Evaluation of Transfer Evidence

Laura Allison

A Dissertation Submitted to the
University of Glasgow
for the degree of
Master of Science

School of Mathematics & Statistics

September 2011

Abstract

The question of whether two sets of measurements which constitute evidence come from the same source is one which is frequently sought to be answered by the forensic community. A common type of evidence comes in the form of glass fragments where the refractive index or elemental composition has been measured The most common way of evaluating trace evidence such as glass fragments is the likelihood ratio, which is a measure of evidential value. A two-level random effects model was used to determine the likelihood ratio for measurements of the refractive index and elemental composition of glass. Two different methods were applied to estimate the between-group distribution of the two datasets; normal approach and kernel density estimation. Both methods were applied to univariate refractive index data as well as to multivariate refractive index and elemental composition data. The effectiveness of each method was assessed in a simulation experiment in which pairs of known origin are compared with different pairs of known origin via the likelihood ratio and the incorrect comparisons are recorded by false negative and false positive rates.

The performed analysis showed that refractive index and elemental composition measurements can be used for identifying same and different-source pairs of glass fragments with a high degree of accuracy. The normal approach for the between-group distribution proved the superior method in both the refractive index and elemental composition sets of glass measurements with 0% false negative and 0.9% false positive rates for the refractive index and 3.4% false negative and 5.5% false positive rates for elemental composition.

Acknowledgements

First and foremost, I am deeply grateful to Dr. Tereza Neocleous who without her valuable support, guidance and technical expertise this thesis would not have been possible.

I would also like to extend my thanks to G. Zadora of the Institute of Forensic Research, Krakow for providing the databases for this project. In addition, I gratefully acknowledge the funding from the School of Mathematics and Statistics which allowed me to undertake this work which I have thoroughly enjoyed.

Thanks is also due to my fellow office friends: Andisheh, Collette, Greg, Kathryn, Mhairi and Stephen for providing me with humour throughout this enjoyable year.

Finally, I owe a great debt of thanks to my parents, Alan and Ann Allison, who have supported me throughout my education and it is down to them how far I have pushed myself academically to reach and surpass many goals. Thank you!

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Forensic science can be thought of as the investigation, explanation and then the evaluation of events of legal relevance to the case in hand which might include identity, origin and life history of humans. The evaluation is done by scientific techniques which allow the scientist to describe, infer and reconstruct the event.

When a crime is committed there is usually trace evidence left at the crime scene which could be fragments of glass, DNA, paint or fibres from clothes for example. The idea that tiny traces of material which can be invisible to the naked eye can be used to investigate crime is powerful. The distinctive characteristics of trace evidence are usually its microscopic size, its ability to transfer from one item to another and subsequently being lost from an item during a crime. Quite often during a crime, as well as trace evidence being left, there is also transfer evidence. This is defined as trace evidence transferred from victim to suspect and vice versa. This evidence is then collected by a team of forensic scientists and taken back to a forensic laboratory for analysis. It is analysed in such a way that numerical values are assigned to the evidence and it is then the role of the forensic statistician to statistically evaluate whether the trace evidence found at the crime scene and the trace

evidence recovered from a suspect are similar to each other.

In the United Kingdom jurisdiction system, the role of the forensic scientist and statistician is vital. The evidence that they evaluate can sometimes persuade a jury that the offender is innocent or guilty. Suppose a crime has been committed where an intruder has broken into a house through a window and blood is found on the floor. The forensic scientist will analyse the fragments of glass found on the floor by some means of a scientific method and compare these fragments to the type of glass in the window. The scientist would then enter the numerical values of the results they had found into a computer and the forensic statistician would then analyse this data by some appropriate statistical technique. The statistician would evaluate the data under two competing hypotheses: the prosecution proposition (H_p) and defence proposition (H_d). The prosecution proposition usually states that the trace evidence found at the crime scene and the trace evidence recovered from a suspect come from the same source. The defence proposition states that the trace evidence found at the crime scene and the trace evidence recovered from a suspect do not come from the same source. This is known as a source-level proposition.

Every proposition in any law, in any country in the world must have a logical relation to the circumstances of the case in hand. In the United Kingdom there is a hierarchy of propositions in the jurisdiction system, namely source, activity and offence level. The source level proposition was described above and it concerns matching of the evidence found at the crime scene to the evidence found on the suspect. The question of interest would be whether the trace evidence recovered from a suspect and the trace evidence that was found at the crime scene, originate from the same source.

In the example of the broken window, the following hypotheses could be formulated at the source level.

$\mathbf{H}_p$: The glass of window from the house that was broken into found on the

suspect matches that of the glass from the crime scene.

$\mathbf{H}_d$: The glass of the window from the house that was broken into found on the suspect does not match that of the glass from the crime scene.

The activity level proposition is slightly more involved than the source level. It requires the forensic scientist to look beyond the evidence which is presented in front of them. In the example of the broken window where there was a bloodstain on the floor, the question of interest to the scientist would be whether the bloodstain found on the floor was consistent with a broken window. These can be formulated as:

$\mathbf{H}_p$: Mr. X smashed the window and left the bloodstain on the floor.

$\mathbf{H}_d$: Someone other than Mr. X smashed the window and left the bloodstain on the floor.

The offence level propositions can be formulated as:

$\mathbf{H}_p$: Mr. X was at the scene of the crime.

$\mathbf{H}_d$: Someone other than Mr. X was at the scene of the crime.

Some propositions can be hard to evaluate, particularly from eye-witness statements. It is hard to evaluate them because a lot of the time they will not have specific numerical values attached to them. Consider the following scenario which was illustrated in [3]. An eye-witness might see a tall, blond haired man who had a tattoo on his neck and wearing a green jacket running away from a crime scene. It can be hard to make sense of portable evidence facts such as that the offender was seen to have blond hair and a green jacket because there are many different shades of blond hair and the colour green. Likewise there are many different styles and cuts of green jackets among the

population. It is hard to assign numerical values to these facts. The fact that the offender had a tattoo on his neck is quite an overwhelming piece of evidence since it could rule out a lot of the population of suspects because this feature is hard to get rid of at short notice. It can be slightly easier with the description of tall because the police could assemble an identity parade of men who are different heights and the witness could point to the man they believed was a similar height to the offender.

Although the previous example deals with evidence to which it is hard to attach numerical values, there are many types of evidence where this is straightforward. For example the elemental composition or the refractive index to a fragment of glass, the complex mixture of pigments, modifiers, extenders, and binders which are commonly found in paint, the chemical composition of fibres determined using infrared spectrophotometry are all numerical in nature and are therefore possible for a statistician to interpret when statistical techniques are used.

For the duration of this thesis, we will only be evaluating evidence under source-level propositions described earlier in Section 1.1. The example given was about glass where one might have measured the refractive index of fragments of the glass or the elemental composition of the glass. There will be numerical values attached to these measurements which will then enable a forensic statistician to employ statistical techniques. The techniques employed calculate the probability that the two sets of fragments were found to be similar assuming that these fragments had come from different sources. If the resulting probability is very low then the different-source proposition is deemed unlikely to be true and the weight of evidence is in favour of the two sets of fragments coming from the same source.

1.2 Statistical Approach

A common approach, which has been used by [13], [3] and [1] to name a few, for evaluating the likelihood of the trace evidence found at a crime scene and the trace evidence recovered from a suspect coming from the same source is the likelihood ratio [3]. This ratio is obtained using Bayes' theorem. The numerator of the likelihood ratio evaluates the probability of the evidence, E, assuming the prosecution proposition is true and the denominator of the likelihood ratio evaluates the evidence, E, assuming the defence proposition is true. The ratio can be considered to be a factor which converts prior odds in favour of the prosecution proposition to the posterior odds in favour of the defence proposition. Another factor which is taken into account when calculating the likelihood ratio is the background information, I, of the case in hand. The likelihood ratio can be formally written as:

$$\underbrace{\frac{Pr(H_p|E,I)}{Pr(H_d|E,I)}}_{\text{Posterior odds}} = \underbrace{\frac{Pr(E|H_p,I)}{Pr(E|H_d,I)}}_{\text{Likelihood ratio}} \times \underbrace{\frac{Pr(H_p|I)}{Pr(H_d|I)}}_{\text{Prior odds}} \tag{1.2.1}$$

The closer the likelihood ratio is to one, the less relevant the evidence but as the value of expression (1.2.1) increases the more weight given to the prosecution proposition that there is a common source for the trace evidence found at the crime scene and the trace evidence recovered from a suspect.
It can be quite hard for the jury members, who might not be statistically minded, to evaluate the weight of evidence given the value of the likelihood ratio which would be presented to them when the forensic scientist would be giving evidence. Table 1.1 given in [3] (page 107) gives some indication of how to interpret the likelihood ratio.

Weight of evidence	**Value of V**
Limited evidence to support H_p	LR $\leq 1 - 10$
Moderate evidence to support H_p	LR 10-100
Moderately strong evidence to support H_p	LR 100-1000
Strong evidence to support H_p	LR 1000-10000
Very strong evidence to support H_p	LR ≥ 10000

Table 1.1: Interpreting the likelihood ratio

1.3 Literature Review

The likelihood ratio defined in expression (1.2.1) was first brought to the forefront of forensic science by Dennis Lindley [13] who considered the problem of deciding whether two sets of measurements of trace evidence come from a common source. Lindley [13] proposed a solution to this in the case where the evidence is glass fragments and the collected univariate data are measurements of refractive indices. Assuming that n_1 measurements have been taken of the refractive index of the glass fragments found at the scene of the crime (control measurements) and n_2 measurements have been taken of the refractive index of the glass fragments which have been recovered from a suspect (recovered measurements), and two sources of variability exist: variability of the refractive index within object and variability of the refractive indices between objects, Lindley proposes two scenarios. The first scenario assumes that both the within-object and between-object distributions are normal. The second scenario continues to assume that the within-object distribution is normal but the between-object distribution is estimated non-parametrically. The second assumption might be more realistic since it is not always feasible to assume normality for the between-object distribution. More recently [1] explore whether two sets of measurements of trace evidence

come from a common source, by considering data coming from a multivariate distribution. In [1] a two-level random effects approach is used to analyse five outcome variables which are measurements of elemental composition of glass. One of the methods used in [1] to obtain the likelihood ratio for the data assumes that between-source variability is modelled by a multivariate normal distribution and another method models the between-source variability with a multivariate kernel density estimate. It is shown that both methods perform better than a univariate approach because they both allow for dependencies between variables whilst in the univariate approach the variables were looked at separately.

A three-level multivariate random effects approach is described in [2] in which there are again five variables of interest which are measurements of elemental composition of glass and the extra level of variation is the measurement error on the individual fragments in addition to the variability within each object and the variability between the objects. Both [2] and [1] are generalisations of Lindley's approach to multivariate data.

1.4 Overview of Thesis

In Chapter 2 we look at implementing Lindley's approach which assumes normality of the between-object distribution in addition to kernel density estimation, to univariate refractive index data.

Chapter 3 applies the multivariate normal model proposed by [1], which is essentially the multivariate version of Lindley's model, and multivariate kernel density estimation to two datasets: refractive index data and elemental composition data. The elemental composition dataset is considerably more complex the the refractive index dataset as it contains seven variables compared to two in the refractive index dataset.

Chapter 4 discusses the findings of the thesis and provides some thoughts of work which might be of future interest.

Chapter 2

Statistical models for the evaluation of evidence in the form of univariate data

In this chapter, two models are applied to univariate data. One assumes normality of the between-object distribution and the other estimates the between-object distribution by kernel density estimation. The two methods are compared in terms of simulations using experimental data which consists of the refractive index measured on various fragments of glass and are of known origin.

2.1 Control and Recovered Data

Suppose a crime was committed where a window was broken and some glass fragments were recovered from a suspect and analysed at a forensic laboratory where the refractive indices of the glass fragments were measured. Suppose that there are $n_1 \geq 1$ replicate measurements taken from the fragments of glass found at the crime scene and further suppose that they all come from source W_1. These measurements are referred to as the control data since a single source is known. Suppose that there are $n_2 \geq 1$ but not necessarily

the same as n_1 replicate measurements taken from the fragments of glass which were recovered from a suspect and these measurements are referred to as source W_2, which assumes that these measurements come from a single source. As defined in Section 1.1, the prosecution proposition states that W_1 and W_2 are the same source and the defence proposition states that W_1 and W_2 are not the same source.

2.2 Lindley's approach assuming normality of the between-object distribution

Lindley [13] proposed the following model for evaluation of glass evidence in the form of univariate data. Given n_1 measurements $(x_{11}, \ldots, x_{1n_1})$ of glass fragments found at the scene of the crime, W_1 and n_2 measurements $(x_{21}, \ldots, x_{2n_2})$ taken of glass fragments recovered from a suspect, W_2, let C be the event that the two sets of fragments come from the same source and $\overline{C}$ the conjugate event. Consider the sample means $\bar{X}_1 = \sum_{i=1}^{n_1} \frac{x_{1i}}{n_1}$ and $\bar{X}_2 = \sum_{i=1}^{n_2} \frac{x_{2i}}{n_2}$ the odds on C will be multiplied by the factor:

$$\frac{f(\bar{X}_1, \bar{X}_2|C)}{f(\bar{X}_1, \bar{X}_2|\overline{C})} \tag{2.2.1}$$

where f is the corresponding probability density. The numerator of (2.2.1) can be expressed as:

$$\int f(\bar{X}_1|\theta) f(\bar{X}_2|\theta) f(\theta) d\theta \tag{2.2.2}$$

because H_p denotes that θ_1 and θ_2 come from the same source, so θ is the common value of θ_1 and θ_2, where θ is a parameter in the joint distribution of $\bar{X}_1$ and $\bar{X}_2$. Since H_d denotes that the two fragments come from different sources, the denominator of (2.2.1) can be expressed as:

$$\int f(\bar{X}_1|\theta_1) f(\theta_1) d\theta_1 \int f(\bar{X}_2|\theta_2) f(\theta_2) d\theta_2 \tag{2.2.3}$$

which is the product of marginal distributions of $\bar{X}_1$ and $\bar{X}_2$ assuming independence. The distributions can be written formally as:

$$\bar{X}_1 \sim \mathrm{N}(\theta_1, \sigma_1^2),\ i = 1, \ldots, n_1$$

$$\bar{X}_2 \sim \mathrm{N}(\theta_2, \sigma_2^2),\ i = 1, \ldots, n_2$$

$$\theta_j \sim \mathrm{N}(\mu, \tau^2),\ j = 1, 2$$

where σ^2 is the within-group variance and σ_1^2 and σ_2^2 are defined to be:

$$\sigma_1^2 = \tau^2 + \frac{\sigma^2}{n_1} \tag{2.2.4}$$

$$\sigma_2^2 = \tau^2 + \frac{\sigma^2}{n_2} \tag{2.2.5}$$

and τ^2 is defined to be the between-group variance. Then, following Lindley [13], $(\bar{X}_1 - \bar{X}_2) \sim N(0, \sigma_1^2 + \sigma_2^2)$ independently of

$$Z = \frac{\sigma_2^2 \bar{X}_1 + \sigma_1^2 \bar{X}_2}{\sigma_1^2 + \sigma_2^2} \tag{2.2.6}$$

and the denominator of the likelihood ratio may be written as:

$$\frac{1}{2\pi\sigma_1\sigma_2} \exp\left[-\frac{(\bar{x_1} - \bar{x_2})^2}{2(\sigma_1^2 + \sigma_2^2)}\right] \exp\left[-\frac{(z - \mu)^2(\sigma_1^2 + \sigma_2^2)}{2\sigma_1^2\sigma_2^2}\right] \tag{2.2.7}$$

