets, the logistic regression model performed worst, and the random forest models performed best. The difference in performance was most pronounced on the USA data set, for which the logistic regression model achieved 71.0% accuracy on same source pairs, and 94.5% accuracy on different source pairs. The random forest model, however, achieved 99.4% accuracy on same source pairs and 100% accuracy on different source pairs.

We noted that across all of the models which were fit, the models applied to Australian casework data performed better in terms of prediction accuracy than the models applied to the USA ribbon data. This difference in the performance of the models applied to the two data sets makes sense, given their respective origins. The Australian data is highly variable, and as such it is "easier" to distinguish between the observations in this data set, than the USA ribbon data, where the samples all came from the same two factories, and were manufactured at similar times. This result suggests that for any casework data, which typically would contain a great deal of variability, a decision tree model would be entirely satisfactory. The added complexity of the random forest offered only a negligible increase in performance over the more easily interpreted decision tree model for the Australian casework, and a difference of the order of 0.01 on the USA ribbon data.

It is also worth noting that the decision tree models performed well "out of the box", that is, with the default model parameters rather than tuned parameters. This is a benefit in terms of ease of implementation, so that models can be fit and applied without the need for extensive education about machine learning methodology.

When it came to the way in which models generalised to predict on different data sets, the models trained on USA data were completely unable to

predict on Australian data. The models trained on Australian data however had some success on the USA data. While the decision tree model performed worse than a uniform random classifier, the random forest was able to make correct predictions 60% of the time, mostly on same source pairs. The performance was, however, far too poor to be used in practice.

This is a very important observation to make, as it suggests that in any given location, whether that be a state or country, for example, a relevant data base would need to be collected which is representative of the types of glass and level of variability present in that location. Models would then need to be trained on this appropriate database in order to be used to effectively classify matches between casework samples. However, we were only able to test the generalisability of two data sets, comprising very different samples and with different levels of variability. As such further research should be conducted into the use of models trained on one database to predict on anther, particularly using multiple sets of realistic casework databases, as compared to the laboratory-like conditions in which the ribbon samples were manufactured.

We have now sees that the decision tree models outperform the likelihood ratio methods and the ellipsoid criterion in predictive accuracy. The ellipsoid criterion, however, with sensitivity and specificity of 0.990 and 0.992 respectively, is not far behind the 0.997 and 1.000 offered by the decision tree model on Australian data. With each of these methods outperforming the LR in classification, one must ask whether these methods can be extended to include a measure of the strength of evidence. Park and Carriquiry (2019) suggest that this can be achieved using what is known as a score-based likelihood ratio (SLR), constructed from any classifier which results in a score. In the next chapter, we will explain and apply this method to the decision tree and ellipsoid criterion scores.

Chapter 8

Score-Based Likelihood Ratios

We have established that the ellipsoid criterion and decision tree models offer better classification performance than likelihood ratios. Now, we aim to extend these methods to measure the strength of the evidence for or against a match. Park and Carriquiry (2019) suggest that the match probabilities generated by ML methods can be used to quantify the strength of evidence in favour of the same or different source hypothesis using the so-called score-based likelihood ratio (SLR) introduced by Davis *et al.* (2012). They note also that this method can be applied to any method which produces a score which is then classified as a match or non-match according to some critical value, namely the interval-based criteria. As such, we will apply this method to the best performing score-based classifiers: the decision tree and the ellipsoid 4σ criterion.

8.1 Procedure

SLRs are calculated by taking the full sets of same source and different source scores, and finding kernel density estimates of the distributions of scores. (We remind the reader than more information about kernel density estimation can be found in Appendix A.4). For the ellipsoid criterion, for which all scores are positive numbers, a gamma kernel was used. For the decision tree, since the scores lie between zero and one, a beta kernel was used for this estimation. In doing so, empirical distributions of same source scores, and of different source scores are obtained. The distribution of same source scores can be considered to represent the density of scores given that the pair of samples originate from the same source, and vice versa for the different source scores. As such these empirical distributions are then used for the numerator and

denominator densities in the SLR, and the score for any given comparison is used to evaluate these densities. The result is therefore a likelihood ratio. The magnitude of the resultant likelihood ratios is determined by the shape of these empirical distributions. As such, it is impacted by the discriminating power of the method used to produce the scores.

8.2 Prior Findings

Park and Carriquiry (2019) demonstrate that the resultant likelihood ratios can vary greatly depending on the method used to calculate the score. In particular, they compare the SLR values for five specific fragments in the USA ribbon data as determined by their random forest model and the standard 4σ interval criterion.

Amongst these samples, they consider two same source pairs and two different source pairs, all of which are correctly predicted by both the RF and 4σ criterion. In each case, the SLR also supports the binary prediction – being greater than one for a match and less than one for a non-match – but the scale of the values differs by orders of magnitude between the methods. This potentially brings into question the idea of calibration again, though one could also argue that the method which yields more extreme values, the RF, simply has a greater discriminating power. The fifth comparison which they make is a different source pair which is incorrectly predicted by both the 4σ criterion (with a score of 3.78, which is less than 4), and the RF (with a score of 0.558, greater than 0.5). For the 4σ criterion, the resultant SLR is greater than one, which aligns with the incorrect prediction of same source. However, the resultant SLR for the RF prediction is less than one, in favour of the different source hypotheses - which contradicts the score. As a result, one can question whether constructing an SLR can potentially correct for errors in the binary prediction of ML methods, but also whether the critical value used should simply be optimised to ensure that correct predictions are maximised. This discrepancy between the class prediction from the original score and the resultant SLR is of interest, and demonstrates that score-based likelihood ratios can lead to some unexpected and potentially confusing results, as also noted by Morrison and Enzinger (2016) and Hepler *et al.* (2012).

We wish to also highlight that the use of score-based likelihood ratios has been subject to some criticism beyond these discrepancies. In particular, we direct the interested reader to Neumann and Ausdemore (2020) for a detailed exposition on this topic, and Morrison and Enzinger (2018) for a discussion

of why SLRs should be constructed to include a measure of typicality of glass samples.

Prior work on SLRs has also found that typically, calculating LRs using a score-based method yields values of smaller magnitude (in terms of the number of factors of ten away from 1), than the traditional LR procedures (Bolck *et al.*, 2015). This is attributed to SLRs being calculated from what can be considered univariate projects of the multivariate data.

8.3 Classification Results

In this section we construct score-based likelihood ratios for the scores generated by the ellipsoid criterion and the match probabilities predicted by the decision tree models for the Australian casework data.

For the decision tree classifier, we note that the predicted probabilities for the Australian casework are almost exclusively ones and zeros, and in particular, that all of the pairs of samples which are predicted to match, received a predicted probability of one. As a result, a probability density function cannot be constructed for the same source distribution. One could construct a probability mass function and simply assign a probability mass of one, to the value 1, but this would mean that the SLR is informed only by the likelihood of the observed probability, given that the fragments originate from different sources. In other words, the numerator would be held constant, and the likelihood ratio would be entirely determined by its denominator.

In order to address this, the decision tree model was tuned in such a way as to ensure that there was more variety in the predicted match probabilities. This was achieved by tuning the minimum impurity decrease required for a node to split. Recall that nodes are split according to either Gini impurity or information gain. This hyperparameter represents the minimum decrease in impurity, whether that be Gini impurity or the entropy in the case of information gain, which is required for a node to be split. If another split would result in an impurity decrease below this minimum, the node will not be split and instead will become a leaf. By default, there is no minimum impurity decrease, and setting higher values will result in a model with less specific leaves. That is, each leaf will account for a larger number of samples, and the predictions at leaves will be less certain, resulting in probabilities strictly between zero and one. Tuning this hyperparameter is one way to perform what is known as pruning a decision tree. Higher values compromise accuracy in order to improve generalisability. This process of tuning leads to our final choices of decision tree models. The specifications for these models

are available on GitHub at github.com/olountain/forensic_models.

For the decision tree models, plots of the distributions of original scores and SLRs were included. The equivalent plots were not included for the ellipsoid criterion scores as the same source and different source scores were on vastly different scales.

8.3.1 USA Ribbon Data

Ellipsoid 4σ Criterion

Once the ellipsoid criterion scores had been converted to SLRs, the classifications agreed with the original classification 97.1% of the time. There were 34 disagreements, of which four were incorrectly predicted as non-matching by the original classification, but correct by SLR; one was incorrectly predicted as matching by the original classification; and the remaining 29 were correctly predicted as non-matching by the original classification but incorrectly by SLR. As a result of these changes, the difference in classification performance is summarised in Table 8.1.

method	accuracy	kappa	sensitivity	specificity
Ellipsoid 4σ	0.950	0.575	0.917	0.951
Ellipsoid 4σ SLR	0.929	0.507	1.000	0.926

Table 8.1: Model fit metrics for ellipsoid 4σ criterion and corresponding SLR applied to USA ribbon data. We note that there is are decreases of approximate 0.02 in overall accuracy and specificity of the SLR as compared to the original classification. The sensitivity, however has improved to correctly predict same source pairs 100% of the time. This difference arises when the SLR gives a different prediction to the original classification.

Decision Tree

After tuning the USA ribbon data decision tree, we found that a minimum impurity decrease of 0.0005 when splitting the nodes using Gini impurity allowed for a range of predicted probabilities while minimising the decrease in accuracy. In Table 8.2 we see that a compromise in performance metrics was required to achieve this. The most substantial decrease was seen in Cohen's kappa coefficient, which fell by 0.091. Accuracy, sensitivity and ROC AUC also saw decreases ranging from approximately 0.05 to 0.08, while specificity only decreased by 0.008.

Park and Carriquiry (2019) found that when treating the resultant SLR as a binary classifier, using one as a critical value, there were a few cases in which the classification by SLR did not agree with the original classification criterion. In our model, however, all predictions agreed.

