EVIDENCE EVALUATION FOR HIERARCHICAL, LONGITUDINAL BINARY DATA USING A DISTANCE MEASURE

Colin G.G. Aitken, Caiyi Huang

School of Mathematics and Maxwell Institute, University of Edinburgh, Peter Guthrie Tait Road, Edinburgh EH9 3FD, UK

Abstract *An approach for evidence evaluation for trace evidence in the form of hierarchical, longitudinal binary data is described. A non-parametric density method is applied to a measure of distance, treated as continuous, between control and recovered data. Training data are available of striation marks from a set of screwdrivers for the estimation of within-source and between-source distances. Correct directions of support are obtained in over 90% of test comparisons.*

Keywords: *Binary data, evidence evaluation, kernel density estimation, trace evidence.*

1. INTRODUCTION

An important role of the forensic scientist in the investigation and prosecution of a crime is the interpretation and evaluation of evidence. Consider measurements on trace evidence, such as the measurements of the refractive index of fragments of glass; denote these measurements as E. In a criminal case, there are two opposing sides, that of the prosecution and that of the defence. Assume they have propositions related to E; denote these as H_p and H_d. It is then desired to measure the effect of the evidence (E) on the probability that the prosecution's proposition (H_p) is true against the probability that the defence's proposition (H_d) is true, *i.e.*,

$$\frac{Pr(H_p \mid E)}{Pr(H_d \mid E)} = \frac{Pr(E \mid H_p)}{Pr(E \mid H_d)} \times \frac{Pr(H_p)}{Pr(H_d)}. \tag{1}$$

The ratio $Pr(E \mid H_p)/Pr(E \mid H_d)$ is the ratio of the probability of the evidence assuming the prosecution's proposition (H_p) is true to the probability

of the evidence assuming the defence's proposition (H_d) is true. This ratio is known as the *likelihood ratio*. The ratio
$Pr(H_p \mid E)/Pr(H_d \mid E)$ is the posterior odds in favour of the prosecution proposition, given the evidence. The ratio $Pr(H_p)/Pr(H_d)$ is the prior odds in favour of the prosecution proposition. A value of the likelihood ratio greater than one is supportive of the prosecution's proposition, a value less than one is supportive of the defence proposition since, in the two situations, the posterior odds in favour of the prosecution's proposition is greater or less, respectively, than the prior odds.

When the evidence takes the form of continuous measurements, the probabilities are replaced by probability density functions. A common example of such evidence is the measurement of the refractive index of a fragment of glass.

Trace evidence, examples of which are body fluids, fragments of glass and gunshot residue, is often in two parts. One part is evidence whose source is known, this is *control evidence*, denoted $\mathbf{x}$. The second part has an unknown source and is known as *recovered evidence*, denoted $\mathbf{y}$. For example, in a crime in which a window has been broken, $\mathbf{x}$ would be a set of measurements of refractive indices of fragments of glass from the broken window. A suspect is found and they have fragments of broken glass on their person. The set of measurements of refractive index of these fragments would be $\mathbf{y}$; the fragments may have come from the crime scene window but may not have. Denote the joint probability density function of $\mathbf{x}, \mathbf{y}$ by $f(\mathbf{x}, \mathbf{y})$ and the likelihood ratio is then

$$\frac{f(\mathbf{x}, \mathbf{y} \mid H_p)}{f(\mathbf{x}, \mathbf{y} \mid H_d)}, \tag{2}$$

where H_p is the prosecution's proposition that the control and recovered fragments have the same source and H_d is the defence proposition that the control and recovered fragments have different sources.

Much has been written about the evaluation of the likelihood ratio when the evidence is in the form of continuous measurements; see, for example, Aitken and Taroni (2004), Aitken and Lucy (2004). However, there is a paucity of methods for discrete data. See Aitken and Gold (2013) for a discussion of two models for the evaluation of evidence in the form of discrete data, motivated by an example in forensic phonetics, in which the data are the numbers of occurrences of a particular vocal characteristic in a

of speech. One model assumes the observations are independent and identically distributed following a Poisson distribution, while the other model assumes the adjacent observations are dependent, leading to a bivariate Bernoulli model.

Another example of the use of discrete data in forensic science, where the data are hierarchical, longitudinal and binary, is discussed in Petraco *et al.* (2012). Classification rates are determined to demonstrate the effectiveness of the procedure for strength of association. These data motivated the development of another approach for the evaluation of evidence.