Under H_p, $\bar{X}_1$ and $\bar{X}_2$ both have means equal to μ, variances σ_1^2 and σ_2^2 and covariance τ^2 and the distribution of $(\bar{X}_1 - \bar{X}_2)$ is $\mathrm{N}(0, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2}))$. Define

$$W = \frac{m\bar{X}_1 + n\bar{X}_2}{n_1 + n_2} \tag{2.2.8}$$

so that $\mathrm{W} \sim \mathrm{N}(\theta_3, \sigma_3^2)$, where

$$\sigma_3^2 = \tau^2 + \frac{\sigma^2}{n_1 + n_2} \tag{2.2.9}$$

and $\theta_3 \sim \mathrm{N}(\mu, \tau^2)$. Also define

$$a^2 = \frac{1}{n_1} + \frac{1}{n_2} \tag{2.2.10}$$

Then $(\bar{X}_1 - \bar{X}_2)$ and W are independent [13] and the numerator can be expressed as:

$$\frac{1}{2\sigma\sigma_3} \exp\left[-\frac{(\bar{x}_1 - \bar{x}_2)^2}{2a\sigma^2}\right] \exp\left[-\frac{(w-\mu)^2}{2\sigma_3^2}\right] \tag{2.2.11}$$

The likelihood ratio is the ratio of (2.2.11) and (2.2.7) and has the form:

$$V = \frac{\sigma_1\sigma_2}{a\sigma\sigma_3} \exp\left[-\frac{(\bar{x}_1 - \bar{x}_2)^2\tau^2}{a^2\sigma^2(\sigma_1^2 + \sigma_2^2)}\right] \exp\left[-\frac{(w-\mu)^2}{2\sigma_3^2} + \frac{(z-\mu)^2(\sigma_1^2 + \sigma_2^2)}{2\sigma_1^2\sigma_2^2}\right]$$

(2.2.12)

and will be referred to from now on as the likelihood ratio from Lindley's model.

The natural logarithm of equation (2.2.12) is

$$\log(V) = \log\left[\frac{\sigma_1\sigma_2}{a\sigma\sigma_3}\right] - \frac{(\bar{x}_1 - \bar{x}_2)^2\tau^2}{a^2\sigma^2(\sigma_1^2 + \sigma_2^2)} - \frac{(w-\mu)^2}{2\sigma_3^2} + \frac{(z-\mu)^2(\sigma_1^2 + \sigma_2^2)}{2\sigma_1^2\sigma_2^2} \tag{2.2.13}$$

In order to implement Lindley's evaluating evidence in the form of univariate data in practice, one needs to estimate the parameters μ, σ^2 and τ^2 from a background database as described in Section 2.3.

2.3 Population database and parameter estimation

Let Ψ denote a population where the data consist of one variable which is measured n times on each of the m items. The background data can then be denoted as $\{y_{ij}\}$, $i = 1, \ldots, m; j = 1, \ldots, n$. In order to obtain the likelihood ratio for a pair of measurements estimates of μ, σ^2 and τ^2 are required. These estimates can be obtained from the population database. The overall mean,

μ, the within-group variance, σ^2, and between-group variance, τ^2, can be estimated by:

$$\hat{\mu} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij} \tag{2.3.1}$$

$$\hat{\sigma}^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{(y_{ij} - \bar{y}_i)^2}{mn - m} \tag{2.3.2}$$

$$\hat{\tau}^2 = \frac{\sum_{i=1}^{m}(y_i - \bar{y})^2}{m-1} - \frac{\hat{\sigma^2}}{m} \tag{2.3.3}$$

With the population database defined and parameter estimates obtained, we can proceed to conduct a simulation experiment to study the performance of Lindley's model which assumes that the between-object distribution is normal (2.2.12) in correctly identifying same-source and different-source pairs of measurements.

2.4 Same-source and different-source comparisons for an experimental dataset of refractive index values before and after annealing using Lindley's normal model

An experimental dataset supplied by G. Zadora was available which consisted of refractive index values measured on several fragments of glass. The refractive index measurements were taken from the fragments before they were subjected to an annealing process and again after the process. A more detailed account of the refractive index dataset can be found in Appendix A. The population database described in Appendix A will be used to obtain parameter estimates and also to supply "control" and "recovered" measurements pairs for a simulation experiment. What follows is a description of how these pairs were obtained.

Simulations were constructed to investigate same-source comparisons that compare the refractive index of two pairs of fragments from the same item with each other, and different-source comparisons that compared the refractive index of two different items with each other. Both simulations were based on the refractive index database described in the previous paragraph and use equation (2.2.12) to return the likelihood ratio.
The same-source comparisons were constructed in the following way:

- Each time the likelihood ratio was calculated for the i^{th} item from the background database, the i^{th} item was removed from the database and its mean and variance components were calculated from the remaining data using equations (2.3.1) - (2.3.3). Out of the four measurements for each object, the first two measurements were taken as the "control" set, (x_{11}, x_{12}), and the second two measurements were taken as the "recovered" set of measurements, (x_{21}, x_{22}). This then enabled the likelihood ratio, V to be calculated for the object using equation (2.2.12).

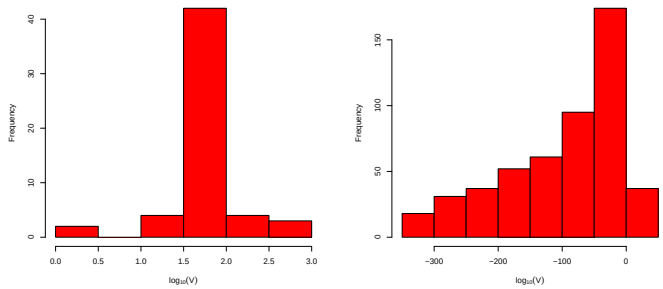
- This resulted in 55 values of the likelihood ratio, one for each of the objects in the database.

For same-source comparisons one would expect the log of the value of evidence, $\log_{10}(V)$, to be greater than 0 (which corresponds to a value of V greater than 1) which indicates that the two sets of fragments are more likely come from the same object.The proportion of false negatives for same-source comparisons is defined to be the percentage of comparisons with $V<1$ or equivalently $\log_{10}(V)<0$, and will be used as a measure of performance for each evidence evaluation method.
The different-source comparisons were set up in a similar way. Since there were 55 items and all possible comparisons were required, this resulted in $\binom{55}{2} = 1485$ possible comparisons. Each time the i^{th} and j^{th} items were compared, they were removed from the background database and the mean and

variance components were calculated from the remaining data using equations (2.3.1) - (2.3.3). In each comparison the four measurements from one object, $(x_{11}, \ldots, x_{14})$, the "control" set, were compared against the four measurements for the second object, $(x_{21}, \ldots, x_{24})$, the "recovered" set. The likelihood ratio , V, was calculated for each comparison using equation (2.2.12). For different-source comparisons one would expect $\log_{10}(V)$ to be less than 0 (which corresponds to a value of V less than 1) which would indicate that the two items are from different sources. The proportion of false positives for different-source comparisons is defined to be the percentage of comparisons with $V>1$ or equivalently $\log_{10}(V)>0$.

2.4.1 Simulation study results for refractive index before and after annealing using Lindley's approach

The simulation experiments described in Section 2.4 were implemented using the refractive index data described in Appendix A. The results for same-source and different-source comparisons for the refractive index before annealing are shown in Figure 2.1 as histograms of the logarithms base 10 of the likelihood ratio.



(a) Same-source comparisons (b) Different-source comparisons

Figure 2.1: $\log_{10}(V)$ values for same-source and different-source omparisons for refractive index before annealing using Lindley's approach

The proportions of false negatives and false positives are shown in the first row of Table 2.1. The fact that the proportion of false negatives for the same source comparisons is 0 suggests that equation (2.2.12) has been effective in correctly identifying same-source pairs of refractive index measurements. It has correctly identified that for every item, the two pairs of fragments do come from the same item. From Figure 2.1(a) for same-source comparisons most of the $\log_{10}(V)$ values are centered between 1.5 and 2 corresponding to the likelihood ratio values (V) between 32 and 100. Referring back to Table 1.1, these values suggest that there is moderate evidence to support the fact the two pairs of fragments come from the same source.

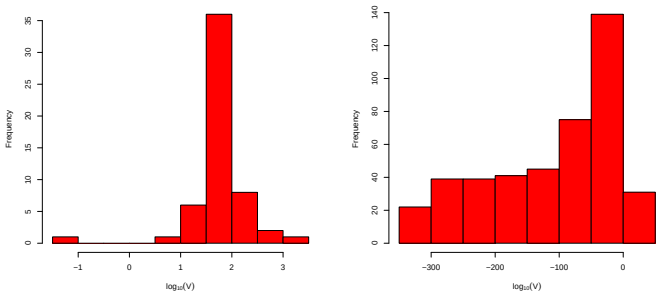
When looking at different-source comparisons the proportion of false positives in Table 2.1 is low. Only 37 out of 1485 comparisons were false positives which suggests that, again, equation (2.2.12) has been effective at identifying that two items do come from different sources. There are 980 comparisons which have a value of $\log_{10}(V)$ to be $-\infty$. This is extremely strong support for the different-source proposition. For the 37 misleading values of the

likelihood ratio, the highest value of $\log_{10}(V)$ is 2.1 which corresponds to a value of V of 126. When one refers back to Table 1.1, this suggests that there is moderately strong evidence to support the proposition that the two pairs of fragments come from the same source. However, there are only seven misleading values of $\log_{10}(V)$ greater than 1.9 (V=79) which suggests there is moderate evidence to support the proposition that the two pairs of fragments come from the same source. These few large misleading values of the likelihood ratio do cause some concern but the value of $\log_{10}(V)$ for the majority of the false positive comparisons is small enough to only lend weak support to the same-source proposition.

Variable Used	**False-negatives (%)**	**False-positives (%)**
Refractive index before annealing	0	2.5
Refractive index after annealing	1.8	2.1

Table 2.1: Table of misclassifications for refractive index before and after annealing using Lindley's model

The simulation experiment was repeated using the refractive indices after annealing. The results for same-source and different-source comparisons consisting of the logarithms (base10) of the likelihood ratio are shown in Figure 2.2 and the proportions of false positives and false negatives are shown in Table 2.1.



(a) Same-source comparisons (b) Different-source comparisons

Figure 2.2: $\log_{10}(V)$ values for same-source and different source omparisons for refractive index after annealing using Lindley's approach

Similarly to the refractive index before annealing, Figure 2.2(a) shows that most of the $\log_{10}(V)$ values are centered between 1.5 and 2 which suggests that the results are relatively in line with with those obtained for the refractive index before the annealing process shown in Table 2.1. Equation (2.2.12) has only returned one false negative item with a value of $\log_{10}(V)$=-1.3. This value has a corresponding value of V to be 0.05, which is equivalent to the pairs of measurements being 20 times more likely to come from different sources than from the same source. When one refers to Table 1.1, this is moderate evidence that the two pairs of measurements come from different sources. With only one false negative result, which is not too concerning, equation (2.2.12) has been effective in correctly identifying same-source pairs of refractive index measurements.

The proportion of false positives for different-source comparisons, located in the second row of Table 2.1, is also low which suggests that using the refractive index after annealing has been fairly effective with only 2.1% false positive comparisons. There are only two values of $\log_{10}(V)$ which are greater

than 2 (V=100), with the largest being V=182. There are 1054 comparisons which have a value of $\log_{10}(V)$ to be $-\infty$ which is equivalent to 71% of all comparisons. This is overwhelming support for the different-source proposition. These results are relatively in line with the different-source comparisons for the refractive index before annealing and show that equation (2.2.12) has been effective once again in identifying different-source comparisons of refractive index measurements.

It is fair to say that Lindley's model appears to perform satisfactorily since there is a very small number of false comparisons.

When one compares the results using the refractive index before the annealing process to those using the refractive index after annealing, there is only a marginal difference. This suggests that both the refractive index before and after annealing can be used to effectively test the same-source hypothesis and that it is reasonable to assume normality for the between-object distribution, at least in terms of the results of this simulation.

2.5 Kernel density estimation

When it becomes difficult to model data by standard distributions, the calculation of their probability density function proves useful. A popular and widely used method is kernel density estimation which estimates the probability density function in a non-parametric fashion from the data.

Kernel density estimation uses the basic principles of the histogram as its underlying theory. As explained in [3] (pages 330-338) a histogram consists of rectangular blocks where each block relates to its observations. Instead of placing a rectangular block to the corresponding observation, kernel density estimation places a probability density curve over the observation (quite often the normal distribution). This is known as the kernel density function. The curve is placed by centering it over the observation it relates to. The estimate of the probability density curve can be thought of as amalgamating all the

curves on the histogram into one. This is formally done by summing all the curves which relate to the observations in the data set and dividing by the total number of observations. The sum of all the curves divided by the number of observations has an area of 1 which makes it a probability density function which is non-negative. Figure 2.3 illustrates this where with eight randomly chosen points on the x-axis.