In Figure 8.1 we see the kernel density estimates of match probabilities by decision tree classification which were used to construct the SLRs. The distribution of the resultant SLRs is then shown in Figure 8.2. To make the plot clearer, we show the distribution of the base-10 logarithm of the SLRs. In each case the grey, vertical dotted line shows the critical value used to make binary classification. We see in both cases that the two densities are mostly distinct from one another which is a visual representation of the discriminating power of this approach.

Data	Accuracy	Kappa	Sensitivity	Specificity	ROC AUC
Original	0.993	0.985	0.985	1.000	0.993
Pruned	0.947	0.894	0.901	0.992	0.947

Table 8.2: Model fit metrics for decision tree model applied to over-sampled Australian casework before and after pruning via the minimum impurity decrease hyperparameter.

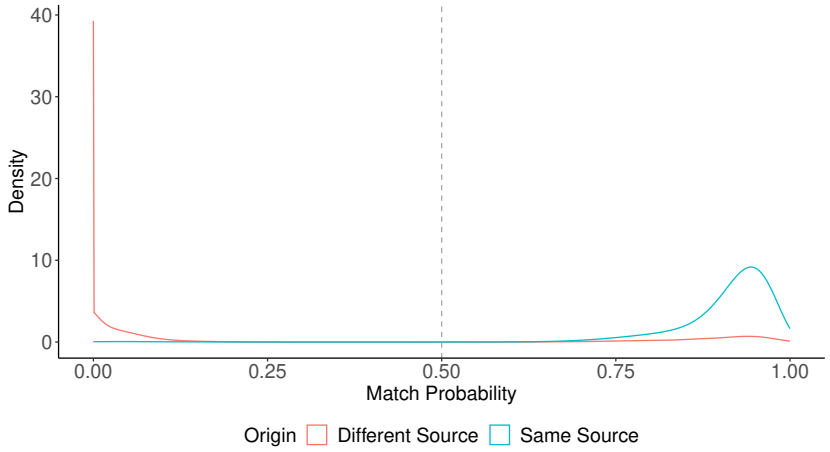


Figure 8.1: Kernel density estimates for same source and different source match probabilities from decision tree model fit to USA ribbon data.

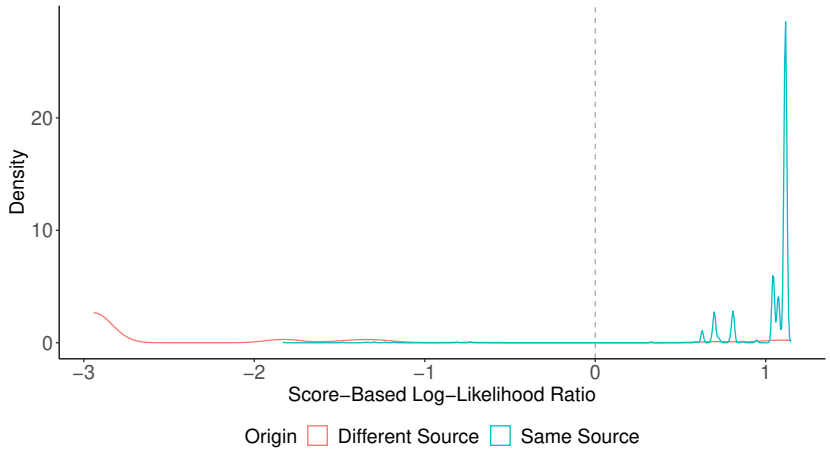


Figure 8.2: Distributions of base 10 logarithm of score-based likelihood ratios constructed from decision tree scores on USA ribbon data.

8.3.2 Australian Casework Data

Ellipsoid 4σ Criterion

In the case of SLRs constructed from the 4σ ellipsoid criterion, we found that the classification by SLR agreed with the original classification 99.5% of the time. Of the observations on which the models disagreed, there was one comparison in which a same source pair was declared as not matching by the ellipsoid criterion, but received an SLR score of approximately 23. There were also 20 different source pairs which were correctly predicted as not matching by the ellipsoid criterion, but received SLR scores greater than one. That is, for the pairs where the classification by SLR did not agree with the original prediction, the original prediction was correct 20 out of 21 times. In Table 8.3 we see that classification performance of the SLR system constructed from the ellipsoid criterion performs slightly worse in overall accuracy, kappa and specificity, but has a perfect score for sensitivity. Overall, this constitutes a minor change in performance.

Decision Tree

After tuning the decision tree model, we found that a minimum impurity decrease of 0.000075 when splitting the nodes using Gini impurity allowed for a range of predicted probabilities while minimising the decrease in accuracy. The performance metrics for this modified model are given in Table 8.4. We see that in each metric, the performance has decreased by at most 0.008.

method	accuracy	kappa	sensitivity	specificity
Ellipsoid 4σ	0.992	0.837	0.990	0.992
Ellipsoid 4σ SLR	0.988	0.772	1.000	0.988

Table 8.3: Model fit metrics for ellipsoid 4σ criterion and corresponding SLR applied to Australia casework data. We note that there is a small decrease in overall accuracy of the SLR as compared to the original classification. This difference arises when the SLR gives a different prediction to the original classification.

Data	Accuracy	Kappa	Sensitivity	Specificity	ROC AUC
Original	0.998	0.997	0.997	1.000	0.998
Pruned	0.994	0.989	0.990	0.999	0.995

Table 8.4: Model fit metrics for decision tree model applied to oversampled Australian casework before and after pruning via the minimum impurity decrease hyperparameter.

Figures 8.3 and 8.4 show the kernel density estimates of match probabilities by decision tree and the system of SLRs respectively. As for the USA data the two densities are mostly distinct from one another, providing is a visual representation of the discriminating power of the approach. We note also that, as was the case for the USA ribbon data model, all of the predictions by the SLR agreed with the original decision tree predictions.

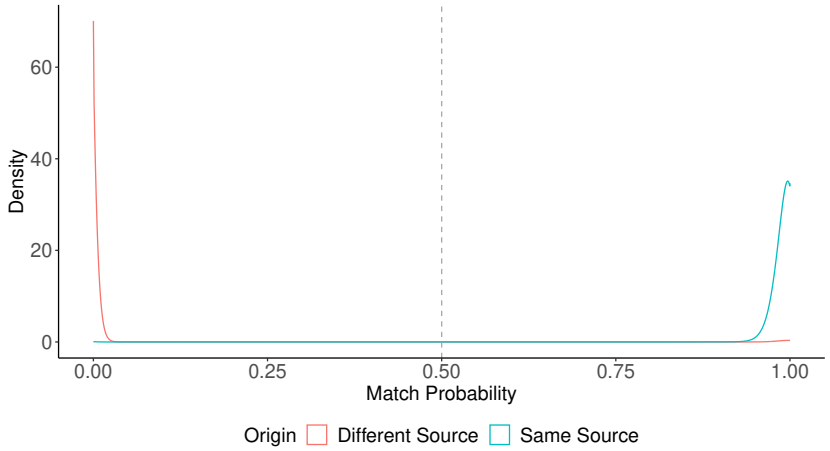


Figure 8.3: Kernel density estimates for same source and different source match probabilities from decision tree model fit to Australian casework data.

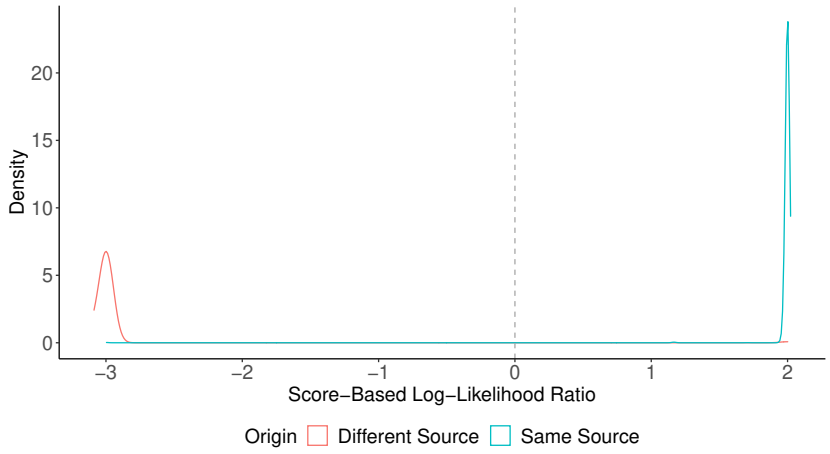


Figure 8.4: Distributions of base 10 logarithm of score-based likelihood ratios constructed from decision tree scores on Australian casework data.

8.4 Calibration Results

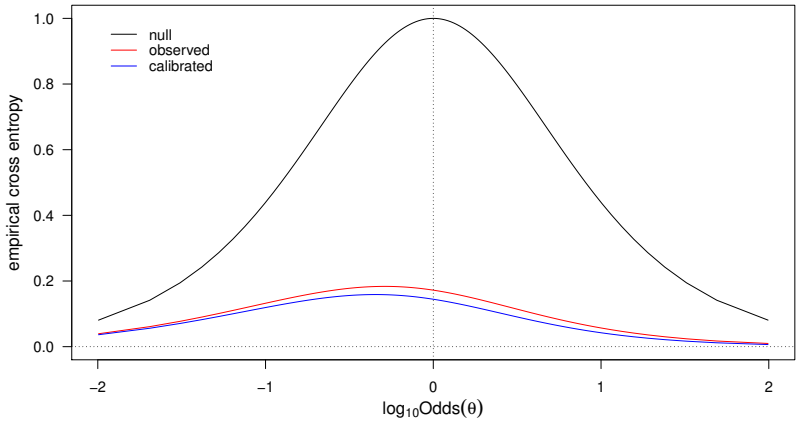
Recall from Part II that an important measure of the performance of a likelihood ratio system is its calibration. The calibration measures the extent to which a likelihood ratio can be properly interpreted as quantifying how many times more likely the same source hypothesis is than the different source hypothesis. The calibration is assessed via empirical cross entropy (ECE) plots, and is explicitly quantified by the cost log-likelihood ratio (C_{llr}).