2. DATA

The source of the discrete data used in Petraco *et al.* (2012) is that of striation marks made by a tool, and the authors describe an experiment to investigate the evidential possibilities of such marks. An experiment was conducted in which nine identical screwdrivers were used. The striation patterns made by each of the nine screwdrivers were recorded. Distances of each line or groove from the left edge of each striation pattern were measured to the nearest 0.05 mm. Each striation pattern was no more than 7 mm wide. For each pattern, a list of 140 pieces of information (7 mm/0.05 mm slots) is created. Each piece of information is a 1 or 0. A 1 is recorded in a slot on the list if a line or groove were present or spans the slot. A 0 is recorded otherwise. The procedure yielded a 140-dimensional binary feature vector for each pattern. In the 140 components of the feature vector, 19 always had value 0 across all recorded striation patterns. Petraco *et al.* (2012) excluded these non-varying components from their analyses. They are retained here for completeness. The analysis described is based on a distance measure between two sets of marks, so the effect of their retention is zero. Methods described by Petraco *et al.* (2012) are based on partial least squares discriminant analysis and principal component analysis with support vector machines. Classification rates of correct assignments of marks to the screwdriver that made the mark of 97% or higher were achieved. Further details are available from Petraco *et al.* (2012). However, none of these methods provides a value for evidence in the form of a likelihood ratio. They show that the detection methods of associating marks with screwdrivers were good.

In a particular case, the evidence E would be the striated marks made by a screwdriver presented in the form of a vector in $B^{140} = \{0,1\}^{140}$. The *control* evidence, $\mathbf{x}$, is the vector of marks for which the source (screwdriver) is known. This could be a screwdriver found in the possession of a suspect, for example. The *recovered evidence*, $\mathbf{y}$, is evidence for which the source is not known. This could be a set of striation marks found at the scene of a crime. These marks could have been made by the screwdriver found in the possession of the suspect; this would be the prosecution's proposition, H_p. Alternatively, these marks could have been made by some other screwdriver; this would be the defence proposition, H_d. The evidence, $\mathbf{x}, \mathbf{y}$, is discrete, so its value can be determined by consideration of

$$\frac{Pr(\mathbf{x}, \mathbf{y} \mid H_p)}{Pr(\mathbf{x}, \mathbf{y} \mid H_d)}. \tag{3}$$

However, some method is needed for the estimation of the associated probability mass function $Pr(\cdot, \cdot)$, over a bivariate B^{140} sample space. This is impractical and a method based on a distance measure $d(\mathbf{x}, \mathbf{y})$ is proposed.

The proposed method is developed with the use of a training set that consists of seventy-five records of striation marks in B^{140} made by the nine screwdrivers. The 75 records are divided into nine groups, one for each screwdriver, indexed by q ($q = 1, \ldots, 9$). There are l_q replicates for screwdriver q with $\mathbf{l} = (l_1 \ldots, l_9)$. The data are $\mathbf{l} = (8, 6, 9, 8, 9, 9, 8, 9, 9)$, with $\sum_{q=1}^{9} l_q = 75$. Therefore each observation in the data set can be represented as $\mathbf{z}_{qk}$ ($k = 1, \ldots, l_q$; $q = 1 \ldots, 9$), and $\mathbf{z}_{qk} \in B^{140}$. The data are hierarchical, there is variability between striation marks made by the same screwdriver (known as *within-source* variation) and variability between striation marks made by different screwdrivers (known as *between-source* variation).

If the measurements for each observation were continuous variables, methods based on multivariate normal distributions or kernel density estimation may be used (Aitken and Lucy, 2004). For binary data, it is hard to estimate a distribution from the training set that describes the observations over B^{140}. There are 2^{140} members of the sample space and only 75 members of the training set. A procedure based on the differences between sets of measurements within each group and between each group is described as an alternative.

Let $\mathbf{x}$ and $\mathbf{y}$ be two sets of binary measurements in B^{140}. The distance

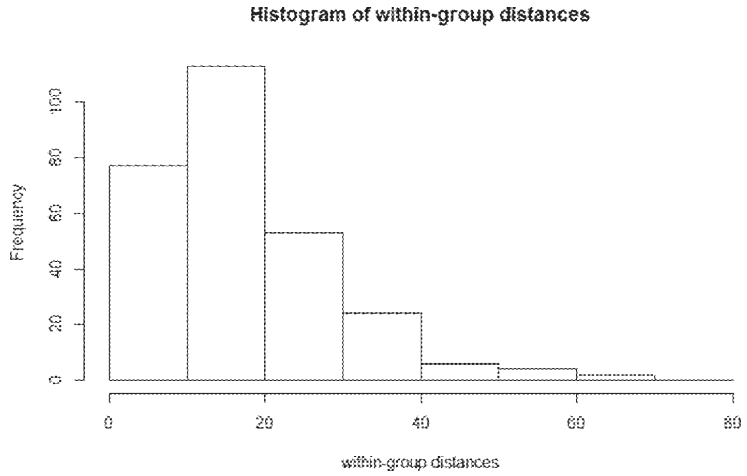

Figure 1: Histogram of within-group distances.