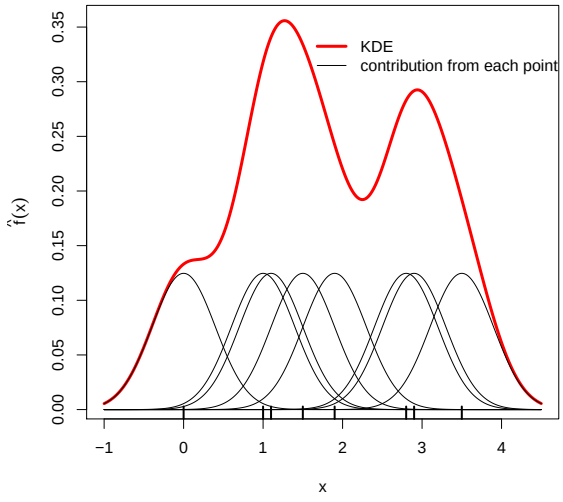


Figure 2.3: An illustration of kernel density estimation

To illustrate how kernel density estimation is simply an extension of the histogram, consider a histogram and a given origin, x_0, and bin width, h, where bins are defined to be the intervals of the histogram. The intervals can be represented by $[x_0 + mh, x_0 + (m+1)h]$, where m is a positive or negative number. The histogram is defined by:

$$\hat{f}(x) = \frac{1}{nh} \times (\text{number of } \mathrm{X}_i \text{ in the same bin as x}) \qquad (2.5.1)$$

The kernel density estimate has a similar formula given by:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{2.5.2}$$

where the K is the kernel function and h is the smoothing parameter. One can see that equation (2.5.2) is an extension of equation (2.5.1).

An important choice in kernel density estimation is the choice of the kernel, K. There are various different kernels which can be selected such as rectangular, triangular, Epanechnikov and Gaussian to name a few. Figure 2.3 illustrates what these kernels look like when they are centered around a single datapoint at zero.

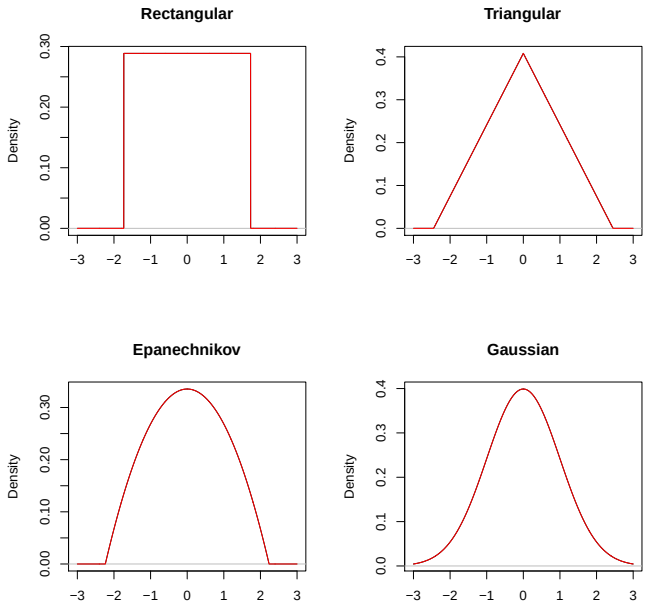


Figure 2.4: Different choices of kernels

In kernel density estimation the choice of the smoothing parameter (or bandwidth), h, is very important. If h is chosen to be too small the resultant curve

can be very spiky and there is too much detail to consider about the underlying distribution. If h is chosen to be too large the resultant curve is very smooth but to the point that quite a lot of information could be lost from the data set. One wants h to be the optimal size which captures most of the data but the curve is not too spiky nor too smooth.

2.5.1 Bandwidth selection

The question of how to make an appropriate choice for the bandwidth is one of the most extensive and controversial problems in the field of kernel density estimation. Several ideas and approaches have been considered over the years. Two popular methods are least-squares cross-validation [4], and Silverman's rule of thumb [18]. These techniques are relatively straightforward and can easily be applied to data. Bandwidth selection can be structured into two groups: subjective and machine-analysed. The subjective approach relies on using one's own eyes to see which bandwidth looks most appropriate for the dataset. For the machine-analysed approach the bandwidths are automatically chosen. Everything depends entirely on the data and one needs no prior experience to obtain a good fit. What follows is a description of machine analysed methods of obtaining optimal bandwidths.

When one chooses a kernel estimator, $\hat{f}$, to model data of the form $Y_1, \ldots, Y_n$ it is useful to assess the discrepancy between $\hat{f}$ and f. The most widely used way of doing this is by calculating the mean integrated square error (MISE) which is formally written as:

$$MISE(\hat{f}) = E\left[\int \left(\hat{f}(x) - f(x)\right)^2 dx\right] \tag{2.5.3}$$

The least-squares cross-validation method which was proposed by [4] intends to choose a value for h which will make the integrated square error as small as possible. The integrated square error is defined as follows:

$$\int (\hat{f}_h - f)^2 = \int \hat{f}_h^{\,2} - 2\int \hat{f}_h f + \int f^2 \tag{2.5.4}$$

Since the last term does not depend on h, it is sufficient just to keep an eye on the first two terms. The first term can be calculated straight from the data and the second term is the expected value of $\hat{f}_h(X)$ where X is a random variable. The expected value is therefore given by:

$$E\left[\hat{f}_h(X)\right] = \frac{1}{n}\sum_{i=1}^{n} \hat{f_{h,-i}}(X_i) \tag{2.5.5}$$

where

$$\hat{f_{h,-i}}(x) = \frac{1}{n-1}\sum_{j=1, i\neq j}^{n} K_j(x - X_j) \tag{2.5.6}$$

denotes a *leave-one-out* estimator where datapoint i is left out of the calculation of $\hat{f_{h,-i}}(x)$. This ensures that the observations for $\hat{f_{h,-i}}(x)$ are independent of X_i. The integrated square error (which is the criterion function that one seeks to minimize with respect to h) can then be re-written as:

$$ISE(h) = \int \hat{f}_h^2 dx - 2E\left[\hat{f}_h(X)\right] + \int f^2 dx \tag{2.5.7}$$

As stated before, the third term does not depend on h so there is no need to worry about it. It is now possible to insert the expression from equation (2.5.5) into equation (2.5.7) and this will produce the cross-validation criterion.

$$LSCV(h) = \int \hat{f}_h^2 dx - \frac{2}{n(n-1)}\sum_i \sum_{j\neq i} K_h(X_i - X_j) \tag{2.5.8}$$

The remaining step is to replace the first term with a sum rather than an integral. It can be shown [11] that:

$$\int \hat{f}_h^2 dx = \frac{1}{n^2 h}\sum_i \sum_j K \star K\left(\frac{X_j - X_i}{h}\right) \tag{2.5.9}$$

where the convolution $K \star K(u)$ is defined as:

$$K \star K(u) = \int K(u-v)K(v)dv \tag{2.5.10}$$

The resulting least-squares cross-validation criterion to be minimized is thus:

$$LSCV(h) = \frac{1}{n^2 h}\sum_i \sum_j K \star K\left(\frac{X_j - X_i}{h}\right) - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x_i) \tag{2.5.11}$$

and the value of $\hat{h}$ that minimizes $LSCV(h)$ is the resulting bandwidth. The $LSCV(h)$ method has one drawback: if there are too many equal values in the data, an overall $LSCV(h)$ will be found at $h = 0$, which would be an unreasonable choice.

Biased cross-validation is similar to least-squares cross-validation but the main difference is that the minimization is based on the asymptotic mean integrated square error (AMISE), formally defined as:

$$AMISE(h) = \frac{R(K)}{nh} + \frac{1}{4}\sigma_K^4 h^4 R(f^{''}) \tag{2.5.12}$$

where $R(K) = \int_{-\infty}^{\infty} K(x)^2 dx$ of the kernel, K, h is the smoothing parameter, n is the sample size, σ_K^4 is the variance of the kernel, K and $f^{''}$ is the second derivative of the underlying density.

The biased cross-validation criterion is derived in the same fashion as $LSCV(h)$ and [16] show that the resulting criterion is given by:

$$BCV(h) = \frac{R(K)}{nh} + \frac{1}{4}\sigma_K^4 h^4 \hat{R}_1 \tag{2.5.13}$$

where $\hat{R}_1 = R(\hat{f}^{''}) = R(\hat{f}^{''}) - \dfrac{R(K^{''})}{nh^5}$ and $\hat{f}^{''}$ is the second derivative of the univariate kernel density estimator with kernel K. As the name suggests the resulting bandwidth is biased and has smaller variance than $LSCV(h)$.

Another popular method of choosing h was developed by Silverman [18] who obtained the following rule of thumb formula for h, under the assumption of normality:

$$h_{opt} = 1.06\sigma n^{-\frac{1}{5}} \tag{2.5.14}$$

This is a quick way of choosing the smoothing parameter which can be done by substituting $\hat{\sigma^2}$ for σ^2, the sample variance. This will tend to oversmooth with multi-modal or skewed data. Another possibility would be to have a more robust estimate of σ, the interquartile range, $\hat{R}$. Because of the fact that R is approximately 1.34 times as high as the standard deviation for normal densities, Silverman adapted equation (2.5.14) to:

$$h_{opt} = 0.79Rn^{-\frac{1}{5}} \tag{2.5.15}$$

Again, this method does tend to oversmooth with multi-modal distributions. Silverman [18] then proposes another formula for h which binds equations (2.5.14) and (2.5.15) together to make the best of both worlds.

$$h = 0.9An^{-\frac{1}{5}} \tag{2.5.16}$$

where A=min(standard deviation, interquartile range/1.34) and n is the sample size. Equation (2.5.16) is thought to be a good rule of thumb because the expression can cope well with unimodal densities and it should cope reasonably well with moderately bimodal densities too, although it can tend to over-smooth in practice.

Another popular and widely used method is that by Sheather and Jones [17] which minimizes the criterion

$$SJ(h) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_h(x_i - x_j) \tag{2.5.17}$$

where K and h respectively are a kernel and bandwidth. This method tends to require extra computation time because of the complex connection of h on the left and right hand side of the equation. This method is also considered to be a good compromise between the Rule of Thumb and least-squares cross-validation.

2.6 Application of kernel density estimation to simulation study for refractive index data

In addition to normal estimation according to Lindley's model described in Section 2.2, kernel density estimation is also considered for the between-object distribution of the refractive indices in the experimental database of Section 2.3. The method assumes the background data, D, has the form $(y_1, \ldots, y_m)$. In the refractive index database y_i would be the mean refractive index for the i^{th} item in the database, $i = 1, \ldots, m$. The within-group variance σ^2 is calculated in the same way as Lindley's [13] approach. The sample variance, s^2, of the refractive indices from different groups is calculated by:

$$s^2 = \sum_{i=1}^{m} \frac{(y_i - \bar{y})^2}{m-1} \tag{2.6.1}$$

where $y_i, i = 1, \ldots, m$ are the background database and m is the number of items in the database.

The kernel, K, is generally chosen to be a unimodal probability density which is symmetric around zero. When the kernel takes a normal distribution, it is of the form:

$$K(\theta|y_i, h) = \frac{1}{hs\sqrt{2\pi}} \exp\left[-\frac{(\theta - y_i)^2}{2h^2 s^2}\right] \tag{2.6.2}$$

where the mean is y_i and the variance is $h^2 s^2$. Then the estimate of the probability density function, $\hat{f}(\theta|D, h)$, at point y_i is given by:

$$\hat{f}(\theta|D, h) = \frac{1}{m} \sum_{i=1}^{m} K(\theta|y_i, h) \tag{2.6.3}$$

Using a normal kernel density estimate in [22] it is shown that the numerator of the likelihood ratio (1.2.1) can be expressed as:

$$\frac{1}{\sqrt{2\pi u_0^2}} \exp\left\{-\frac{(\bar{x_1}-\bar{x_2})^2}{2u_0^2}\right\}$$
$$\times \frac{1}{m}\sum_{i=1}^{m} \frac{1}{\sqrt{2\pi\frac{\sigma^2}{n_1+n_2}+h^2s^2}} \exp\left\{-\frac{(w-y_i)^2}{2\left(\frac{\sigma^2}{n_1+n_2}+h^2s^2\right)}\right\} \tag{2.6.4}$$

where $u_0 = \sigma^2\left(\frac{1}{m}+\frac{1}{n}\right)$ and $\bar{y}_i$ is the mean of the i^{th} item. Similarly, [22], the denominator of the likelihood ratio (1.2.1) can be expressed as:

$$\frac{1}{\sqrt{2\pi\left(\frac{\sigma^2}{n_1}+h^2s^2\right)}}\frac{1}{m}\sum_{i=1}^{m}\exp\left\{-\frac{(\bar{x}_1-y_i)}{2\left(\frac{\sigma^2}{n_1}+h^2s^2\right)}\right\}$$
$$\times \frac{1}{\sqrt{2\pi\left(\frac{\sigma^2}{n_2}+h^2s^2\right)}}\frac{1}{m}\sum_{i=1}^{m}\exp\left\{-\frac{(\bar{x}_2-y_i)}{2\left(\frac{\sigma^2}{n_2+h^2s^2}\right)}\right\} \tag{2.6.5}$$

Equations (2.6.4) and (2.6.5) can be simplified to obtain the likelihood ratio:

$$V = \frac{K\exp\left[-\frac{(\bar{x}_1-\bar{x}_2)^2}{2u_0}\right]\sum_{i=1}^{m}\exp\left[-\frac{(n_1+n_2)(w-y_i)^2}{2[\sigma^2+(n_1+n_2)s^2h^2]}\right]}{\sum_{i=1}^{m}\exp\left[-\frac{n_1(\bar{x}_1-y_i)^2}{2(\sigma^2+n_1s^2h^2)}\right]\sum_{i=1}^{m}\exp\left[-\frac{n_2(\bar{x}_2-y_i)^2}{2(\sigma^2+n_2s^2h^2)}\right]} \tag{2.6.6}$$

where

$$K = \frac{p\sqrt{n_1+n_2}\sqrt{\sigma^2+n_1s^2h^2}\sqrt{\sigma^2+n_2s^2h^2}}{a\sigma\sqrt{n_1n_2}\sqrt{\sigma^2+(n_1+n_2)s^2h^2}} \tag{2.6.7}$$

as shown in [3].

2.6.1 Bandwidth selection for refractive index before and after annealing

The bandwidth selection methods described in Section 2.4.1 were implemented for the refractive index data described in Section 2.3 where each of the m items had a mean $\bar{y}_i$. Kernel density estimates of the between-object distributions of refractive index before annealing, with various bandwidth choices and Gaussian kernels are shown in Figure 2.5. These were obtained considering the 55 object means as the data. The plots were obtained from R [14] using the `density` function found in the `MASS` [19] package.