8.4.1 USA Ribbon Data

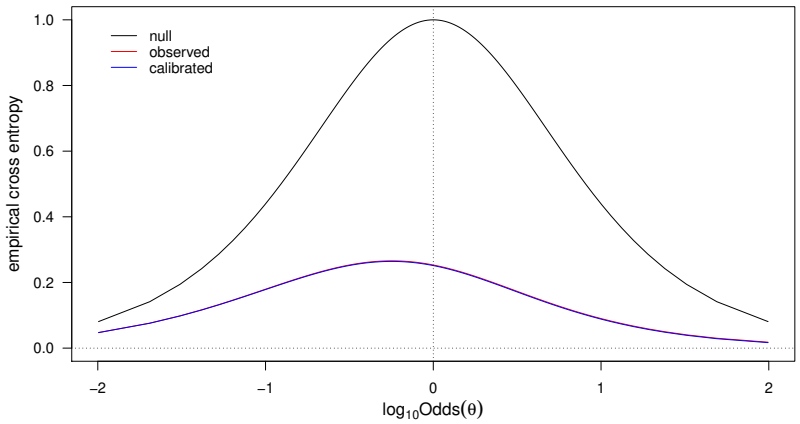
Figures 8.5a and 8.5b show the ECE plots for the USA ribbon data ellipsoid criterion and decision tree SLR systems respectively. In both cases we see that the observed ECE curve is well below the null curve, and that the C_{llr} values are very small, particularly for the decision tree, suggesting that the SLR systems are well-calibrated and require no adjustment.

8.4.2 Australian Casework Data

Similarly, Figures 8.6a and 8.6b show the ECE plots for the Australian casework ellipsoid criterion and decision tree SLR systems respectively. As for the USA ribbon data, both plots show that the observed ECE curve is well below the null curve, and that the C_{llr} values are very small. The decision tree SLR in particular, has the smallest C_{llr} of any likelihood ratio system presented



(a) Ellipsoid criterion. $C_{llr} = 0.02769501$.



(b) Decision tree. $C_{llr} = 0.001878512$

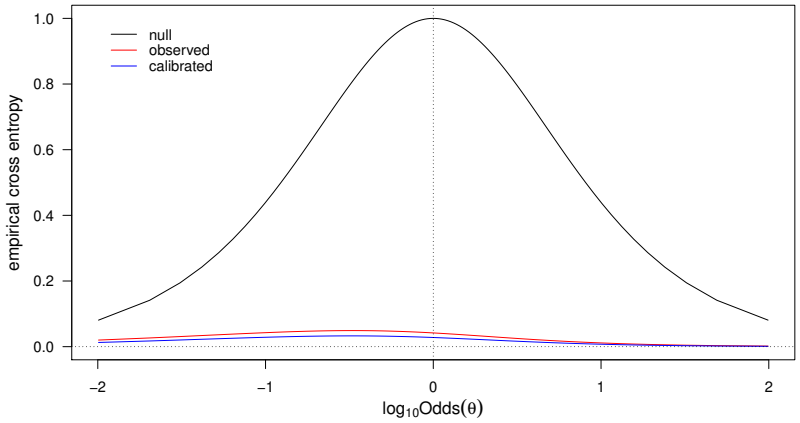
Figure 8.5: ECE plots of score-based likelihood ratio systems constructed from oversampled decision tree models. We see in both cases that the observed ECE curve is well below the null curve, and that very little difference can be seen between the observed and calibrated curves, suggesting very good levels of calibration. This is quantified by the C_{llr} score which is very low in both cases.

in this thesis. This suggests that all of the SLR systems are well-calibrated and require no adjustment.

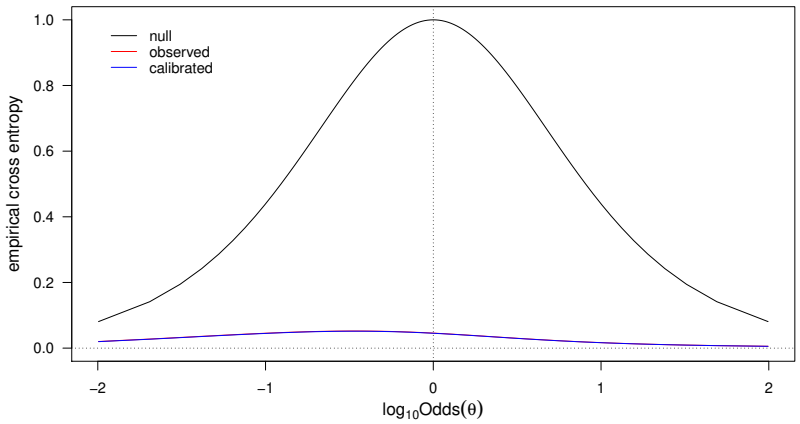
8.5 Summary

In this chapter, we have taken the best performing score-based models: the decision tree and the ellipsoid 4σ criterion, and constructed score-based likelihood ratio systems. This procedure has maintained the high accuracy of these methods and improved upon them by allowing for a measure of "strength of evidence" otherwise enjoyed by only the traditional likelihood procedures. Doing so required a small compromise in the predictive accuracy of the decision tree model by adjusting the minimum impurity decrease at each split in the decision tree to ensure that a wider distribution of match probabilities were generated. SLRs were then created using kernel densities estimates for these distributions. We found that the classification by SLR was entirely in line with that original decision tree classification, and included a small number of discrepancies in the case of the ellipsoid criterion. We found that the values of the SLRs were smaller in magnitude than the traditional LRs, which is in line with what was found by Bolck *et al.* (2015). This can be seen in particular by comparing Figures 8.2 and 8.4 with Figures B.1 to B.4 in Appendix B.

We also assessed the calibration of these SLR systems, and found that they were indeed well-calibrated, suggesting that this procedure provides a robust way to quantify the strength of evidence from ML classification. Overall, on the Australian data the decision tree SLR appears to be the clear winner with the highest overall accuracy, high sensitivity and specificity, and the smallest C_{llr} value. On the USA ribbon data, the decision tree SLR system compromises sensitivity in favour of specificity, while the opposite is true for the ellipsoid SLR system. In general, we wish to maximise specificity, and in doing so minimise incorrect evidence in favour of the prosecution, and so the decision tree would likely remain the most favourable choice.



(a) Ellipsoid criterion. $C_{llr} = 0.0139$.



(b) Decision tree. $C_{llr} = 0.000447$

Figure 8.6: ECE plots of score-based likelihood ratio systems constructed from oversampled decision tree models. We see in both cases that the observed ECE curve is well below the null curve, and that very little difference can be seen between the observed and calibrated curves, suggesting very good levels of calibration. This is quantified by the C_{llr} score which is very low in both cases.

Part IV

Discussion and Final Remarks

Chapter 9

Comparison of Methods

We have now presented three broad types of method for making comparison between glass samples: interval-based criteria (current practice), likelihood ratios, and machine learning classifiers. The question is therefore: which is best? In this chapter we seek to answer this question by balancing the accuracy, parsimony and practical usefulness of the models considered.

9.1 Binary Classification Performance

The best likelihood ratio model was the multivariate kernel procedure after the max-mean transformation had been applied, and of the machine learning models the decision tree model achieved the best compromise of accuracy and parsimony. Comparing these models with the multivariate adaptation to the interval criteria, Table 9.1 shows that in all of the metrics for binary classification performance except sensitivity, the decision tree model performs best. In sensitivity, the standard 4σ interval criterion performs perfectly, but the decision tree is only worse by 0.003. We note also that the ellipsoid σ criterion is not far behind the decision tree, with sensitivity and specificity scoring only 0.007 and 0.008 less respectively. This begs the question of whether the machine learning procedure is worthwhile, since it is a far more complex and involved methodology, which offers accuracy improvements of less than 0.01 on both same and different source pairs.

On the USA data, however, the difference was more substantial. In Table 9.2 we see that the decision tree model provided improvements of 0.068 and 0.049 over the ellipsoid criterion for sensitivity and specificity respectively. Further research should be conducted into the use of ML methods on different data sets.

Model	Accuracy	Kappa	Sensitivity	Specificity
Decision Tree	0.998	0.997	0.997	1.000
Ellipsoid 4σ Criterion	0.992	0.837	0.990	0.992
Max-mean MVK LR	0.964	0.788	0.897	0.971
Standard 4σ Criterion	0.961	0.497	1.000	0.960

Table 9.1: Classification performance metrics for best models applied to Australian casework data. The decision tree model performs best overall, with the highest score in all metrics except for sensitivity, in which it scored 0.003 lower than the standard 4σ interval criterion. The decision tree offered improvements of 0.007 and 0.008 over the Ellipsoid 4σ criterion in sensitivity and specificity respectively. The LR methods performed worst out of the newly investigated methods.

Model	Accuracy	Kappa	Sensitivity	Specificity
Decision Tree	0.993	0.985	0.985	1.000
Ellipsoid 4σ Criterion	0.950	0.575	0.917	0.951

Table 9.2: Classification performance metrics for the decision tree and ellipsoid 4σ criterion applied to the USA ribbon data. The decision tree model scored higher in all metrics, with an overall accuracy 0.043 higher than the ellipsoid criterion.

Interestingly, in terms of classification performance, the LR methods with the best accuracy were outperformed by both the standard and ellipsoid 4σ criteria. While this may seem to suggest that the LR methods offer no benefit over the interval-based methods, we must remember that the likelihood ratio can be used to describe strength of evidence. This does, however, suggest that LR methods only offer improvement when they are used to quantify the strength of evidence, and should not be employed only as a binary classification method.