$d(\mathbf{x}, \mathbf{y})$ between them is defined as

$$d(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T(\mathbf{x} - \mathbf{y}) = \sum_{i=1}^{140}(x_i - y_i)^2; \;\; \{x_i, y_i\} \in \{0, 1\}. \tag{4}$$

For group q ($q = 1, \ldots, 9$), determine

$$d(\mathbf{z}_{qk_1}, \mathbf{z}_{qk_2}) \text{ for } k_1 = 1, \ldots, l_q - 1; \; k_2 = k_1 + 1, \ldots, l_q, \tag{5}$$

so there are $\frac{1}{2}\sum_{q=1}^{9} l_q(l_q - 1) = 279$ within-group distances.

Similarly, the distance between the pairs of observations from different groups is given by

$$d(\mathbf{z}_{q_1k_1}, \mathbf{z}_{q_2k_2}) \tag{6}$$

for $q_1 = 1, \ldots, 8; \; q_2 = q_1 + 1, \ldots, 9; k_1 = 1, \ldots, l_{q_1}; \; k_2 = 1, \ldots, l_{q_2}$.

There are $\frac{1}{2}\sum_{q_1=1}^{8}\sum_{q_2=q_1+1}^{9} l_{q_1}l_{q_2} = 2496$ between-group distances.

Histograms of within- and between-group distances are shown in Figures 1 and 2. Most of the within-group distances are below 40 whereas the between-group distances cluster in the interval 30 to 70.

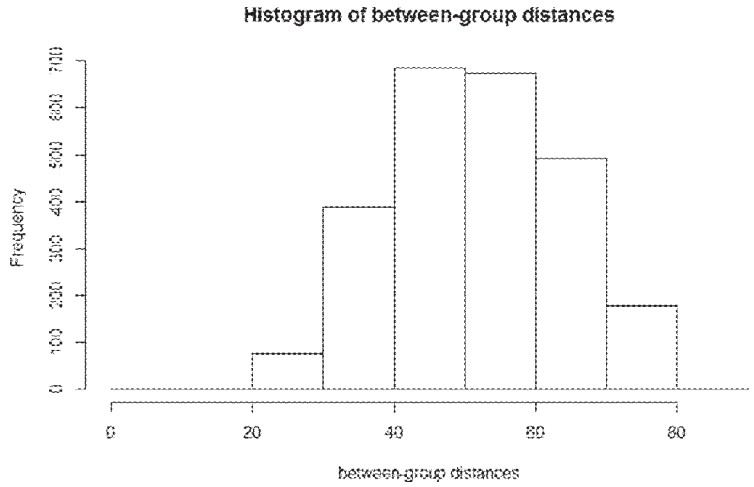

Figure 2: Histogram of beetween-group distances.

3. ESTIMATION OF THE LIKELIHOOD RATIO BASED ON DISTANCES

A likelihood ratio for the evaluation of the evidence of the striation marks based on the distances is described. The evidence E is two sets of measurements $(\mathbf{x}, \mathbf{y})$ of striation marks from B^{140}. These two sets are then summarised as $d(\mathbf{x}, \mathbf{y})$. The propositions are H_p, the two sets of measurements came from the same source (screwdriver in this context), and H_d, the two sets of measurements came from different sources. The likelihood ratio is then

$$\frac{f(d(\mathbf{x}, \mathbf{y}) \mid H_p)}{f(d(\mathbf{x}, \mathbf{y}) \mid H_d)}, \tag{7}$$

where f in the numerator is a probability density function modelling within-group distances and f in the denominator is a probability density function modelling between-group distances. Thus, the problem of the evaluation of evidence for discrete data has been transformed into one for continuous data, treating the 141 possible distances $0, \ldots, 140$, as a continuous variable.

These two probability density functions are estimated using the kernel method of density estimation. This method is a distribution-free method for the estimation of a probability density function. See Silverman (1986) for a general description of the kernel density estimation procedure and Aitken and Taroni (2004) for examples of its application in the evaluation of evidence for continuous data. The kernel density estimates are developed

from two training sets, the within-group set with 279 distances for the within-group density and between-group set with 2496 distances for the between-group density.

In general, let D be a training set $\mathbf{z} = (z_1, \ldots, z_n)$ of size n of univariate data from which it is desired to estimate a probability density function f. Then the kernel density estimate $\hat{f}$ of f at a point w is given by

$$\hat{f}(w; h, D) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{w - z_i}{h}\Big), \tag{8}$$

where K is a function, known as a *kernel function*, satisfying $K(u) \geq 0$ and $\int K(u)du = 1$. The parameter h (> 0) is a smoothing parameter (also known as *bandwidth*). The larger h is, the smoother the estimate is. Define $K_h(u) = K(u/h)$. Then the estimate may be written as

$$\hat{f}(w; h, D) = \frac{1}{n} \sum_{i=1}^{n} K_h(w - z_i). \tag{9}$$

The kernel function chosen here is the standard normal distribution

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{u^2}{2}\Big) \tag{10}$$

and the smoothing parameter is determined in an optimal manner using the R package 'KernSmooth'.