Figure 2.5 shows that there is not much difference between the four methods of automatically choosing the bandwidth. By inspection, the Sheather-Jones and Least-Squares Cross-Validation methods appear to capture the shape of the data slightly better than the other two but not by a considerable amount. All four methods will be considered in the simulations for same-source and different-source comparisons.

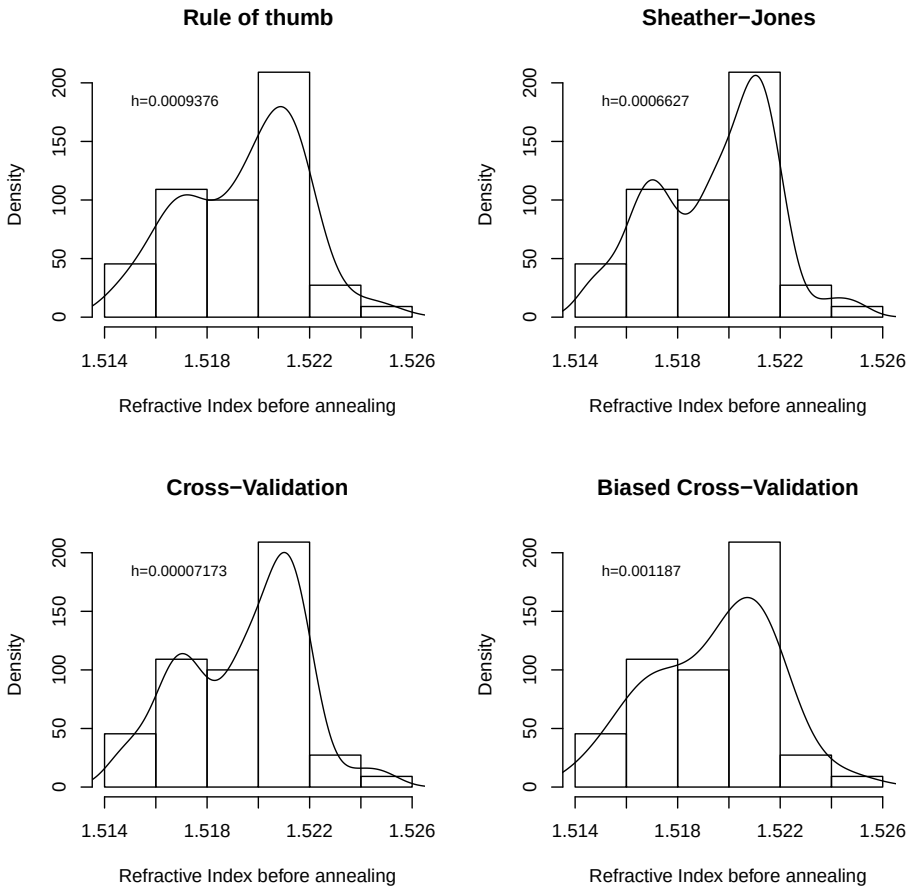


Figure 2.5: Comparing different bandwidths in kernel density estimation of the between-object distribution for refractive index before annealing

It would also be useful to look at the bandwidth methods implemented for the refractive index after annealing data. Figure 2.6 clearly shows that there is not much variation between the four methods of bandwidth selection. As for the bandwidth selection for refractive index before annealing, it appears perfectly valid to use any one of the bandwidth selection methods to estimate the between-object distribution.

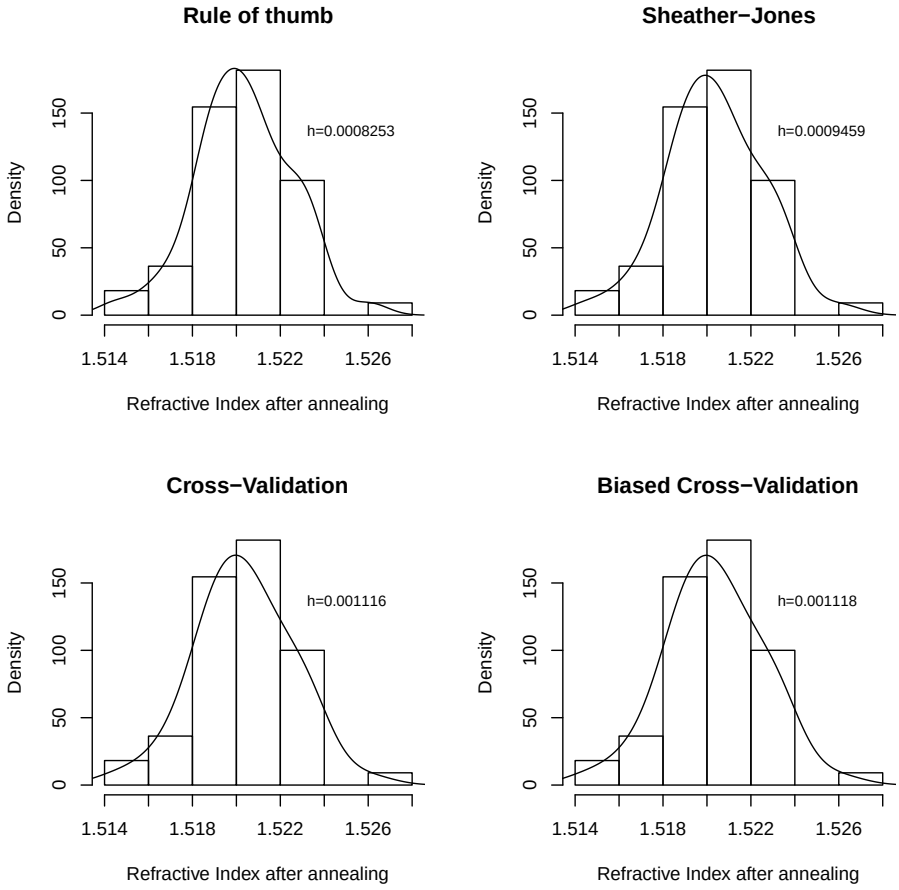


Figure 2.6: Comparing different bandwidths in kernel density estimation of the between-object distribution for refractive index after annealing

2.6.2 Simulation study results for refractive index before and after annealing using kernel density estimation

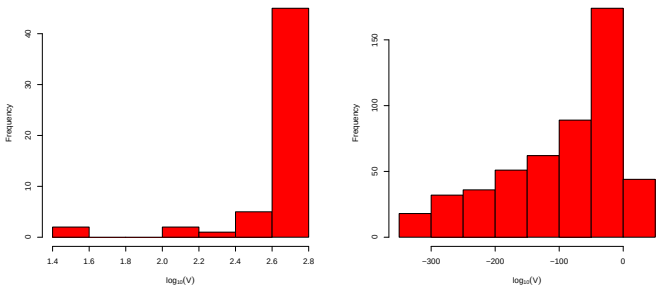
Same-source and different-source comparisons were implemented for the refractive index data in the simulation experiments similar to the one described in Section 2.4 where the likelihood ratio for each comparison was obtained using kernel density estimation as given in equation (2.6.6). The histograms

of same-source and different-source comparisons to logarithm (base 10) of likelihood ratios for comparisons for the refractive indices before annealing is shown in Figure 2.7 and the summary of results shown in the first row of Table 2.2.

Silverman's method is presented for illustrative purposes because there was no difference between the four methods of bandwidth selection in terms of the percentages of false positives and false negatives. Figure 2.7(a) shows that most of the $\log_{10}(V)$ values lie between 2.6 and 2.8 which have corresponding values of V to be between 398 and 631 and when one refers back to Table 1.1, this suggests there is moderately strong evidence to support the proposition that the two sets of measurements come from the same source. The other three methods of bandwidth selection produced more larger values of $\log_{10}(V)$ between 2.8 and 3.0 than Silverman's method, which might suggest that Silverman's method is slightly more conservative than the other three. The false negative rate of 0% suggests that equation (2.6.6) has performed extremely well in identifying that two sets of measurements come from the same source.

Variable Used	**False-negatives (%)**	**False-positives(%)**
R.I. before annealing	0	3.0
R.I. after annealing	1.8	2.6

Table 2.2: Table of misclassifications for refractive index before and after annealing using kernel density estimation and implementing all four methods of bandwidth selection described in Section 2.5.1



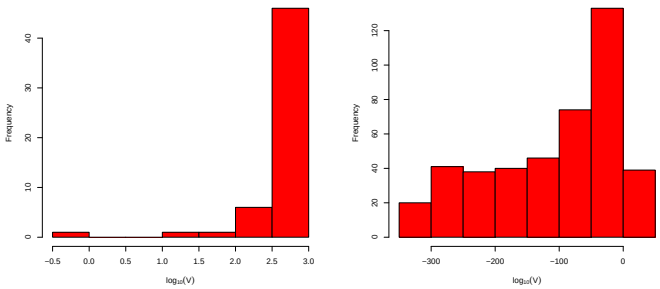
(a) Same-source comparisons (b) Different-source comparisons

Figure 2.7: $\log_{10}(V)$ values for same-source and different-source comparisons for refractive index before annealing using kernel density estimation for the between-object distribution

There were 971 different-source comparisons which had a value of $\log_{10}(V)$ to be $-\infty$ which suggests there is overwhelming support for different-source proposition. The fairly low false positive rate of 3% suggests that, again for refractive index before annealing, equation (2.6.6) has performed extremely well.

Same-source and different-source comparisons were also implemented for the refractive index data after annealing. The resulting histograms of logarithm (base 10) ratios are shown in Figure 2.8 and the misclassifications are shown in the second row of Table 2.2.

There was no difference between the four methods of selecting the bandwidth, so Silverman's method is shown in Figure 2.8 for illustrative purposes. All four methods returned a false negative result for item 12, classifying the two sets of fragments as coming from different sources. With a small false positive rate of 1.8%, equation (2.6.6) has performed very well.



(a) Same-source comparisons (b) Different-source comparisons

Figure 2.8: $\log_{10}(V)$ values for same-source and different-source omparisons for refractive index after annealing using kernel density estimation for between-object distribution

Looking at Figure 2.8(b), there are 1054 comparisons which have a value of $\log_{10}(V)$ (equivalent to 71% of all possible comparisons) to be $-\infty$ which suggests there is overwhelming evidence that the two pairs of measurements come from different sources. At the other end of the scale, the highest value of the false positives is 3 which has an equivalent value of V to be 1000. This is only equivalent to moderate evidence to support the same-source proposition which is not too concerning. It is plausible to say that equation (2.6.6) has performed well once again.

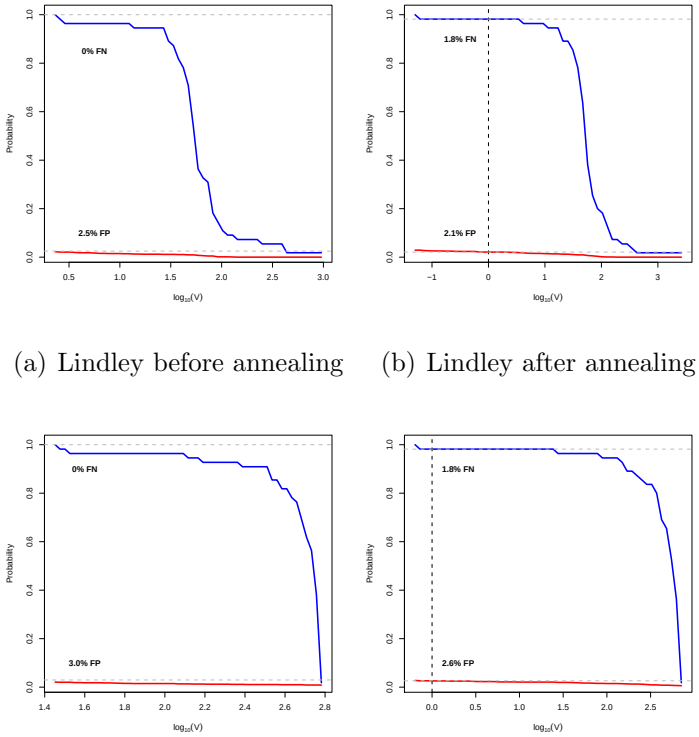
When one compares the results of comparisons using the refractive index before annealing to the refractive index after annealing using kernel density estimation, there are very few differences. It is plausible to use either the refractive index before annealing or after annealing to test the same-source and different-source hypothesis and it is reasonable to use any of the four methods of bandwidth selectors.

2.7 Tippett plots for univariate refractive index dataset

The Tippett plot is a diagnostic plot which contains two curves, one for the log likelihood ratios for the same-source comparisons (H_p) and one for the different-source comparisons (H_d). Tippett plots are commonly used for interpretation of evidence in forensic speaker recognition. The separation between the two curves, where the vertical dashed line at zero lie, is indicative of how well the plot has discriminated between comparisons which correspond to H_p and H_d for a given variable. The blue line in the Tippett plots represent the log likelihood ratios for H_p and the red line in each of the plots represent the log likelihood ratio for H_d.

The Tippett plots in Figure 2.9 are for the univariate refractive index dataset when one has employed Lindley's method and kernel density estimation for the between-object distribution.

For every plot in Figure 2.9, the separation between the lines for same-source and different-source comparisons is fairly wide which suggests that both Lindley's normal approach and kernel density estimation to the between-object distribution have been successful in discriminating same-source and different-source comparisons.



(a) Lindley before annealing (b) Lindley after annealing

(c) KDE before annealing (d) KDE after annealing

Figure 2.9: Tippett plots for univariate refractive index dataset using both Lindley's normal approach and kernel density estimation to estimate the between-object distribution

2.8 Summary

In this chapter, a two-level random effects model was applied to evaluating evidence in the form of univariate refractive index data.

Two approaches to estimating the between-object distribution were explored, normal and kernel density estimation. Both methods performed extremely well as they resulted in very low false-negative rates for the same-source comparisons and very low false-positive rates for the different-source comparisons. Very low false rates suggest that both methods have been extremely effective

at identifying same and different-source pairs of refractive index measurements.

The fact that both methods for estimating the between-object distribution have been extremely effective suggests that they are both plausible to use for refractive index data, however it may be simpler to use Lindley's normal approach in practice than the more involved kernel density estimation method. To use Lindley's method in practice, one only needs to estimate μ, σ^2 and τ^2, whereas for kernel density estimation one needs the aforementioned estimates as well as selecting a kernel and choosing from a vast array of bandwidth selection methods. It is recommended to use a normal model for the between-object distribution of refractive index since it is simple to use. Although the false rates are very similar for both the refractive index before annealing and the refractive index after annealing, if one had to use one variable, it is recommended to use a normal model for the between-object distribution with the refractive index before annealing as this variable has the lowest total of false rates.