9.2 Benefits and Shortcomings of Models

As we noted in the previous section, the likelihood ratio procedures perform worse than both the machine learning models and the ellipsoid criterion in terms of binary classification performance. However, the LR methodology does have the edge in that it offers more information than the other methods by measuring the strength of evidence.

The interval-based methods are by far the easiest approaches to implement. For the standard criterion, one needs only to compute the mean and standard deviation of each element in the control sample, and the mean of

each element in the recovered sample. Similarly, for the ellipsoid criterion, one needs to compute the covariance matrix of the control sample rather than only the individual standard deviations. For these methods, no background database is used, and the models do not need to be trained. The likelihood ratios and machine learning models, however, are much more involved. Both of these approaches require a background database on which to train the models, and the ML methods in particular, can be quite computationally expensive to train. The mathematics involved in the LR calculations is more complex than any of the other methods, and therefore potentially less accessible. For the ML methods, while the algorithms used may not be accessible for all, the fundamental approach can be easily understood. With this in mind, it is impressive that the ellipsoid criterion performed only slightly worse than the decision tree models, given that the implementation is far less complicated, and no background database is required. With this in mind, for the purposes of classification alone, it might be reasonable to recommend the use of the ellipsoid criterion if a relevant population database cannot be obtained, and the decision tree model if such a database can be obtained.

While the ellipsoid criterion and decision tree models performed best in terms of classification, the strength of the evidence in favour of either hypothesis obtained by likelihood ratios can be a very important piece of information for a jury when glass evidence is being used in combination with many other pieces of evidence. Given the substantially better performance offered by these two methods over the LRs, we consider the score-based likelihood ratio systems that can be obtained from the ellipsoid and decision tree models. With this in mind, our focus shifts from simply looking to optimise classification performance metrics such as accuracy, to include the cost log-likelihood ration (C_{llr}), which measures the level of calibration.

We note that the two score-based likelihood ratio systems performed the best when applied to the USA ribbon data (Table 9.3). The decision tree SLR performed best in overall accuracy, beating the ellipsoid SLR by 0.018. The decision tree SLR, however, favoured specificity (different source accuracy) over sensitivity (same source accuracy), while the reverse was true for the ellipsoid criterion. The SLR systems also achieved the highest levels of calibration, with the decision tree model in particular, achieving an SLR more then ten times less than the ellipsoid method.

Amongst the LR procedures applied to the Australian casework, we see that the decision tree SLR receives the best score in all metrics (Table 9.4). It achieves an overall accuracy 0.03 higher than the highest of the standard LRs, sensitivity at 0.99, 0.072 higher than any model, and near perfect specificity of 0.999. Further to this, it received a C_{llr} 100 times smaller than any

Model	Accuracy	Kappa	Sensitivity	Specificity	C_{llr}
Decision Tree SLR	0.947	0.894	0.901	0.992	0.00188
Ellipsoid 4σ SLR	0.929	0.507	1.000	0.926	0.0277
Min-trade-off MVK	0.841	0.077	0.842	0.841	0.095
Max-mean MVK	0.803	0.065	0.907	0.802	0.089

Table 9.3: Performance metrics for the best likelihood ratio procedures applied to the USA ribbon data. We see that the decision tree score-based likelihood ratio outperforms all other techniques, with near-perfect sensitivity and specificity, as well as a cost log-likelihood ratio score 10 times smaller than any other procedure.

of the other LR techniques, suggesting a much better level of calibration. For the MVK LR, max-mean transformation outperforms the min-trade-off transformation in accuracy by 0.048, but has a C_{llr} value much larger than the SLR procedures.

Model	Accuracy	Kappa	Sensitivity	Specificity	C_{llr}
Decision Tree SLR	0.994	0.989	0.990	0.999	0.000447
Ellipsoid 4σ SLR	0.988	0.772	1.000	0.988	0.0139
Min-trade-off MVK	0.916	0.604	0.917	0.916	0.207
Max-mean MVK	0.964	0.788	0.897	0.971	0.504

Table 9.4: Performance metrics for the best likelihood ratio procedures applied to the Australian casework data. We see that the decision tree score-based likelihood ratio outperforms all other techniques, with near-perfect sensitivity and specificity, as well as a cost log-likelihood ratio score 100 times smaller than any other procedure.

Overall, on the Australian data the decision tree appears to be the clear winner between the methods tested. The results suggest that the standard likelihood ratio procedures cannot outperform the ML techniques in terms of classification accuracy. On the USA ribbon data, however, the decision tree and ellipsoid SLR systems trade sensitivity and specificity. This data set, is of course, unrealistic though in its low level of variability, and in practice it is unlikely that this trade-off would be substantial. The main selling point of the likelihood ratio approach was therefore the added benefit of quantifying the strength of evidence, but with the implementation of a score-based likelihood ratio constructed from ML results (and the ellipsoid criterion), the standard LR procedures no longer have an edge over the ML techniques.

Chapter 10

Summary and Future Research

In this thesis we have investigated the use of several statistical methods to make comparison between forensic glass samples measured by elemental composition, specifically using laser ablation-inductively coupled plasma mass spectrometry. We found that the simple extension to the current practice of an ellipsoid criterion taking advantage of the multivariate structure of the data improves the predictive accuracy, and that machine learning methods improve upon it further. We found that likelihood ratio based approaches do not perform as well in prediction, but offer the ability to quantify the strength of evidence. In order to introduce this concept into the machine learning framework, we have considered the use of score-based likelihood ratios, which maintain the predictive accuracy, and allow for the strength of match. The complete procedure to implement the models in practice and incorporate SLRs is given in Appendix E.

We have also considered the performance of these methods on a diverse data set with a great deal of variation, and a homogeneous data set with much less spread. We found that the models were more accurate when predicting on the diverse data set, but only by approximately five percentage points. The best performing model, the decision tree, achieved an accuracy of 0.994 on the diverse data set, and 0.947 on the homogeneous data set, as well as a specificities of 0.999 and 0.992 respectively.

To potentially improve upon these results, there are a number of areas which warrant further research following the results presented in this thesis, both in the context of likelihood ratios and score-based methods. Firstly, alternative methods could be considered to recalibrate likelihood ratio systems, whether that be different ways of optimising the critical value, or using a completely different transformation on the system. The critical value optimisation method has the benefit of being invertible, and any alternative ap-

proaches should also retain this property. For any score-based method, such as the interval criteria or machine learning classifiers, alternative choices of critical value could be entertained to maximise performance. For example, a different number of standard deviations (rather than four) could be used for interval criteria, and perhaps a cut-off of 0.5 does not allow for the best compromise between sensitivity and specificity in machine learning models. Finally, some alterations could be made to the machine learning approaches. While the decision tree method provides a good compromise between accuracy and parsimony, alternative models could be considered. For example, the Bayesian additive regression tree which were considered by (Park and Carriquiry, 2019) along with their analysis of random forests.

With these areas for possible further research in mind, the results presented in this thesis demonstrate that a great deal of predictive accuracy can be achieved by taking full advantage of the multivariate structure of elemental glass measurements. Further, the results support the idea that the models can perform very well on different data sets, and have the potential to be very effective if implemented by forensic practitioners. While glass evidence only constitutes a single component of a legal case, it is important to ensure that the methods used to evaluate the data are high in accuracy. In particular, the statistical procedures used for this comparison should adhere to the philosophy of "innocent until proven guilty", and the models we have presented perform well in minimising the rate at which samples are incorrectly classified as matching.

Appendices

Appendix A

Mathematical Details

A.1 Logistic Regression

In Part III of the thesis we investigated the use of logistic regression to compare glass samples. Here we provide a review of the theory behind logistic regression. Consider grouped data of the form.

$$(n_1, y_1, \boldsymbol{x}_1), (n_2, y_2, \boldsymbol{x}_2), \ldots, (n_m, y_m, \boldsymbol{x}_m)$$

for m groups of size n_i. Consider also the model where

$$Y_i \sim \mathrm{B}(n_i, \pi_i)$$

independently for each $i = 1, \ldots, m$. Logistic regression aims to relate the success probability π_i to the predictor x_i, specifically

$$\pi_i = \pi(x_i).$$

We wish to define this model analogously to linear regression, which can theoretically predict any real number, but faces the hurdle that $0 \leq \pi_i \leq 1$. In order to address this, we now introduce one more term, the logit η_i:

$$\eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right).$$

The logit now satisfies the requirement of linear regression that $-\infty < \eta_i < \infty$, and is invertible such that π_i can be easily calculated as

$$\pi_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}.$$

The logistic regression model is then defined as a linear regression model as

$$\eta_i = \boldsymbol{\beta}^T \boldsymbol{x}_i.$$

Now what remains is to estimate the values of the coefficients $\boldsymbol{\beta}$. This is done by applying maximum likelihood estimation to $\ell(\boldsymbol{\beta} \mid \boldsymbol{y})$, where ℓ is the product of m binomial likelihoods. Appendix A.2 provides the details of maximum likelihood estimation.

Logistic regression gives an estimate of an empirical success probability – or for our purpose, an empirical probability of match between two glass samples. To make classification, one must choose a threshold probability above which two samples are said to match. As mentioned, we will use 0.5 as this threshold, meaning that pairs of fragments are classified as matching if they are predicted to have a greater than 50% chance of being a match.

A.2 Maximum Likelihood Estimation

In many circumstances, such as regression, it may be desirable to choose an optimum value for a certain parameter. One such method of optimisation is maximum likelihood estimation.

Definition A.2.1. Let $\boldsymbol{X}$ be a random variable from a distribution with parameters $\boldsymbol{\theta}$, and let f be the probability density (if $\boldsymbol{X}$ is continuous) or mass (if $\boldsymbol{X}$ is discrete) function of $\boldsymbol{X}$. The likelihood function of $\boldsymbol{\theta}$ is given by

$$\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) = f(\boldsymbol{x}).$$

We often simply write $\mathcal{L}(\boldsymbol{\theta})$ for $\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x})$.