Chapter 3

Statistical models for evaluation of evidence in the form of multivariate data

In this chapter, the random effects models presented in chapter 2 are extended to a multivariate setting. It is of interest to model data in a multivariate as well as univariate form. In the refractive index example for instance this is possible since there are two variables of interest, namely the refractive index before annealing and the refractive index after annealing. These variables might not be independent and it is necessary to allow for dependence between them. We apply multivariate random effects models both using between-object normality and kernel density estimation to two databases; the refractive index dataset described in Appendix A analysed in Chapter 2 and an elemental composition dataset which is described in Appendix B. Simulation experiments are conducted to assess the performance of these methods in evaluating evidence in the form of multivariate data.

3.1 The multivariate normal model

As in Chapter 2, consider two sets of measurements but now these are multivariate in nature. Consider a vector of n_1 control measurements $\mathbf{x}_1 = (\mathbf{x}_{11}, \ldots, \mathbf{x}_{1p})$ and a vector of n_2 recovered measurements $\mathbf{x}_2 = (\mathbf{x}_{21}, \ldots, \mathbf{x}_{2p})$. Their means are $\bar{\mathbf{x}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{x}_{1i}$ and $\bar{\mathbf{x}}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2}\mathbf{x}_{2i}$. The control and recovered measurements are assumed to have normal distribution with means $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ respectively. The within-group variance-covariance matrices are assumed be of form: $\mathbf{D}_1 = \frac{1}{n_1}\mathbf{U}$ such that $\bar{\mathbf{x}}_1 \sim N(\boldsymbol{\theta}_1, \mathbf{D}_1)$ and $\mathbf{D}_2 = \frac{1}{n_2}\mathbf{U}$ such that $\bar{\mathbf{x}}_2 \sim N(\boldsymbol{\theta}_2, \mathbf{D}_2)$.
Under the assumption of between-group normality, let the between-object variance covariance matrix be $\mathbf{C}$, so that

$$\bar{\mathbf{x}}_1 \sim N(\boldsymbol{\mu}, \mathbf{C} + \mathbf{D}_1)$$

$$\bar{\mathbf{x}}_2 \sim N(\boldsymbol{\mu}, \mathbf{C} + \mathbf{D}_2)$$

For the numerator of the likelihood ratio, $\boldsymbol{\theta}_1$=$\boldsymbol{\theta}_2$=$\boldsymbol{\theta}$ and the numerator has the form

$$\int_\theta f(\bar{\mathbf{x}}_1|\boldsymbol{\theta}, \mathbf{D}_1) f(\bar{\mathbf{x}}_2|\boldsymbol{\theta}, \mathbf{D}_2) f(\boldsymbol{\theta}|\boldsymbol{\mu}, \mathbf{C}) d\boldsymbol{\theta} \tag{3.1.1}$$

The three instrumental parts to equation (3.1.1) are all multivariate normals and their expressions are substituted into the probability density function for the multivariate normal. The resulting numerator is then given by [1]:

$$|2\pi\mathbf{U}|^{-0.5(n_1+n_2)}|2\pi\mathbf{C}|^{-0.5}|2\pi\left[(n_1+n_2)\mathbf{U}^{-1}+\mathbf{C}^{-1}\right]^{-1}|^{0.5}\times\exp\left[-\frac{1}{2}(H_1+H_2+H_3)\right] \tag{3.1.2}$$

where

$$H_1 = \sum_{l=1}^{2} trace(\mathbf{S}_l\mathbf{U}^{-1}) \tag{3.1.3}$$

$$H_2 = (\bar{\mathbf{x}}^* - \boldsymbol{\mu})^T \left(\frac{\mathbf{U}}{n_1 + n_2} + \mathbf{C} \right)^{-1} (\bar{\mathbf{x}}^* - \boldsymbol{\mu}) \tag{3.1.4}$$

$$\bar{\mathbf{x}}^* = \frac{n_1 \bar{\mathbf{x}}_1 + n_2 \bar{\mathbf{x}}_2}{n_1 + n_2} \tag{3.1.5}$$

$$\mathbf{S}_l = \sum_{j=1}^{n} (\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)(\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)^T \tag{3.1.6}$$

where $j = 1, \ldots, n$ are the number of measurements per object, $l = 1, 2$ and $\mathbf{x}_{lj}$ are p-dimensional vectors.

The expression for equation (3.1.2) is simplified slightly by [22] to:

$$\begin{aligned} \text{numerator} = (2\pi)^{-p} |\mathbf{D}_1 + \mathbf{D}_2|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T (\mathbf{D}_1 + \mathbf{D}_2)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \right] \\ \times \left| \frac{\mathbf{U}}{n_1 + n_2} \right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} (\bar{\mathbf{x}}^* - \boldsymbol{\mu})^T (\frac{\mathbf{U}}{n_1 + n_2})^{-1} (\bar{\mathbf{x}}^* - \boldsymbol{\mu}) \right] \end{aligned} \tag{3.1.7}$$

For the denominator $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ and $\bar{\mathbf{x}}_\mathbf{1}$ and $\bar{\mathbf{x}}_\mathbf{2}$ are assumed to be independent as they are believed to come from different sources. The denominator thus takes the form:

$$\int_\theta f(\bar{\mathbf{x}}_1 | \boldsymbol{\theta}, \mathbf{U}) f(\boldsymbol{\theta} | \boldsymbol{\mu}, \mathbf{C}) d\boldsymbol{\theta} \int_\theta f(\bar{\mathbf{x}}_2 | \boldsymbol{\theta}, \mathbf{U}) f(\boldsymbol{\theta} | \boldsymbol{\mu}, \mathbf{C}) d\boldsymbol{\theta} \tag{3.1.8}$$

As for the numerator, the four component parts of (3.1.8) are substituted by the probability density function for the multivariate normal and following [1] the resulting denominator is defined to be:

$$|2\pi\mathbf{C}|^{-0.5} |2\pi(n_1\mathbf{U}^{-1} + \mathbf{C}^{-1})^{-1}|^{0.5} |2\pi(n_2\mathbf{U}^{-1} + \mathbf{C}^{-1})^{-1}|^{0.5} \times \exp\left[-\frac{1}{2} (H_4 + H_5) \right] \tag{3.1.9}$$

where

$$H_4 = (\boldsymbol{\mu} - \boldsymbol{\mu}^*)^T \left[(\mathbf{D}_1 + \mathbf{C})^{-1} + (\mathbf{D}_2 + \mathbf{C})^{-1} \right] (\boldsymbol{\mu} - \boldsymbol{\mu}^*) \tag{3.1.10}$$

$$H_5 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T(\mathbf{D}_1 + \mathbf{D}_2 + 2\mathbf{C})^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \tag{3.1.11}$$

$$\boldsymbol{\mu}^* = \left[(\mathbf{D}_1 + \mathbf{C})^{-1} + (\mathbf{D}_2 + \mathbf{C})^{-1}\right]^{-1}\left[(\mathbf{D}_1 + \mathbf{C})^{-1}\bar{\mathbf{x}}_1 + (\mathbf{D}_2 + \mathbf{C})^{-1}\bar{\mathbf{x}}_2\right] \tag{3.1.12}$$

A simplified version of expression (3.1.9) given by [22] is:

$$\begin{aligned}\text{denominator} = (2\pi)^{-\frac{p}{2}}|\mathbf{D}_1 + \mathbf{C}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\bar{\mathbf{x}}_1 - \boldsymbol{\mu})^T(\mathbf{D}_1 + \mathbf{C})^{-1}(\bar{\mathbf{x}}_1 - \boldsymbol{\mu})\right] \\ \times |\mathbf{D}_2 + \mathbf{C}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\bar{\mathbf{x}}_2 - \boldsymbol{\mu})^T(\mathbf{D}_2 + \mathbf{C})^{-1}(\bar{\mathbf{x}}_2 - \boldsymbol{\mu})\right]\end{aligned} \tag{3.1.13}$$

The likelihood ratio is given by the ratio of (3.1.7) and (3.1.13):

$$V = \frac{\text{numerator}}{\text{denominator}} \tag{3.1.14}$$

In order to implement the multivariate normal model and subsequent multivariate models for evaluating evidence in the form of multivariate data in practice, we need to estimate the parameters $\boldsymbol{\mu}$, $\mathbf{U}$ and $\mathbf{C}$ from a background database, as described in Section 3.2.

3.2 Population database for multivariate data

Let Ψ denote a population where the data consist of p variables of interest and are measured n times on each of the m items. The background data can then be denoted as $\mathbf{y}_{ij} = (y_{ij1}, \ldots, y_{ijp})^T; i = 1, \ldots, m; j = 1, \ldots, n$. In order to obtain the likelihood ratio for a pair of measurements which come from a multivariate database one needs to estimate $\boldsymbol{\mu}$, $\mathbf{U}$ and $\mathbf{C}$. These estimates can be made from the background database. The mean $\boldsymbol{\mu}$ is the mean vector over the p variables. The mean vector is estimated as follows:

$$\hat{\boldsymbol{\mu}} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{y}_{ij} \tag{3.2.1}$$

There are two sources of variation within the parameters being estimated. Firstly, there is variation between the replicates within each item and it is assumed that it is constant variation and is normally distributed. There is also variation between items, as in the univariate case, and this variation is similarly assumed to be normally distributed or can be estimated by kernel density estimation.

The within-group variance covariance matrix, $\mathbf{U}$, is estimated from the population database by:

$$\hat{\mathbf{U}} = \frac{\mathbf{S}_w}{N-m} \tag{3.2.2}$$

where

$$\mathbf{S}_w = \sum_{i=1}^{m}\sum_{j=1}^{n}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)^T \tag{3.2.3}$$

and $\bar{\mathbf{y}}_i = \frac{1}{n}\sum_{j=1}^{n}\mathbf{y}_{ij}$ is the mean for the i^{th} item.

The between-group variance covariance matrix, $\mathbf{C}$, is estimated in a similar way from the population database:

$$\hat{\mathbf{C}} = \frac{\mathbf{S}^*}{m-1} - \frac{\mathbf{S}_w}{n(N-m)} \tag{3.2.4}$$

where

$$\mathbf{S}^* = \sum_{i=1}^{m}(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})^T \tag{3.2.5}$$

The resulting $\hat{\mathbf{U}}$ and $\hat{\mathbf{C}}$ covariance matrices will be $p \times p$ since there are p variables involved. The diagonal terms in the matrices indicate the variance of the p^{th} variable. The off-diagonal entries describe the covariance which is a measure of association between the two variables that makes them statistically dependent.

With the multivariate population database defined, we can proceed to conduct a simulation experiment to study the performance of the multivari-

ate normal model and any subsequent models in correctly identifying same-source and different-source pairs of measurements. The simulation experiments are similar to those described in Section 2.4 but use the multivariate version of the population database described in Section 3.2 and the likelihood ratio equation obtained by dividing equation (3.1.7) by equation (3.1.13) with parameters estimates for $\boldsymbol{\mu}$, $\mathbf{U}$ and $\mathbf{C}$ given by expression (3.1.14).

3.3 Multivariate kernel density estimation

The univariate case for kernel density estimation can be extended when one wants to account for possible dependencies between variables. The monographs of [5] and [18] provide an overview of the research which has already been carried out in the field of multivariate kernel density estimation.
The general form of the p-dimensional multivariate kernel density estimator is given by

$$\begin{aligned}\hat{\mathbf{f}}_{\mathbf{H}}(x) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\det(\mathbf{H})}\mathcal{K}\left\{\mathbf{H^{-1}}(\mathbf{x}-\mathbf{X_i})\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathcal{K}_{\mathbf{H}}(\mathbf{x}-\mathbf{X_i}) \qquad (3.3.1)\end{aligned}$$

where $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{ip})$, $\mathbf{H}$ is the bandwidth matrix and $\mathcal{K}$ is the multiplicative kernel function. The kernel function can be derived simply by either multiplying the univariate kernels together or by "rotating" the univariate kernel in p-dimensional space, which is known as a radially symmetric kernel. As for the univariate case of kernel density estimation, there are different choices of kernel such as Gaussian, Epanechnikov, triangular and rectangular which were all mentioned in Section 2.5. The most common choice in multivariate kernel density estimation is the Gaussian. The multivariate Gaussian kernel has the form:

$$K_{\mathbf{H}}(\mathbf{x}) = 2\pi^{-\frac{p}{2}}\exp(\mathbf{x}^T\mathbf{x}) \qquad (3.3.2)$$

where $\mathbf{x}=(x_1,\ldots,x_p)$.

The choice of the bandwidth matrix, as in the univariate case, is an important one. Two common choices of bandwidth matrices are the diagonal and the unconstrained matrix. The diagonal bandwidth matrix takes the form:

$$\mathrm{H}= diag(\mathrm{h}_1, h_2, \ldots, h_p)= \begin{bmatrix} h_1 & 0 & \cdots & 0 \\ 0 & h_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & h_p \end{bmatrix} \tag{3.3.3}$$

The diagonal entries relate to which bandwidth selection method has been chosen. The unconstrained bandwidth matrix which has no restrictions on **H** is defined to be:

$$\mathrm{H}= \begin{bmatrix} h_1 & h_{12} & \cdots & h_{1p} \\ h_{12} & h_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & h_{p-1,p} \\ h_{1p} & \cdots & h_{p-1,p} & h_p \end{bmatrix} \tag{3.3.4}$$

where **H** is a positive definite symmetric matrix and the off-diagonal entries allow the kernels to have an arbitrary orientation. Alternatively, the bandwidth matrix can take the form of $h^2\mathbf{C}$, where h is a single value and C is the variance-covariance matrix of the data, i.e. the resulting **H** is a multiple of the variance-covariance matrix.

What follows are descriptions of commonly used multivariate bandwidth selectors.

3.4 Bandwidth selection for multivariate kernel density estimation

Härdle and Müller [12] outline a proof of the least-squares cross-validation method for the multivariate case. As for the univariate case the method for least-squares cross-validation, $LSCV(\mathbf{H})$, aims to choose an optimal bandwidth matrix which will minimize the integrated square error (ISE).

$$ISE(\mathbf{H}) = \int \hat{f}^2_{\mathbf{H}} dx - 2\int \hat{f}_{\mathbf{H}}(x) f(x) dx + \int f^2(x) dx \tag{3.4.1}$$

This is the same equation as (2.5.4) where the last term of (3.4.1) can be ignored since it does not depend on $\mathbf{H}$. The first term can be calculated straight from the data and it is only the second term of (3.4.1) which needs to be estimated. The second term $\int \hat{f}_{\mathbf{H}}(x) f(x) dx$ is equivalent to $E\left[\hat{f}_{\mathbf{H}}(X)\right]$ where

$$E\left[\hat{f}_{\mathbf{H}}(X)\right] = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{\mathbf{H},-\mathbf{i}(X_i)} \tag{3.4.2}$$

and

$$\hat{f}_{\mathbf{H},-\mathbf{i}}(X_i) = \frac{1}{n-1}\sum_{i\neq j, j=1}^{n} \mathcal{K}_{\mathbf{H}}(X_j - x) \tag{3.4.3}$$

where $\mathcal{K}$ denotes a multivariate kernel function which operates over p-dimensional data.