In other words, the likelihood function is equal to the probability density function, but framed as a function of the parameters of that distribution, rather than the random variable. With this in mind, the likelihood function can be used to find an optimum value of $\boldsymbol{\theta}$, for which the probability density/mass function is maximised. This is known as the maximum likelihood estimator:

Definition A.2.2. Let $\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x})$ be a likelihood function. The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is given by

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}}\, \mathcal{L}(\boldsymbol{\theta}).$$

It is often the case that the likelihood function is not concave, which makes maximising the function practically difficult. However, it is more often the case that log-likelihood functions are concave. This leads us to the notion of a log-likelihood function:

Definition A.2.3. Let $\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x})$ be a likelihood function. The log-likelihood function is given by

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{x}) = \log \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}).$$

Since log is an increasing function, maximising the likelihood is equivalent to maximising the log-likelihood and practically, it is often easier to maximise the log-likelihood. As such, we instead define the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}}\, \ell(\boldsymbol{\theta}).$$

The maximum likelihood estimate can then be obtained by solving the equations

$$\frac{\partial \ell}{\partial \theta_i} = 0 \text{ for } i = 1, \ldots, n,$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$.

A.3 Graph Theory

The decision tree and random forest models rely on the mathematical notion of a graph, consisting of nodes and edges. Here, we provide the fundamental definitions required to describe a tree graph.

Definition A.3.1. A *graph* $G(N, E)$ is a set of nodes N along with a set of edges $E \subseteq N \times N$ which represent connections between the nodes.

It is worth noting that the basic definition of a graph does not allow for multiple edges between two nodes nor does it allow edges from a node to itself (known as self-loops). A graph which allows for these properties is known as a multigraph.

Definition A.3.2. A graph is said to be *directed* if its edges $(i, j) \in E$ represent a one-directional connection from node i to node j. If a connection exists in both directions between nodes i and j, then $(i, j) \in E$ and $(j, i) \in E$.

Graphically, we represent nodes in a graph as circles (often labelled with names or numbers) and represent the edges as lines between the nodes. In the case of a directed graph, we represent the directed edges with an arrow in the direction of the edge.

Definition A.3.3. Let $G(N, E)$ be a graph and let $i, j \in N$. There is said to be a *path* between i and j if there exists a sequence of edges

$$(i, v_1), (v_1, v_2), \ldots, (v_{n-1}, v_n), (v_n, j).$$

That is, when traversing the graph, there is a set of edges such that j can be reached from i.

Definition A.3.4. A graph $G(N, E)$ is said to be *connected* if there exists a path between every pair of nodes.

The idea of connectedness simply means that there are no outlying nodes, nor can the graph be considered to be made up of two or more separate graphs. We next consider the idea of a cycle within a graph.

Definition A.3.5. A directed graph is said to be *acyclic* if for each node $i \in N$, there does not exist a path of length greater than or equal to one from i to itself. That is, no self-loops exist and when traversing the graph, when a node is left it cannot be returned to.

With this in mind, we can now define a tree:

Definition A.3.6. A (*rooted*) *tree* is a connected acyclic graph. That is, a graph in which every pair of nodes is connected by exactly one path. One node is designated as the *root* and if the graph is directed, all paths originate from the root node. The nodes at which the paths terminate are called *leaves*.

Figure A.1 provides a visualisation of a directed rooted tree $G(N, E)$ with nodes and edges given by

$$\begin{aligned} N &= \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}, \\ E &= \{(1, 2), (1, 3), (2, 4), (2, 5), (3, 6), (4, 7), \\ &\qquad (4, 8), (4, 9), (6, 10), (6, 11), (11, 12)\}. \end{aligned}$$

A.4 Kernel Density Estimation

Kernel density estimation (KDE) is a method of estimating the probability density function of a random variable based on a sample of data. To begin with, we consider the notion of a kernel:

Definition A.4.1. A *kernel* K is non-negative, real valued integrable function. That is, for some $S \subseteq \mathbb{R}$, $K : S \to \mathbb{R}$ such that

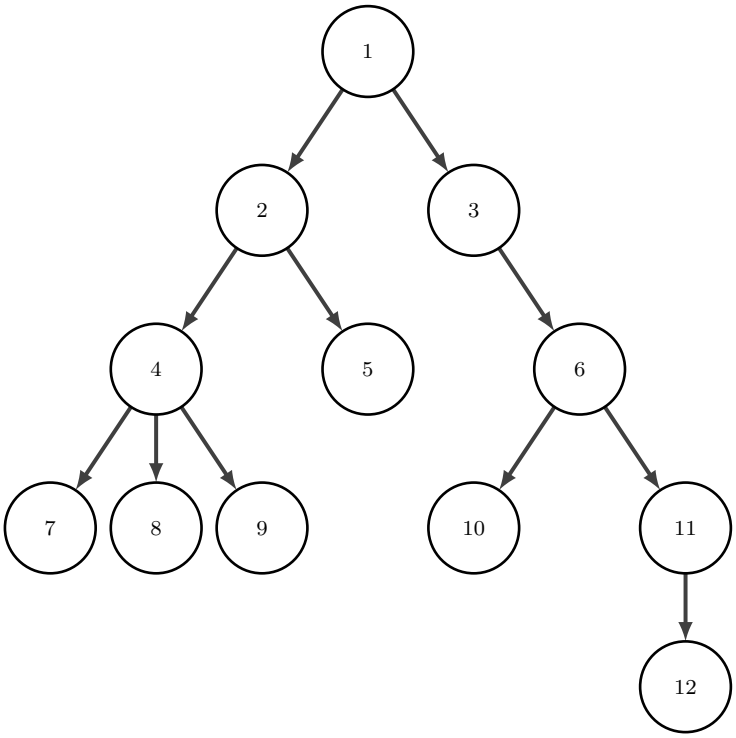


Figure A.1: Example of a directed rooted tree.

1. $K(s) \geq 0$ for all $s \in S$, and
2. $\int_{\mathbb{R}} K(s)\mathrm{d}s < \infty$.

There are some common choices of kernel. For example, for symmetric data taking values on the real line, a normal kernel may often be chosen:

$$K_{N(\mu,\sigma^2)}(x) = e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Similarly, a gamma kernel may be chosen for positive data, or a beta kernel for data which lies between zero and one. With this in mind, we can define a kernel density estimator as follows:

Definition A.4.2. Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be a random sample from some univariate distribution for which the density f is unknown. The *kernel density estimator* with bandwidth $h > 0$ is give by

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right).$$

The bandwidth h can be considered as a smoothing parameter. One typically wishes to choose h as small as possible, but there are several factors involved in the optimisation of this parameter. A detailed discussion of this can be found in Park and Marron (1990), Sheather (1992), Cao *et al.* (1994), Jones *et al.* (1996), Agarwal and Aluru (2010) and Xu *et al.* (2015).

Appendix B

Distributions of Likelihood Ratios

In this appendix we include plots of the distributions of log-likelihood ratio systems before and after calibration transformations have been applied. We include also, for convenience, the distributions of the log of the score-based likelihood ratio systems calculated from the decision tree models.

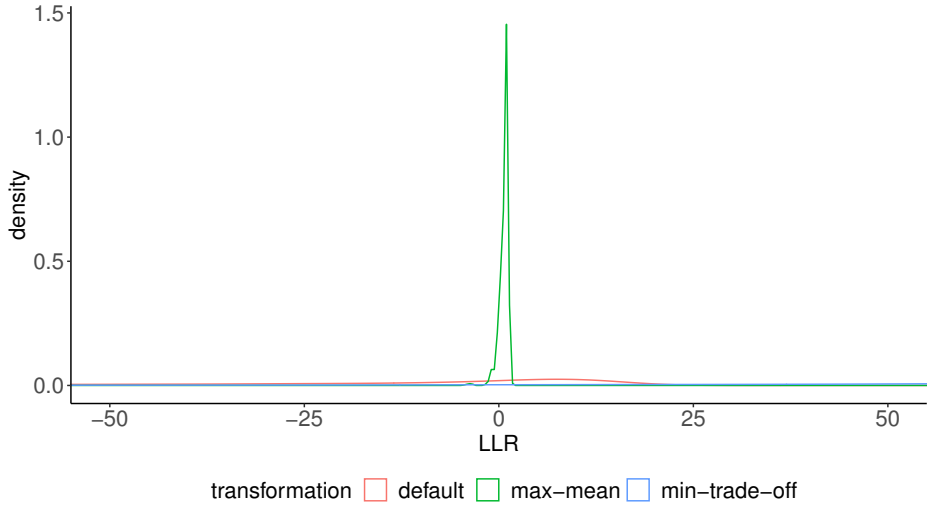


Figure B.1: Distributions of same source log-likelihood ratios constructed from the USA Ribbon Data

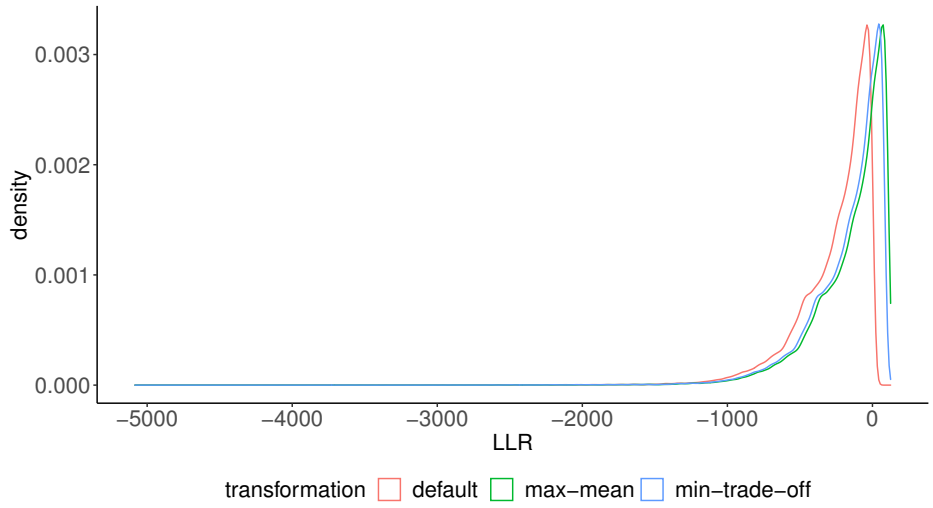


Figure B.2: Distributions of different source log-likelihood ratios constructed from the USA Ribbon Data