Equation (3.4.2) is the multivariate version of the leave-one-out estimator for the univariate equation (2.5.5) where datapoint i is left out of the calculation to ensure that $\hat{f}_{\mathbf{H},-\mathbf{i}}$ is independent of X_i. By substituting (3.4.2) into (3.4.1) to obtain the multivariate $LSCV(\mathbf{H})$ criterion to minimize given by

$$LSCV(\mathbf{H}) = \frac{1}{n^2 \det(\mathbf{H})} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathcal{K} \star \mathcal{K} \left[\mathbf{H}^{-1}(X_j - X_i)\right] - \frac{2}{n(n-1)} \sum_{i=1}^{m} \sum_{i \neq j, j=1}^{n} \mathcal{K}_{\mathbf{H}}(X_j - X_i) \tag{3.4.4}$$

where the convolution $\mathcal{K} \star \mathcal{K}(u)$ is the multivariate form of $K \star K(u)$ which was defined in (2.5.10). The multivariate form of the least-squares cross-validation method retains the characteristics of its univariate counterpart such that it is easy to implement and interpret.

As in the univariate case, the multivariate version of biased cross-validation seeks to find an optimal bandwidth matrix to minimize the asymptotic mean integrated square error (AMISE). Sain, Baggerly and Scott [15] illustrate a proof of the biased cross-validation method when one considers the case of two-dimensional data. The resulting criterion can be extended to p-dimensional data:

$$BCV(h_1, \ldots, h_p) = \frac{1}{(2\sqrt{\pi})^p n h_1, \ldots, h_p} + \frac{1}{4n(n-1)h_1, \ldots, h_p} \times \sum_{i=1}^{n} \sum_{i \neq j} \left[\left(\sum_{k=1}^{p} \Delta_{ijk}^2 \right) - (2p-4) \left(\sum_{k=1}^{p} \Delta_{ijk}^2 \right) + (p^2 + 2p) \right] \times \Pi_{k=1}^{p} \phi(\Delta_{ijk}) \tag{3.4.5}$$

where ϕ is taken to be the standard normal density and $\Delta_{ijk} = \dfrac{x_{ik} - x_{jk}}{h_k}$.

In a simulation study for unimodal data [15] equation (3.4.5) performed well as it produced slightly lower standard deviations than the $LSCV(\mathbf{H})$ method. A second simulation for bimodal data and (3.4.5) appeared to have a much lower standard deviation than the $LSCV(\mathbf{H})$ method. These results suggest that biased cross-validation method for the bandwidth matrix could be good as it does not tend to oversmooth too much.

An alternative to the $LSCV$ and BCV methods is the smoothed cross-validation. This was first proposed by Hall, Marron and Park [10]. The resulting criterion to be minimized is:

$$SCV_g(\mathbf{H}) = (n\mathbf{H})^{-1}R(K) + \hat{B}_g(\mathbf{H}) \tag{3.4.6}$$

where

$$\begin{aligned}\hat{B}_g(\mathbf{H}) &= n^{-1}R(K)|\mathbf{H}|^{-\frac{1}{2}}(n-1)^{-1} \\ &\quad + n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{(K_{\mathbf{H}} \star K_{\mathbf{H}} - 2K_{\mathbf{H}} + K_0)L_{\mathbf{G}} \star L_{\mathbf{G}}\right\}(\mathbf{X}_i - \mathbf{X}_j)\end{aligned} \tag{3.4.7}$$

where K_0 is the Dirac delta function, kernel functions K and L are not necessarily the same and bandwidth matrices $\mathbf{H}$ and $\mathbf{G}$ are not necessarily the same.

In the case of a diagonal bandwidth matrix $\mathbf{H}$=$(h_1, \ldots, h_p)$ (3.3.3), Silverman [18] proposed a smoothing parameter, $\hat{h}_i$, to give the optimal window with for the smoothing of normally distributed data with unit variance:

$$\hat{h}_i = \left[\frac{4}{p+2}\right]^{\frac{1}{p+4}} \sigma_i n^{-\frac{1}{p+4}} \tag{3.4.8}$$

where p is the number of variables and an estimator $\hat{\sigma}_i$ is necessary for σ_i (the variance for each i^{th} item) in application. Equation (3.4.8) retains a lot of its characteristics as its univariate counterpart as it will cope well with unimodal densities and reasonably well with multimodal densities.

There are several packages in R [14] which allow one to analyse the implementation of bandwidth selection methods to data. There is the function `sm.density` in Bowman's `sm` [6] package which implements a variety of bandwidth selection methods. A drawback to [6] is that it can only handle data up to three dimensions and it only uses diagonal bandwidth matrices. Another popular function to visualise the data, which also uses diagonal bandwidth matrices is `kde2d` which is found in the `MASS` [19] `R` package. This function has attractive graphical features but is unable to cope with data that is higher than 2 dimensions. A more recent R package is Duong's `ks` [8]. This library claims it can deal with one-dimensional to six-dimensional data.

It implements both diagonal and unconstrained bandwidth matrices which would allow comparisons of both the diagonal and unconstrained bandwidth matrices to see if there is much difference between them.

3.5 Application of multivariate kernel density estimation to simulation study

Further to the multivariate normal model described in Section 3.1, kernel density estimation is also considered for the between-object distribution of the experimental databases described in the Appendix section to see how the false rates differ. The method assumes the background data has the form of $\mathbf{y}_{ij} = (y_{ij1}, \ldots, y_{ijp})^T; i = 1, \ldots, m; j = 1, \ldots, n;$. The mean, $\hat{\boldsymbol{\mu}}$, within-group covariance matrix, $\hat{\mathbf{U}}$, between-group covariance matrix, $\hat{\mathbf{C}}$ were calculated from the population databases described in the Appendix section.

The estimate of the probability density function at point y_i is given by:

$$\hat{f}_{\mathbf{H}}(y) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{\mathbf{H}} \mathcal{K}\left(\mathbf{y} - \mathbf{Y}_i\right) \tag{3.5.1}$$

where $\mathbf{H}$ is the bandwidth matrix, $\{\mathbf{Y}_i : i = 1, \ldots, m\}$ are the item means and $\mathcal{K}$ is the multiplicative kernel.

As for the univariate case of kernel density estimation Section 2.6, we assume there are $\mathbf{y}_{1j} = (\mathbf{y}_{1j1}, \ldots, \mathbf{y}_{1jp})$, j=1, …, n_1 control measurements taken assumed to have come from one source and similarly $\mathbf{y}_{2j} = (\mathbf{y}_{2j1}, \ldots, \mathbf{y}_{2jp})$, j=1, …, n_2 recovered measurements also coming from a single source. The means of the control and recovered measurements are $\bar{\mathbf{y}}_{\mathbf{l}} = \frac{1}{n_l} \sum_{j=1}^{n_l} \mathbf{y}_{lj}$ for $l = 1, 2$.

Following [22], using multivariate normal kernel density estimation the numerator of the likelihood ratio (1.2.1) is given by:

$$\begin{aligned}\text{numerator} = (2\pi)^{-\frac{p}{2}}|\mathbf{D}_1+\mathbf{D}_2|^{-\frac{1}{2}} \\ \times \exp\left[-\frac{1}{2}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)(\mathbf{D}_1+\mathbf{D}_2)^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\right] \\ \times \frac{1}{m}\sum_{i=1}^{m}(2\pi)^{-\frac{p}{2}}\left|\frac{\mathbf{U}}{n_1+n_2}+\mathbf{H}\right| \\ \exp\left[-\frac{1}{2}(\bar{\mathbf{x}}^*-\bar{\mathbf{y}}_i)\left(\frac{\mathbf{U}}{n_1+n_2}+\mathbf{H}\right)^{-1}(\bar{\mathbf{x}}^*-\bar{\mathbf{y}}_i)^T\right] \qquad (3.5.2)\end{aligned}$$

where $\bar{\mathbf{x}}^*=\frac{n_1\bar{\mathbf{x}}_1+n_2\bar{\mathbf{x}}_2}{n_1+n_2}$, $\mathbf{D}_l=\frac{\mathbf{U}}{n_l}$ for l=1,2 and $\bar{\mathbf{x}}_i$ are the means for each item in the population database. Similarly the denominator of the likelihood ratio (1.2.1) is given by:

$$\begin{aligned}\text{denominator} = (2\pi)^{-\frac{p}{2}}|\mathbf{D}_1+\mathbf{H}|^{-\frac{1}{2}}\frac{1}{m}\sum_{i=1}^{m}\left\{-\frac{1}{2}(\bar{\mathbf{x}}_\mathbf{1}-\bar{\mathbf{y}}_\mathbf{i})^T(\mathbf{D}_1+\mathbf{H})^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{y}}_i)\right\} \\ \times (2\pi)^{-\frac{p}{2}}|\mathbf{D}_2+\mathbf{H}|^{-\frac{1}{2}}\frac{1}{m}\sum_{i=1}^{m}\left\{-\frac{1}{2}(\bar{\mathbf{x}}_2-\bar{\mathbf{y}}_i)^T(\mathbf{D}_2+\mathbf{H})^{-1}(\bar{\mathbf{x}}_2-\bar{\mathbf{y}}_i)\right\} \qquad (3.5.3)\end{aligned}$$

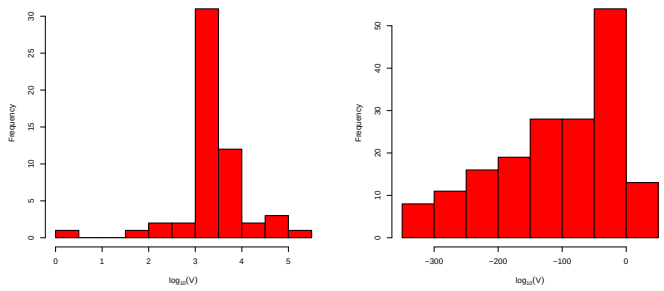
The likelihood ratio equation obtained by dividing equation (3.5.2) by equation (3.5.3):

$$V = \frac{\text{numerator}}{\text{denominator}} \qquad (3.5.4)$$

3.6 Simulation study results for refractive index data using the multivariate normal model

Same-source and different-source comparisons were implemented for the refractive index data in the simulation experiments similar to the one described in Section 2.4. The results for same-source and different-source comparisons

for the refractive index before annealing and after annealing are shown in Figure 3.2 and consist of the logarithms (base 10) of the likelihood ratio.



(a) Same-source comparisons (b) Different-source comparisons

Figure 3.1: $\log_{10}(V)$ values for same-source and different-source comparisons for the refractive index dataset using the multivariate normal model

The majority of the comparisons in Figure 3.1(a) have $\log_{10}(V)$ values between 3 and 4 and when one refers back to Table 1.1 there is strong evidence to support the proposition that two pairs of fragments come from the same source. Figure 3.1(a) shows there have been no false negatives with the same-source comparisons. This suggests that expression (3.1.14) has been extremely effective in identifying that two pairs of measurements do come from the same source.

Looking at the different-source comparisons, there have only been 0.9% of comparisons wrongly considered to have come from the same source. There are two out of the 13 false positive comparisons which have a value of $\log_{10}(V)$ greater than 3 (corresponding value of V=1000). Even though these are large misleading values of the likelihood ratio and there is strong evidence to support the same-source proposition, this is only accounts for 0.13% of the dataset and it could be said that bivariate normal model has performed

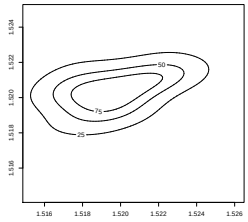
extremely well with the remaining 99.87% of the dataset. Another factor which suggests that expression (3.1.14) has performed extremely well is the fact that 1308 of all possible comparisons have a resulting value of V to be $-\infty$ which gives very strong evidence to support the different-source proposition.

When we compare the values of V for the bivariate normal model to the univariate models for the data in Chapter 2, the bivariate normal model produces slightly lower values of V than would have been obtained by combining the likelihood ratios for the refractive index dataset before annealing and the refractive index after annealing assuming that the two are independent. By definition, if the two variables are assumed to be independent their values of V can be multiplied together, for example, if we look at item 25: this item had a likelihood ratio based on the refractive index before annealing value of 48 and a likelihood ratio on the refractive index after annealing value of 66. The product of these two values is 3168. When we compare this value to the corresponding value for item 25 for the multivariate normal model value of 2957, it is slightly higher. The reason for the multivariate normal model having a slightly lower value of V could be that the model allows for dependencies between the variables.

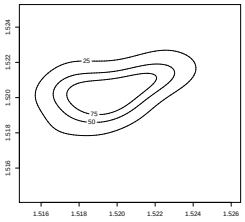
3.6.1 Bandwidth selection for the refractive index dataset

The bandwidth selection methods described in Section 3.4 were applied to the refractive index data described in Section 3.2. Kernel density estimation of the between-object distributions for the refractive index data is performed on the bivariate item means and the resulting contour plots are shown in Figure 3.4. We have considered the different bandwidth selection methods using the unconstrained bandwidth matrix (3.3.4), the diagonal bandwidth matrix (3.3.3) and Silverman's [18] rule of thumb (3.4.8) where the bandwidth is of the form h^2C and all use Gaussian kernels. Contour plots 3.4(a)-(f) have been obtained using the `ks` [8] package in R [14] where the upper 25%, 50% and 75% contours are the highest density regions within in the data and contour plot (g) has been obtained from the `sm` [6] package.

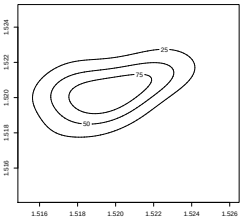
Figure 3.2(a)-(f) shows that there is very little difference in terms of the shape of the contour plots when one considers using either the unconstrained bandwidth matrix or the diagonal matrix, however the unconstrained matrices appear to describe the data slightly better. The shape of the contour plot for Figure 3.2(g) does not differ that much from the plots which precede it and its 25%, 50% and 75% contours are similar to the other diagonal bandwidth matrix options. Therefore, as for the univariate case in kernel density estimation, it is unlikely that there will be much of a difference, if any, in the simulation results depending on which bandwidth selection method was used. The least-squares cross-validation, biased cross-validation, smoothed cross-validation bandwidth selection methods using the unconstrained matrices and Silverman's rule of thumb were thus chosen to be implemented.



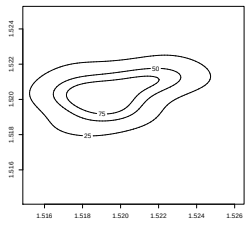
(a) Least-squares CV unconstrained



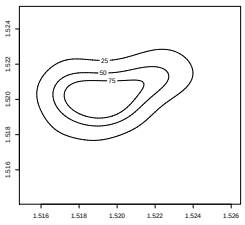
(b) Biased CV unconstrained



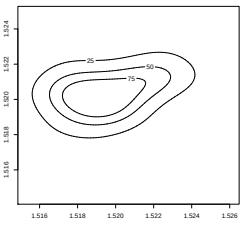
(c) Smoothed CV unconstrained



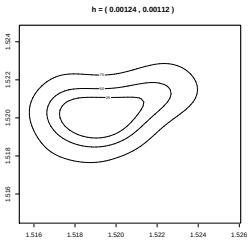
(d) Least-squares CV diagonal



(e) Biased CV diagonal



(f) Smoothed CV diagonal



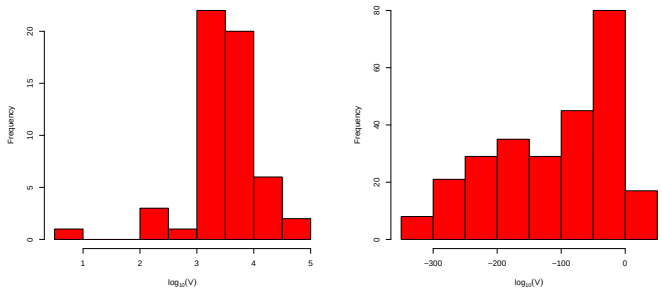
(g) Rule of thumb

Figure 3.2: Comparing different bandwidth choices for the refractive index dataset using multivariate kernel density estimation

3.6.2 Simulation study results for refractive index data using multivariate kernel density estimation

Same-source and different-source comparisons were implemented for the refractive index data in a simulation experiment similar to the one described in Section 2.4 where the likelihood ratio for each comparison was obtained

using multivariate kernel density estimation. The histograms of same-source and different-source comparisons consisting of logarithm (base10) of likelihood ratios for comparisons for the refractive index data are shown in Figure 3.3.



(a) Same-source comparisons (b) Different-source comparisons

Figure 3.3: $\log_{10}(V)$ values for same-source and different-source comparisons for the refractive index dataset using multivariate kernel density estimation

The smoothed cross-validation method is shown in Figure 3.3(a) for illustrative purposes since there was no difference between the four methods of bandwidth selection. For all four methods of bandwidth selection there were no false negative results and most of the values of $\log_{10}(V)$ lie between 3 and 4 which is strong evidence to support the proposition that the two pairs of measurements come from the same source. A couple of minor differences between the four bandwidth selection methods is that $LSCV$ classifies one object with a value of V to be 100,000,000 whereas the maximum of the other two methods is 100,000 and for Silverman's method, most of the values for $\log_{10}(V)$ lie between 4 and 4.5 which is slightly higher than the other methods. Figure 3.3(a) suggests that equation (3.5.4) described in Section 3.5 has

been extremely effective in identifying that the two pairs of measurements do come from the same source since the rate of false negatives is 0%.

Looking at Figure 3.3(b) there are only 13 false positive comparisons out of a possible 1485, for all choices of unconstrained bandwidth selection matrices for the different-source comparisons. There are four out of these 13 comparisons which have a value of $\log_{10}(V)$ greater than 3 (corresponding value of V=1000). Although this suggests there is strong evidence to support the same-source proposition, this only equates to 0.3% of all possible comparisons. Silverman's method had two more false positive comparisons than the other methods which is equivalent to a 1% false positive rate. Eleven of these comparisons had a value of $\log_{10}(V)$ to be 4. Although this is equivalent to very strong evidence to support the same source proposition, this only accounts for 0.7% of the dataset. There are, at the other end of the scale for all bandwidth selection methods, 1308 comparisons which have a value of V to be $-\infty$. The results for different-source comparisons suggest that equation (3.5.4) described in Section 3.5 has been very effective in identifying that two pairs of measurement come from two separate sources since the false-positive rate is extremely low at a maximum of 1%.

In terms of the false-positive and false-negative rates, the multivariate kernel density estimation has performed slightly better than the univariate version of kernel density estimation. This suggests that the multivariate population database could be favoured since the false positive rate is slightly lower at 0.9% compared to its univariate counterpart of 2.6% and 3.0%.

When we allow possible dependencies between the refractive index before annealing and the refractive index after annealing, there is a slight improvement in percentages of false positive and false negative rates for both approaches. This suggests that modelling the data multivariately may be more favourable than looking at each variable separately. The Tippett plots in Figure 3.4(a)-(b) provide further justification as the separation between the two lines is slightly wider than for the univariate data. Certainly, with this dataset, it

appears more appropriate to use the multivariate methods of analysis rather than the univariate approaches described in Chapter 2. Both multivariate methods produced exactly the same false positive and false negative results which suggests that either method is appropriate to the refractive index data, however it might be wiser to use the multivariate normal model since it is less complex in nature than multivariate kernel density estimation.

3.6.3 Simulation study results for the elemental composition data using the multivariate normal model

As shown in Sections 2.6 and 3.6 the refractive index is an extremely good variable for identifying same-source and different-source comparisons of glass fragments. There are times when the refractive index of glass fragments is not available and instead the elemental compositions are available. In order to study how effective elemental compositions are in identifying same-source and different-source glass fragments, we repeat the simulation experiment of Section 2.4 on an experimental database of elemental compositions supplied by G. Zadora of the Institute of Forensic Research in Krakow. Glass comprises of sodium (Na), magnesium (Mg), aluminium (Al), silicon (Si), potassium (K) calcium (Ca), iron (Fe) and oxygen (O). The database consists of seven variables which are log(-logratios) where oxygen was taken to be the baseline element. A more detailed account of the database and how the variables were formed can be found in Appendix B.

Many different subsets of the data were investigated. We combined variables which look to have reasonably constant variation between the types of glass and variables which had different amounts of variation between the types of glass. The goal was to establish whether the elemental composition of glass is as good as the refractive index dataset in distinguishing between same-source and different-source pairs. This might not be easy to obtain since the elemental composition dataset has more variables measured which makes it

naturally more complicated than the refractive index dataset.

Same-source and different-source comparisons were implemented to the elemental composition dataset described in Section 4 for all the variables combined. The results of these comparisons are shown in Figure 3.5 and consist of the logarithms (base10) of the likelihood ratio. The false negative and false positive rates can be found in Table 3.1 for various combinations of variables.

Variables Used	**False negatives (%)**	**False positives(%)**
Na, Si, Ca	2.5	37.6
Mg, Al, Fe	2.5	10.0
Na, K, Ca	3.4	18.0
Na, Al, K, Fe	3.4	9.0
Na, Al, Si, Ca	3.2	17.8
Na, Al, Si, K	4.0	11.1
Na, Mg, Si, Ca, K	3.8	10.3
Na, Al, Si, Ca, K	4.1	11.0
Na, Mg, Al, Si, K, Ca	3.4	6.6
Mg, Al, Si, K, Ca, Fe	3.4	5.5
Na, Mg, Si, K, Ca, Fe	2.8	6.6
Na, Mg, Al, Si, K, Ca, Fe	3.4	5.5

Table 3.1: Table of false results for subsets of the elemental composition data using the multivariate normal model

Looking at Table 3.1, the false positive rates are reasonably high for lower dimensions of the data and so it appears that we need to model the data in six or seven dimensions to obtain both suitably low false negative and false positive rates with the best and simplest variable combination for the normal model highlighted in yellow.

When we compare these results to the refractive index data, particularly the

false positive rates are a lot higher. With the bivariate refractive index data it was possible to obtain a perfect false negative rate of 0% for the same-source proposition and a very good false positive rate of 0.9% for the different-source proposition. However, when the same technique is applied to the elemental composition dataset, the false rates are not so promising particularly the false positive rates. The elemental composition dataset is naturally more complicated than the refractive index dataset with more variables measured, so it is no surprise that it does not perform as well with the multivariate normal model as we have to model the data in at least six dimensions to get reasonable false rates.

It is worth investigating whether using multivariate kernel density estimation would improve these false rates using the same subsets of the data.

3.6.4 Simulation study results for elemental composition dataset using multivariate kernel density estimation

As for the multivariate normal model, we applied multivariate kernel density estimation to the elemental composition dataset and the goal was to establish whether this dataset is as good as the refractive index at distinguishing between same-source and different-source pairs.

Same-source and different source comparisons were implemented for the elemental composition dataset in a simulation experiment which is similar to the one described in Section 2.4. Table 3.2 shows the results of Silverman's bandwidth method applied to multivariate kernel density estimation.

Variables Used	False negatives (%)	False positives(%)
Na, Si, Ca	2.8	62.5
Mg, Al, Fe	2.5	10.8
Na, K, Ca	2.5	24.1
Na, Al, K, Fe	2.5	14.6
Na, Al, Si, Ca	1.9	37.0
Na, Al, Si, K	4.1	17.9
Na, Mg, Si, Ca, K	2.8	21.2
Na, Al, Si, Ca, K	3.2	35.8
Na, Mg, Al, Si, K, Ca	4.1	11.8
Mg, Al, Si, K, Ca, Fe	3.4	10.3
Na, Mg, Si, K, Ca, Fe	3.4	11.8
Na, Mg, Al, Si, K, Ca, Fe	3.2	7.6

Table 3.2: Table of false results for subsets of the elemental composition dataset using multivariate kernel density estimation

The least-squares cross-validation bandwidth selection method was implemented to the same subsets of data but, naturally, the computation of each result for the LSCV method is far more involved than Silverman's rule of thumb as it incorporates the determinant of $\mathbf{H}$, the convolution of multivariate kernels and the inverse of $\mathbf{H}$ whereas Silverman's rule of thumb only deals with the number of variables involved, the total number in the sample and the variance of each variable. It follows that the computation of LSCV method comparisons take a lot longer than Silverman's rule of thumb.

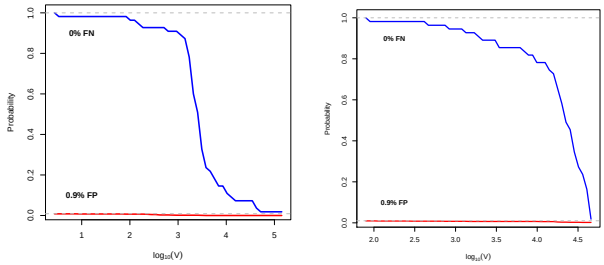
Error messages were returned for different-source comparisons because of computational problems in `R`. If time had permitted, one would have looked at the function for the likelihood ratio when using kernel density estimation and the structure of the simulations in order to make them more efficient with a view to obtain values for different-source comparisons when using LSCV.

Table 3.2 shows that Silverman's bandwidth selection method has performed

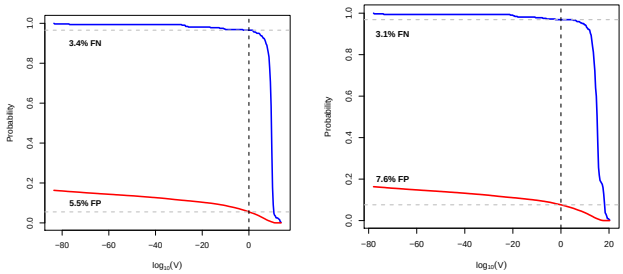
fairly well, particularly with same-source pairs which are comparable to the multivariate normal method. The false positive rates are, in general, higher than the multivariate normal model. It is interesting that three variables Mg, Al and Fe perform very well particularly with different-source comparisons. When we compare the results of multivariate kernel density estimation applied to the elemental composition dataset to the bivariate refractive index dataset, the false rates are nowhere near as good, hence when available the refractive index is a preferable variable to use for evidence evaluation in the form of glass fragments.

3.7 Tippett plots for multivariate data

Tippett plots in Figure 3.4(a)-(b) are for the multivariate refractive index database when a multivariate normal distribution and multivariate kernel density estimation have been used to estimate the between-group distribution. The Tippett plots in Figure 3.4(c)-(d) are for the elemental composition database which are highlighted in Table 3.1 and Table 3.2. These are considered to be the best subsets of the data which produce reasonably low false rates.



(a) Refractive index MVN (b) Refractive index KDE



(c) Normal - Mg,Al,Si,K,Ca,Fe (d) Kernel density - Na, Mg,Al,Si,K,Ca,Fe

Figure 3.4: Tippett plots for multivariate datasets using both multivariate normal distribution (MVN) and multivariate kernel density estimation (KDE) to estimate the between-group distribution

The separation between the two lines in both Figures 3.4(a) and (b) is very wide which suggests that both approaches for estimating the between-group distribution have been highly effective in identifying same-source and different-source pairs for the refractive index database. The separation between the two curves in Figures 3.4(c)-(d) is narrower than (a) and (b) which suggests that the refractive index database was easier to work with and produced better results.

3.8 Summary

Chapter 3 presented an extension of the two-level random effects model which was described in Chapter 2 as the model was applied to evaluating evidence in the form of multivariate data.

A multivariate normal model as well as multivariate kernel density estimation was employed for the between-object distribution. Both methods performed extremely well with the refractive index dataset as very low false rates were produced. This suggests that both methods have been effective at identifying same and different-source pairs of refractive index measurements.

Both methods were also applied to the elemental composition dataset. The false rates were not as good as for the refractive index dataset. It was found that the dataset needed to be modelled in higher dimensions to achieve suitably low false rates. The multivariate normal model yielded lower false rates, in general, than kernel density estimation especially for different-source pairs. The best and simplest combination of variables for multivariate normal model to produce reasonably low false rates was magnesium, aluminium, silicon, potassium, calcium and iron. However, for multivariate kernel density estimation all variables were needed to produce reasonably low false rates for both same-source and different-source pairs.