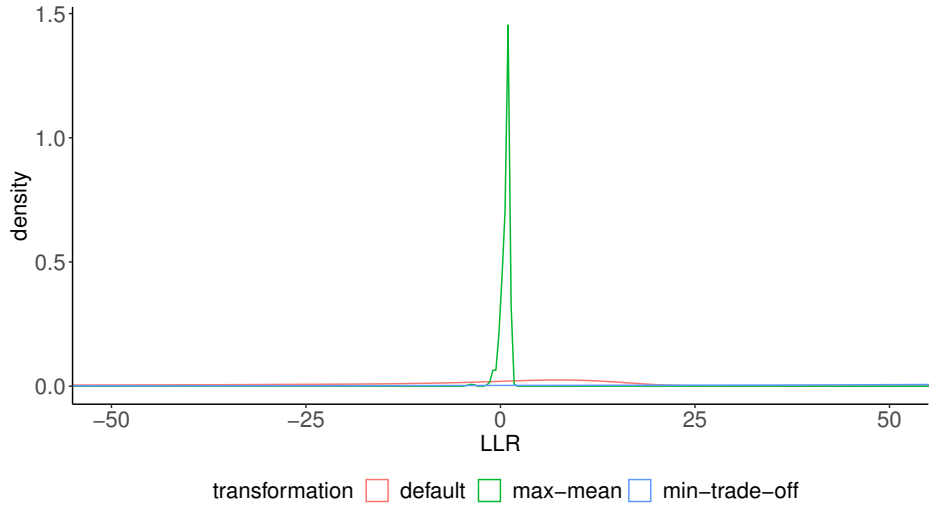


Figure B.3: Distributions of same source log-likelihood ratios constructed from the Australian Casework Data

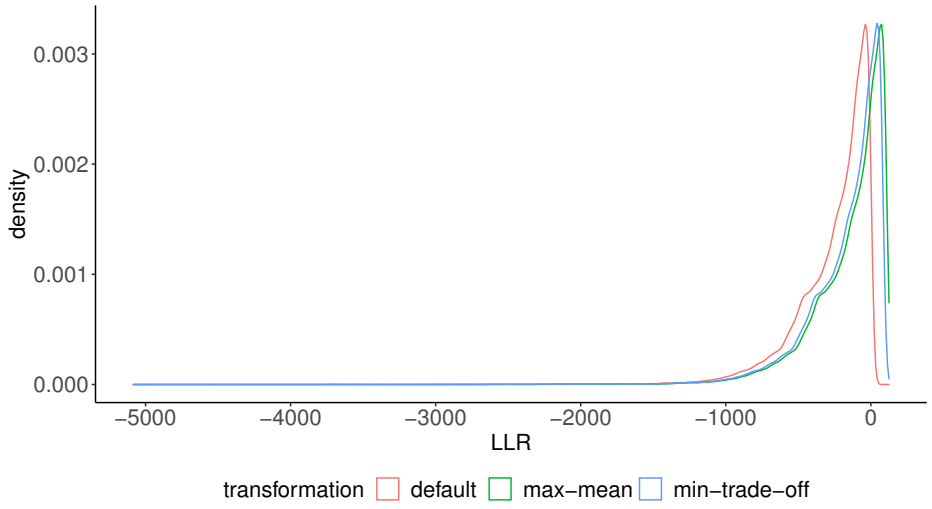


Figure B.4: Distributions of different source log-likelihood ratios constructed from the Australian Casework Data

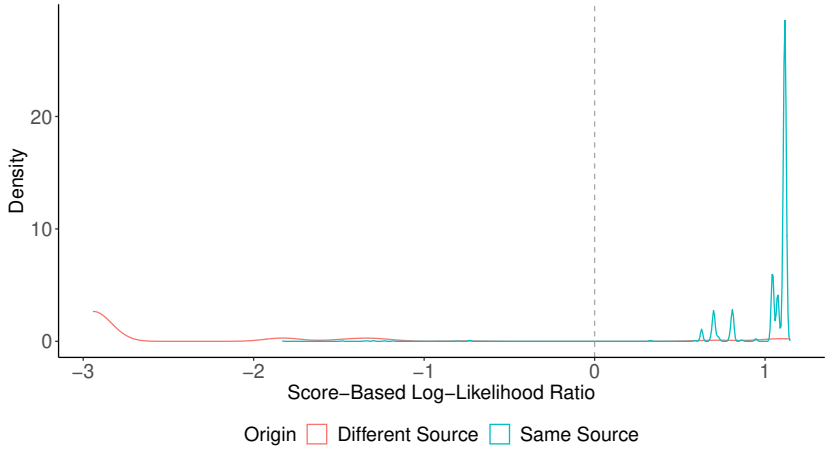


Figure B.5: Distributions of base 10 logarithm of score-based likelihood ratios constructed from decision tree scores on Australian casework data.

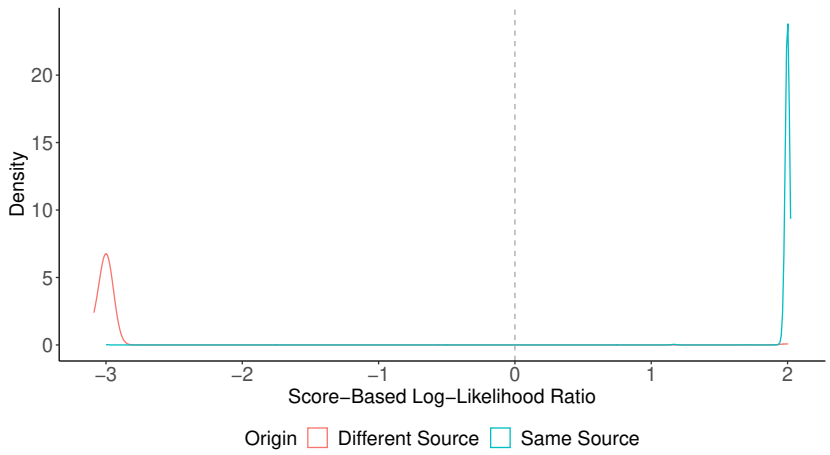


Figure B.6: Distributions of base 10 logarithm of score-based likelihood ratios constructed from decision tree scores on Australian casework data.

Appendix C

Algorithms

Algorithm C.1: PAV algorithm to obtain a set of calibrated LRs

Input : Set of observed likelihood ratios and ground truth classes $\{(LR_i, C_i)\}_{i=1}^{n}$

Output: Set of calibrated likelihood ratios and ground truth classes $\{(LR_i^{cal}, C_i)\}_{i=1}^{n}$

`/* Number of same and different source samples */`

Define $n_{ss} \leftarrow \text{sum}\left(\mathbb{1}_{\{C_i=\text{same source}\}}\right)$

Define $n_{ds} \leftarrow \text{sum}\left(\mathbb{1}_{\{C_i=\text{different source}\}}\right)$

`/* Sort the` LR_i `into ascending order */`

Set $\{LR_i\} \leftarrow \text{ascending}\left(\{LR_i\}\right)$

for $i \leftarrow 1$ **to** n **do**

  Define $p_i \leftarrow \mathbb{1}_{\{C_i=\text{same source}\}}$

end

Define $\boldsymbol{P} \leftarrow ([p_1, p_2], [p_2, p_3], \ldots, [p_{n-m}, p_n])$

while *there exist out of order blocks* **do**

  Select a pair of out of order blocks $[p_q, p_r]$ and $[p_r, p_s]$

  Replace them with $[p_q, p_s]$

  Compute θ_{qs}

  Update $\boldsymbol{P}$

end

Define $O_i^{\text{prior}} \leftarrow n_{ss}/n_{ds}$

Define $O_i^{\text{post}} \leftarrow p_i/(1-p_i)$

Compute $LR_i^{cal} \leftarrow O_i^{\text{post}}/O_i^{\text{prior}}$

return $\{(LR_i^{cal}, C_i)\}_{i=1}^{n}$

Algorithm C.2: SMOTE

Input : Number of minority class samples T, Amount of SMOTE $N\%$, Number of nearest neighbours k

Output: $TN/100$ synthetic synthetic minority class samples

if $N < 100$ **then**
  Randomise the T minority class samples
  Set $T \leftarrow TN/100$
  Set $N \leftarrow 100$
end
Set $num\_attr \leftarrow$ number of attributes
$Sample[\,][\,]$ `// array for original minority class samples`
Set $new\_index \leftarrow 0$ `// number of synthetic samples generated`
$Synthetic[\,][\,]$ `// array for synthetics samples`
for $i \leftarrow 1$ **to** T **do**
  Compute k nearest neighbours for i, save indices in nn_array
  Populate(N, i, nn_array)
end

`/* Function to generate the synthetic samples */`

Function $Populate(N, i, nn\_array)$
  while $N \neq 0$ **do**
    Sample $nn \sim \mathcal{U}\{1, k\}$ `// randomly choose a neighbour`
    for $attr \leftarrow 1$ **to** num_attr **do**
      $dif \leftarrow Sample[nn\_array[nn]][attr] - Sample[i][attr]$
      Sample $gap \sim \mathcal{U}(0, 1)$
      $Synthetic[newindex][attr] \leftarrow Sample[i][attr] + gap \times dif$
    end
    $new\_index \leftarrow new\_index + 1$
    $N \leftarrow N - 1$
  end
return

Appendix D

Machine Learning Model Tuning

For the decision tree and random forest models, we consider a number of important hyperparameters to tune. We first consider each hyperparameter separately. To do this, we fit and apply the decision tree model with all hyperparameters left at their default values except for the parameter in question. In each case we apply the model to the testing data and the training data, and plot the ROC AUC against the parameter values, to establish how performance varies. In doing so, we can establish which parameter values lead to better performance, and also assess the level of overfitting at each value. We then establish a range of values for each hyperparameter to consider in our tuning. We use a random grid search cross validation procedure across these potential values to obtain an optimal hyperparameter set.

D.1 Decision Trees

To begin, we consider four important hyperparameters to tune in the decision tree model: the minimum samples required to split a node, the minimum samples required at each leaf node, the maximum tree depth, and the maximum number of features considered at each split. The ROC AUC versus parameter value plots for each hyperparameter are displayed in Figure D.1.

D.1.1 Australian Casework Data

Minimum Samples to Split a Node

The minimum samples required to split a node (min-samples-split) was tuned at values ranging from 0.00001 and 0.15, where the value represents the proportion of the samples in the training set. Specifically, given n samples and a proportion p, min-samples-split will be calculated as $\lceil np \rceil$. We see in Figure D.1a that the decision tree performs worse and overfits to the training data at larger values of min-sample-split, and seems to reach a minimum performance level at which it remains as the value of min-samples-split increases. In particular the model performs better on the testing data than the training data at higher values. As such, the parameter will be tuned for values ranging from 0.00001 to 0.02.

Minimum Samples at Leaf Nodes

The minimum samples required at leaf nodes (min-samples-leaf) was also tuned at values ranging from 0.00001 and 0.15. Again, the value represents the proportion of the samples in the training set. In Figure D.1b we note that the overall performance decreases, and again note that at higher values, the model performs better on the testing data. As such, we will tune this parameter at values ranging from 0.00001 to 0.0001.

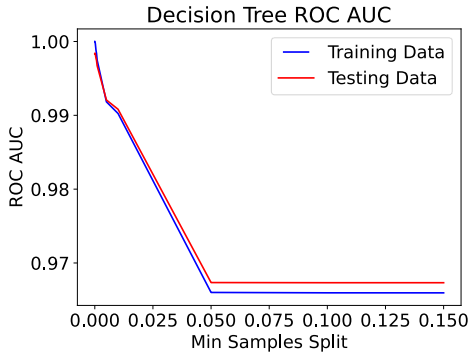
Maximum Tree Depth

The maximum tree depth is the maximum length of the path from the root node to the leaves. In Figure D.1c we plot the ROC of the model applied to the training and testing data for maximum tree depths ranging from one to 40. We note that the model performs better at larger values, but that there is very little change in the level of overfitting, and so leave this parameter at its default setting of none, meaning that there is no explicit limit on the tree depth.

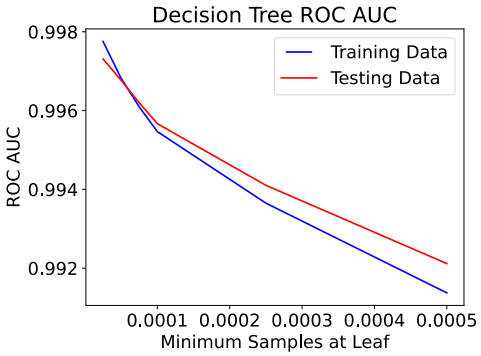
Maximum Features

Finally, we consider the maximum features considered at each split in the tree. In Figure D.1d we plot the performance at all possible values, that is 1 to 17, that is the total number of predictors. We note very little change in the performance at different values, and so tune this parameter with two

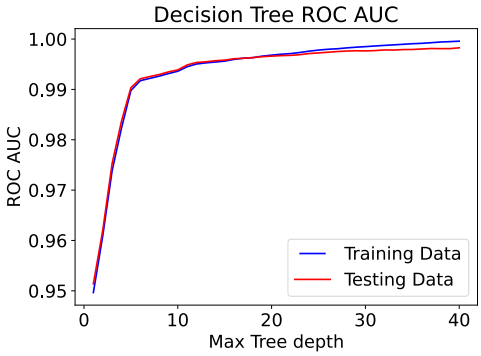
of the most common values: total number of features, the square root of the total number of features, rounded to the nearest integer.



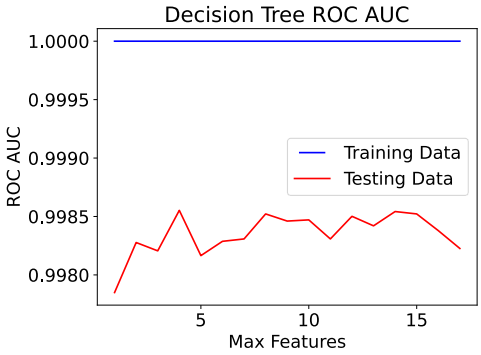
(a) Minimum samples to split a node.



(b) Minimum samples at each leaf.



(c) Maximum tree depth.



(d) Maximum features at each split.

Figure D.1: ROC AUC vs hyperparameter values for decision trees trained on Australian casework evaluated on training and testing data.

Cross Validation Results

Using 100 iterations of three-fold cross-validation procedure over a random grid of the hyperparameter values. We also tuned over the min impurity decrease parameter to avoid overfitting. This random grid was constructed over the potential values given in Table D.1.

After running the cross-validation, we obtained optimal values as given in Table D.2.

Hyperparameter	Values
Min-samples-split	(0.02, 0.01, 0.005, 0.001, 0.00075, 0.0005, 0.00025, 0.0001, 0.000075, 0.00005, 0.000025, 0.00001)
Min-samples-leaf	$(0.0001, 0.000075, 0.00005, 0.000025, 0.00001)$
Min impurity decrease	$(0.0, 0.0000001, 0.000001)$
Max features	None, square root
Splitting criterion	Information gain, Gini impurity

Table D.1: Potential hyperparameter values for decision tree fit to oversampled Australian casework data.

Hyperparameter	Value
Min-samples-split	0.000025
Min-samples-leaf	0.00001
Min impurity decrease	0.0000001
Max features	None
Splitting criterion	Information gain

Table D.2: Optimal hyperparameter values for decision tree fit to oversampled Australian casework data.

D.2 USA Ribbon Data

Applying the same procedure and tuning over the same values as for the Australian casework data, optimal hyperparameter values were determined for the decision tree model fit to oversampled USA ribbon data. These values are shown in Table D.3:

Hyperparameter	Value
Min-samples-split	0.00001
Min-samples-leaf	0.00001
Min impurity decrease	0.0000001
Max features	Square root
Splitting criterion	Gini impurity

Table D.3: Optimal hyperparameter values for decision tree fit to oversampled USA ribbon data.

D.3 Random Forest Models

For the random forest models, the same hyper-parameters were tuned over random grids spanning the same values as for the decision tree models. The only addition to this was the number of decision trees used in each model. This hyper-parameter was tuned with values ranging from 100 to 500 trees.

D.3.1 Australian Casework Data

Applying the random grid cross-validation procedure yielded the following optimal hyper-parameter values (Table D.4).

Hyperparameter	Value
Number of trees	100
Min-samples-split	0.000075
Min-samples-leaf	0.00001
Min impurity decrease	0.0000001
Max features	None
Splitting criterion	Gini impurity

Table D.4: Optimal hyperparameter values for random forest fit to oversampled Australian casework data.

D.3.2 USA Ribbon Data

Table D.5 displays the optimal hyper-parameter values for the random forest applied to the USA ribbon data.

Hyperparameter	Value
Number of trees	100
Min-samples-split	0.00005
Min-samples-leaf	0.000025
Min impurity decrease	0
Max features	None
Splitting criterion	Information gain

Table D.5: Optimal hyperparameter values for random forest fit to oversampled USA ribbon data.

Appendix E

Implementing the Models in Practice

In this appendix we describe the procedures required to implement the ellipsoid criterion and decision tree models in practice, and use the scores produced to calculate score based likelihood ratios.

E.1 Ellipsoid Criterion

Our ellipsoid 4σ criterion method was implemented in R (R Core Team, 2020). Implementing this model for practical use requires the following steps:

1. Obtain a data base of elemental glass measurements for a given location, labeled by their known sources.

2. Consider a list each possible pair of samples, labeled by whether they originate from the same or different sources. For each pair, randomly label one of the observations in the pair as the control sample and the other as recovered.

3. Apply a resampling technique to the data to balance the classes, as the different source class will have many more observations that the same source class.

4. Calculate the Mahalanobis distance between each set of pairs in the resampled data using the covariance matrix of the control sample. A shrinkage estimator may need to be used for the covariance matrix. Now label each pair as predicted to be matching or non-matching.

5. Assess the performance of the model by comparing the truth to the prediction.

Next, we must obtain the kernel density estimates required to produce score-based likelihood ratios. This can be done as follows:

1. Separate the set of predicted scores for same source and different source pairs.
2. For each set, obtain a kernel density estimate of the set of Mahalanobis distances. We recommend a gamma kernel since the scores must be greater than zero.
3. Using the same source KDe as the numerator, and the different source KDE as the denominator, calculate SLRs for each of the scores in the testing data.
4. Assess the performance and calibration of the SLR system.

Now all of the necessary steps have been taken to build the model, and it can be used in practice. The following steps can be used to analyse new observations:

1. For any new control and recovered observations which you wish to compare, calculate the Mahalanobis distance between the points using the covariance matrix of the control sample.
2. Using this Mahalanobis distance, evaluate the score-based likelihood ratio for this pair of observations.