It would be advisable to use the multivariate normal model for both types of data due to its simplicity since one only needs to compute $\hat{\boldsymbol{\mu}}$, $\hat{\mathbf{U}}$ and $\hat{\mathbf{C}}$ whereas with multivariate kernel density estimation one needs to consider the type of kernel, bandwidth selection method and the choice of the bandwidth matrix in addition to the parameter estimates.

Chapter 4

Discussion

The common problem in forensic science of whether two sets of measurements come from the same source is one which frequently arises during police investigations. This thesis has dealt with two databases of glass fragment evidence containing information on the refractive index and the elemental composition of glass. Simulation studies were constructed for both datasets in an attempt the answer the problem in forensic science of whether two sets of measurements come from the same source. This was assessed by the likelihood ratio which is a measure of evidential value. A two-level random effects model was used to obtain the likelihood ratio, where the levels of variability were the between glass objects (between-group) and that within glass objects (within-group).

Two different and commonly used methods for estimating the between-group distribution were employed: the normal approach and kernel density estimation. The performance of these methods was assessed in simulation experiments by means of false rates. For same and different-source comparisons, the source of each pair of fragments was known. Incorrectly identified pairs for same-source comparisons were known as false negatives and for different-source comparisons it was false positives. The refractive index database proved an easier dataset to work with since the false rates for both methods were low for the refractive index before and after annealing when the data

was modelled univariately and the false rates were even lower for each method when the refractive index before annealing and the refractive index after annealing were combined in Chapter 3. The elemental composition database in Chapter 3 did not perform as well as the refractive index database as it produced higher false negative and false positive rates for both methods estimating the between-group distribution.

When analysing glass fragments it may be required to select a method which leaves fragments available for future testing especially in forensic settings. The method for determining the refractive index, described in Appendix A, has the potential to destroy glass fragments [21] rendering these unusable in the future. The method for determining the elemental composition of glass fragments is not as destructive because the fragments can be cleaned up after the process and so satisfy the requirement that they can be used again in the future, so there might be instances where elemental compositions need to be used instead of refractive indices. In these cases, it is still possible to distinguish between same and different-source pairs with reasonably good accuracy using multivariate normal model with six variables of elemental compositions but not as good accuracy as for refractive indices using either the multivariate normal or the kernel density estimation models.

To conclude, it is favourable to use the refractive index of glass when it is readily available. However, this is not always the case and it is acceptable to use the elemental composition of glass as shown in [21]. The method of obtaining the elemental composition produces fairly low false rates when simulation studies are set up in attempt to answer the forensic question of whether two sets of measurements come from a common source when six or more variables are combined.

Since there were no results recorded for the LSCV method in Section 3.6.4, it would be of interest in the future to simplify both the function used for obtaining the likelihood ratio and the structure of the simulations with a view to comparing these results with Silverman's bandwidth selection method.

Appendix A

Refractive index database

An experimental dataset of refractive indices was available to be used as a population database. The dataset was supplied by G. Zadora and is described in [21]. The data consists of 55 glass objects, with four measurements of the refractive index for each object (four fragments). The refractive index was determined by a method known as thermo-immersion. Each fragment of glass was placed on its own clean microscopic slide and was then covered with silicone oil. When the match temperature (MT) of refractive index of the silicone oil is equal to the refractive index of the glass, the value of the refractive index was determined by the calibration model:

$$\text{Refractive Index} = -3.74 \times 10^{-4} MT + 1.54491$$

a more in depth discussion of which can be found in [21].

The database had m=55 items measured (32 car and 23 building windows) where each item had n=4 fragments measured, resulting in a total of N=mn=220 observations. Each fragment was subjected to an annealing process and the refractive index was measured again after annealing. Thus, there are 220 measurements of refractive index before annealing and 220 measurements of refractive index after annealing.

The population database will be used to obtain parameter estimates and also to supply "control" and "recovered" measurement pairs for a simulation

experiment.

Table A.1 shows the estimates of $\hat{\mu}$, $\hat{\sigma}^2$ and $\hat{\tau}^2$ for both the refractive index before annealing and the refractive index after annealing.

Variable used	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\tau}^2$
RI before annealing	1.52	2.6×10^{-9}	5.39×10^{-6}
RI after annealing	1.52	1.72×10^{-9}	4.79×10^{-6}

Table A.1: Parameter estimates for the refractive index dataset

Figure A.1 shows refractive index after annealing against the refractive index before annealing.

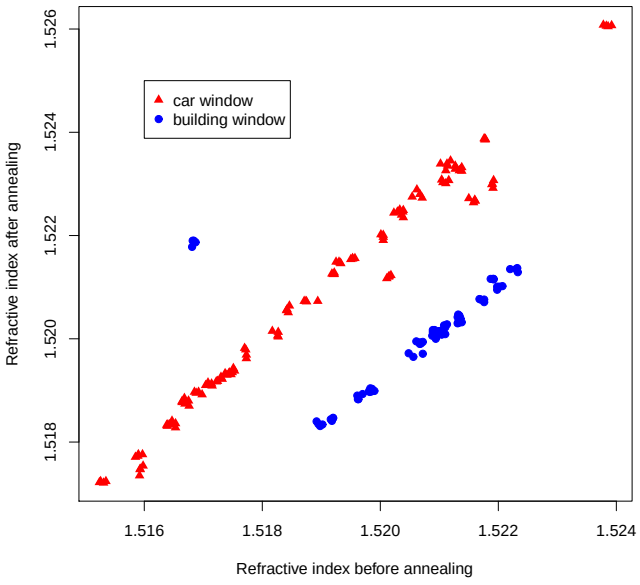


Figure A.1: Scatterplot of refractive index before and refractive index after annealing

Figure A.1 shows a positive relationship between the refractive index before annealing and the refractive index after annealing. The lower the refractive index of a fragment before annealing, then it is likely that fragment will have a reasonably low refractive index after annealing. Figure A.1 also shows that

building windows tend to have lower refractive indices than car windows with the exception of a few outliers.

When the variables were combined in Chapter 3, the estimate of the $\boldsymbol{\mu}$ parameter was as follows:

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} 1.520 & 1.519 \end{bmatrix} \tag{A.0.1}$$

The within-group variance-covariance matrix, $\mathbf{U}$, was estimated to be:

$$\hat{\mathbf{U}} = \begin{bmatrix} 1.72 & 7.00 \\ 7.00 & 2.60 \end{bmatrix} \tag{A.0.2}$$

The between-group variance-covariance matrix, $\mathbf{C}$, was estimated to be:

$$\hat{\mathbf{C}} = \begin{bmatrix} 5.85 & 1.43 \\ 1.43 & 4.73 \end{bmatrix} \tag{A.0.3}$$

Appendix B

Elemental composition database

In addition there was another experimental database available which was also supplied by G. Zadora and is described in [23]. The dataset consisted of transformed elemental compositions of glass from 320 objects (26 bulbs, 94 car windows, 16 headlamps, 79 containers and 105 building windows). Eight variables were measured for each item of glass namely: oxygen (O), sodium (Na), magnesium (Mg), aluminium (Al), silicon (Si), potassium (K), calcium (Ca) and iron (Fe). Initially, each variable represented its percentage weight of the total weight of the glass object, so the eight variables summed to 100%. However it was sufficient to know the weight of seven of the variables because the eighth variable can be worked out as 100% - (sum of the other seven variables). A common transformation for compositional data is the logratio. This was performed to the dataset where oxygen was taken to be the baseline element. The logratios were of the form:

$$Na' = \max\left(\log_{10}\left(\frac{Na}{O}\right), \log_{10}\left(\frac{0.0001}{O}\right)\right) \tag{B.0.1}$$

$$\vdots$$

$$Fe' = \max\left(\log_{10}\left(\frac{Fe}{O}\right), \log_{10}\left(\frac{0.0001}{O}\right)\right) \tag{B.0.2}$$

It was necessary to take the maximum of the two numbers for each variable because for some readings of Fe, the readings were zero and one cannot evaluate the logarithm of zero. Further, the log(-logratio) was taken for each variable:

$$Na'' = \log_{10}(-Na' + 0.1) \tag{B.0.3}$$

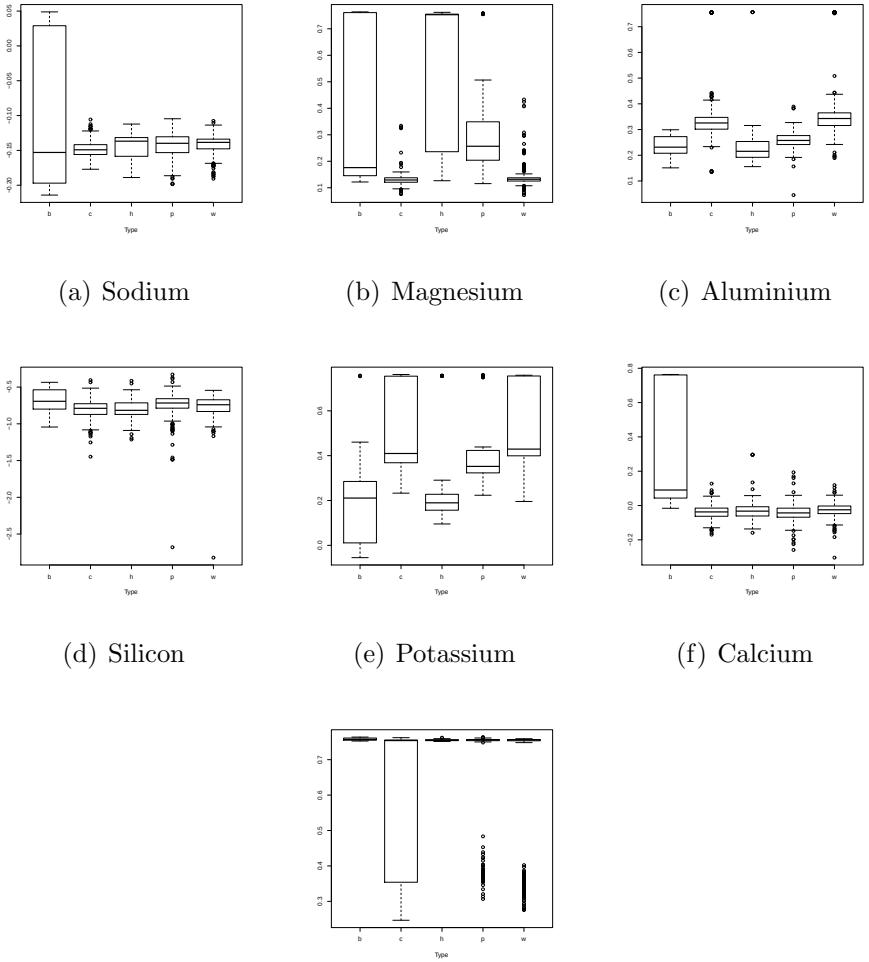
$$\vdots$$

$$Fe'' = \log_{10}(-Fe' + 0.1) \tag{B.0.4}$$

There was a slight adjustment of 0.1 because oxygen had the highest elemental concentration in all but two measurements where silicon concentration was higher than oxygen. When one takes the logratio of (Si/O), there would be negative values in all but two measurements. Hence, it is necessary to add a small constant so that all measurements are positive when one is required to take the log(-logratio).

The final database resulted in transformed elemental composition dataset of sodium (Na), magnesium (Mg), aluminium (Al), silicon (Si), potassium (K), calcium (Ca) and iron (Fe). The database had m=320 items measured where each item had n=4 fragments measured, resulting in a total of N=mn=1280 observations.

The following plots show how the seven variables differ from one another and the variation across the objects.



(a) Sodium (b) Magnesium (c) Aluminium

(d) Silicon (e) Potassium (f) Calcium

(g) Iron

Figure B.1: Investigating each element by the type of glass

Figure B.1(g) shows that there is very little difference between bulbs, headlamps, containers and building windows in terms of the amount of Iron but there is a large amount of variability in Iron coming from car windows. Aluminium and Silicon appear to have fairly constant variability across the different glass types, as does calcium with the exception of the glass from bulbs.

Bibliography

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

[2] Aitken, C.G.G, Lucy, D, Zadora, Curran, J.M. (2005). Evaluation of transfer evidence for three-level multivariate data with the use of graphical models. *Computational Statistics and Data Analysis*, 50, 2571-2588.

[3] Aitken, C. and Taroni, F. (2004). *Statistics and the evaluation of evidence for forensic scientists*. Wiley, 2nd edition.

[4] Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71, 353-360.

[5] Bowman, A.W. and Azzalini, A. (1997). *Applied smoothing techniques for data analysis*. Oxford Statistical Science Series.

[6] Bowman, A. W. and Azzalini, A. (2010). R package "sm": nonparametric smoothing methods (version 2.2-4).

[7] Duong, T. (2007). ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *Journal of Statistical Software* 21, 7.

[8] Duong, T. (2010). ks:Kernel Smoothing. *R package version 1.6.12*.

[9] Fraley, C. and Raftery, A.E. (2002). Model based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*. 97, 611-631.

[10] Hall, P., Marron, J.S. and Park, B.U. (1992). Smoothed cross validation. *Probability theory and related fields*. 92. 1-20.

[11] Härdle, W. (1991). *Smoothing techniques, with implementation in S*, Springer, New York.

[12] Härdle, W and Müller, M. (1989). *Nonparametric and Semiparametric Models*, Springer.

[13] Lindley, D. V. (1977). A problem in forensic science. *Biometrika*, 64, 207-213.

[14] R Development Core Team (2010). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. ISBN 3-900051-07-0, URL http://www.R-project.org.

[15] Sain, S. R. and Baggerly, K. A. and Scott, D. W. (1994). Cross-validation of multivariate densities. *Journal of the American Statistical Association*. 89, 807-817.

[16] Scott, D. W. and Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation. *Journal of American Statistical Association*, 82, 1131-1146.

[17] Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 53, 683690.

[18] Silverman, B. W. (1986). *Density estimation for Statistics and Data Analysis*, Chapman and Hall.

[19] Venables, W.N. and Ripley, B.D. (2002). Modern applied statistics with S. Fourth Edition. *Springer*. New York.

[20] Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing*, Chapman and Hall, London.

[21] Zadora, G. and Neocleous, T. (2009). Likelihood ratio model for classification of forensic evidence. *Analytica Chimica Acta*. 642, 266-278.

[22] Zadora, G. and Neocleous, T. (2010). Evidential value of physicochemical data comparison of methods of glass database creation. *Analytica Chimica Acta*. 24, 367-378.

[23] Zadora, G. and Neocleous, T. and Aitken, C. (2010). A two level model for evidence evaluation in the presence of zeros. *Journal of Forensic Sciences*. 55, 371-382.