E.2 Decision Tree

Our decision tree model was implemented in Python (Van Rossum and Drake Jr, 1995) using the scikit-learn package (Pedregosa *et al.*, 2011). Implementing this model for practical use requires the following steps:

1. Obtain a database of elemental glass measurements for a given location, labeled by their known sources.
2. Apply the process of pairwise differencing to this data set, as described in Section 6.2. This will result in a transformed database of differenced samples labeled according to whether they originate from the same or different sources.

3. Apply a resampling technique to the data to balance the classes, as the different source class will have many more observations that the same source class.

4. Split the data set into a training and a testing set.

5. Train the decision tree classier on the training set and tune the model.

6. Assess the performance of the model on the testing data.

At this stage, the model has now been fit to the data. Next, we must obtain the kernel density estimates required to produce score-based likelihood ratios. This can be done as follows:

1. Separate the set of predicted scores for same source and different source pairs in the testing data.

2. For each set, obtain a kernel density estimate of the set of scores. We recommend a beta kernel since the scores lie between zero and one.

3. Using the same source KDE as the numerator, and the different source KDE as the denominator, calculate SLRs for each of the scores in the testing data.

4. Assess the performance and calibration of the SLR system.

Now all of the necessary steps have been taken to build the model, and it can be used in practice. The following steps can be used to analyse new observations:

1. For any new control and recovered observations which you wish to compare, take the element-wise difference of the observations.

2. Apply the decision tree classifier to this differenced observation to obtain a probability of match (score).

3. Using this score, evaluate the score-based likelihood ratio for this pair of observations.

Bibliography

Nitin Agarwal and N. R. Aluru. A data-driven stochastic collocation approach for uncertainty quantification in MEMS. *International Journal for Numerical Methods in Engineering*, 83(5):575–597, mar 2010.

Ravindra K Ahuja and James B Orlin. A fast scaling algorithm for minimizing separable convex functions subject to chain constraints. *Operations Research*, 49(5):784–789, 2001.

C. G. G. Aitken and D. Lucy. Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(1):109–122, jan 2004.

Colin Aitken, Anders Nordgaard, Franco Taroni, and Alex Biedermann. Commentary: Likelihood ratio as weight of forensic evidence: A closer look. 9, jun 2018.

J. R. Almirall and T. Trejos. Advanced in the forensic analysis of glass fragments with a focus on refractive index and elemental analysis. *Forensic science review*, 18:73–96, July 2006.

Jose Almirall and Tatiana Trejos. Analysis of glass evidence. pages 228–272. John Wiley & Sons, Ltd, oct 2015.

ASTM-E2330-12. Test method for determination of concentrations of elements in glass samples using inductively coupled plasma mass spectrometry (ICP-MS) for forensic comparisons, 2012.

ASTM-E2927-16. Test method for determination of trace elements in soda-lime glass samples using laser ablation inductively coupled plasma mass spectrometry for forensic comparisons, 2016.

Annabel Bolck, Haifang Ni, and Martin Lopatka. Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison. 14(3):243–266, sep 2015.

Niko Brümmer and Johan Du Preez. Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2-3):230–275, 2006.

Gareth P Campbell and James M Curran. The interpretation of elemental composition measurements from forensic glass evidence iii. *Science & Justice*, 49(1):2–7, 2009.

Ricardo Cao, Antonio Cuevas, and Wensceslao González Manteiga. A comparative study of several smoothing methods in density estimation. *Computational Statistics & Data Analysis*, 17(2):153–176, feb 1994.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, jun 2002.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Roger Cook, Ian W Evett, Graham Jackson, PJ Jones, and Jim A Lambert. A model for case assessment and interpretation. *Science and Justice*, 38(3):151–156, 1998.

J.M. Curran, C.M. Triggs, J.R. Almirall, J.S. Buckleton, and K.A.J. Walsh. The interpretation of elemental composition measurements from forensic glass evidence: II. *Science & Justice*, 37(4):245–249, oct 1997.

James Michael Curran, Tacha Natalie Hicks Champod, and John S Buckleton. *Forensic interpretation of glass evidence*. CRC Press, 2000.

Linda J Davis, Christopher P Saunders, Amanda Hepler, and JoAnn Buscaglia. Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios. *Forensic science international*, 216(1-3):146–157, 2012.

Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

Joseph Fleiss. *Statistical methods for rates and proportions*. Wiley, New York, 1973.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Amanda B Hepler, Christopher P Saunders, Linda J Davis, and JoAnn Buscaglia. Score-based likelihood ratios for handwriting evidence. *Forensic science international*, 219(1-3):129–140, 2012.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer-Verlag GmbH, 2017.

M. C. Jones, J. S. Marron, and S. J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407, mar 1996.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159, mar 1977.

Olivier Ledoit and Michael Wolf. Honey, I shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4):110–119, jul 2004.

David Lucy, James Curran, and Agnieszka Martyna. *comparison: Multivariate Likelihood Ratio Calculation and Evaluation*, 2020. R package version 1.0-5.

Nicola Lunardon, Giovanna Menardi, and Nicola Torelli. ROSE: a package for binary imbalanced learning. *The R Journal*, 6(1):79, 2014.

Steven P. Lund and Hari K. Iyer. Likelihood ratio as weight of forensic evidence: A closer look. April 2017.

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

Geoffrey Stewart Morrison and Ewald Enzinger. What should a forensic practitioner's likelihood ratio be? *Science & Justice*, 56(5):374–379, 2016.

Geoffrey Stewart Morrison and Ewald Enzinger. Score based procedures for the calculation of forensic likelihood ratios – scores should take account of both similarity and typicality. 58(1):47–58, jan 2018.

Geoffrey Stewart Morrison. Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, 51(3):91–98, sep 2011.

D.F. Nelson and B.C. Revell. Backward fragmentation from breaking glass. *Journal of the Forensic Science Society*, 7(2):58–61, apr 1967.

Cedric Neumann and Madeline Ausdemore. Defence against the modern arts: the curse of statistics—part II: 'score-based likelihood ratios'. 19(1):21–42, mar 2020.

Soyoung Park and Alicia Carriquiry. Learning algorithms to evaluate forensic glass evidence. *The Annals of Applied Statistics*, 13(2):1068–1102, 2019.

Byeong U. Park and J. S. Marron. Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, 85(409):66–72, mar 1990.

Soyoung Park, Alicia Carriquiry, L. Kenneth Horkley, and David W. Peate. A database of elemental compositions of architectural float glass samples measured by LA-ICP-MS. *Data in Brief*, 30:105449, jun 2020.

Judea Pearl and Dana Mackenzie. *The Book of Why*. Hachette Book Group USA, 2020.

Judea Pearl. The limitations of opaque learning machines. *Possible minds: twenty-five ways of looking at AI*, pages 13–19, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

Daniel Ramos and Joaquin Gonzalez-Rodriguez. Reliable support: Measuring calibration of likelihood ratios. *Forensic Science International*, 230(1-3):156–169, jul 2013.

Daniel Ramos, Joaquin Gonzalez-Rodriguez, Grzegorz Zadora, and Colin Aitken. Information-theoretical assessment of the performance of likelihood ratio computation methods. *Journal of Forensic Sciences*, 58(6):1503–1518, jul 2013.

E.M. Rounds. A combined nonparametric approach to feature selection and binary decision tree design. *Pattern Recognition*, 12(5):313–317, jan 1980.

Leonard J Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Juliane Schfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1), jan 2005.

Simon J Sheather. *The performance of six popular bandwidth selection methods on some real data sets*. University of New South Wales, Australian Graduate School of Management, 1992.

Tatiana Trejos, Robert Koons, Stefan Becker, Ted Berman, JoAnn Buscaglia, Marc Duecking, Tiffany Eckert-Lumsdon, Troy Ernst, Christopher Hanlon, Alex Heydon, Kim Mooney, Randall Nelson, Kristine Olsson, Christopher Palenik, Edward Chip Pollock, David Rudell, Scott Ryland, Anamary Tarifa, Melissa Valadez, Peter Weis, and Jose Almirall. Cross-validation and evaluation of the performance of methods for the elemental analysis of forensic glass by μ-XRF, ICP-MS, and LA-ICP-MS. *Analytical and Bioanalytical Chemistry*, 405(16):5393–5409, may 2013.

Tatiana Trejos, Robert Koons, Peter Weis, Stefan Becker, Ted Berman, Claude Dalpe, Marc Duecking, JoAnn Buscaglia, Tiffany Eckert-Lumsdon, Troy Ernst, Christopher Hanlon, Alex Heydon, Kim Mooney, Randall Nelson, Kristine Olsson, Emily Schenk, Christopher Palenik, Edward Chip Pollock, David Rudell, Scott Ryland, Anamary Tarifa, Melissa Valadez, Andrew van Es, Vincent Zdanowicz, and Jose Almirall. Forensic analysis of glass by μ-XRF, SN-ICP-MS, LA-ICP-MS and LA-ICP-OES: evaluation of the performance of different criteria for comparing elemental composition. *Journal of Analytical Atomic Spectrometry*, 28(8):1270, 2013.

Guido Van Rossum and Fred L Drake Jr. *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.

Peter Vergeer, Yara van Schaik, and Marjan Sjerps. Measuring calibration of likelihood-ratio systems: A comparison of four metrics, including a new metric devPAV. *Forensic Science International*, 321:110722, apr 2021.

Peter Weis, Marc Dücking, Peter Watzke, Sonja Menges, and Stefan Becker. Establishing a match criterion in forensic comparison analysis of float glass using laser ablation inductively coupled plasma mass spectrometry. *Journal of Analytical Atomic Spectrometry*, 26(6):1273, 2011.

Xiaoyuan Xu, Zheng Yan, and Shaolun Xu. Estimating wind speed probability distribution by diffusion-based kernel density method. *Electric Power Systems Research*, 121:28–37, apr 2015.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002.