Let $\mathbf{x}$ and $\mathbf{y}$ be two sets of striation marks with distance $d(\mathbf{x}, \mathbf{y})$. The kernel density estimation procedure is used to estimate the probability density function of the distances $d(\mathbf{x}, \mathbf{y})$ between sets $\mathbf{x}$ and $\mathbf{y}$. The likelihood ratio for $d(\mathbf{x}, \mathbf{y})$ is the ratio of the estimates of the probability density functions, first, for the numerator based on the training set of within-group distances, and second, for the denominator based on the training group of between-group distances.

3.1 ESTIMATION DENSITY OF WITHIN-GROUP DISTANCES

Let h_w be the optimal bandwidth based on within-group distances. The associated density function is used for estimation of the density function

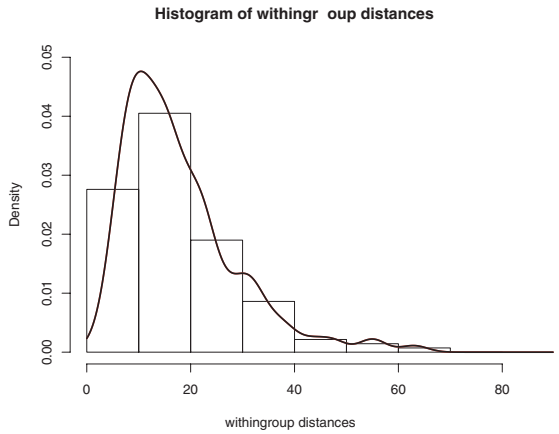

Figure 3: Histogram and kernel density estimate of distribution of within-group distances.

in the numerator. It is estimated using the training set D_w of within-group distances. This set consists of 279 within-group distances $\mathbf{z}_w = (z_{w1}, \ldots, z_{w,279})$. The kernel density estimate $\hat{f}_w$ for $d(\mathbf{x}, \mathbf{y})$ is then

$$\hat{f}_w(d(\mathbf{x}, \mathbf{y}); h_w, D_w) = \frac{1}{279\ h_w} \sum_{i=1}^{279} K\Big(\frac{d(\mathbf{x}, \mathbf{y}) - z_{wi}}{h_w}\Big). \tag{11}$$

The kernel density estimate, superimposed on the histogram of within-group distances, is shown in Figure 3.

3.2 ESTIMATION DENSITY OF BETWEEN-GROUP DISTANCES

Let h_b be the optimal bandwidth based on between-group distances. The associated density function is used for estimation of the density function in the denominator. It is estimated using the training set D_b of 2496 between-group distances $\mathbf{z}_b = (z_{b1}, \ldots, z_{b,2496})$. The kernel density estimate $\hat{f}_b$ for $d(\mathbf{x}, \mathbf{y})$ is then

$$\hat{f}_b(d(\mathbf{x}, \mathbf{y}); h_b, D_b) = \frac{1}{2496\ h_b} \sum_{i=1}^{2496} K\Big(\frac{d(\mathbf{x}, \mathbf{y}) - z_{bi}}{h_b}\Big). \tag{12}$$

The kernel density estimate, superimposed on the histogram of within-group distances, is shown in Figure 4.

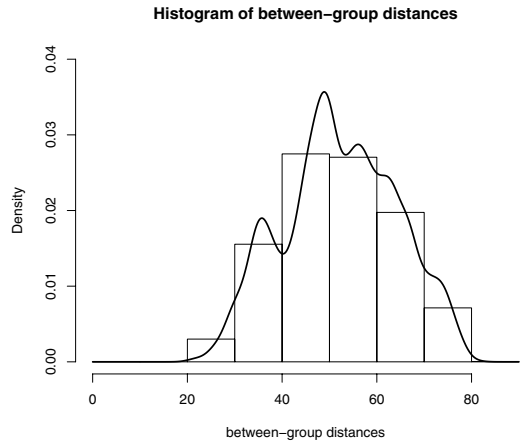

Figure 4: Histogram and kernel density estimate of distribution of between-group distances.

3.3 LIKELIHOOD RATIO AND RESULTS

The likelihood ratio is the ratio of the within-group density estimate and the between-group density estimate:

$$\frac{\hat{f}_w(d(\mathbf{x},\mathbf{y});h_w)}{\hat{f}_b(d(\mathbf{x},\mathbf{y});h_b)} = \frac{\frac{1}{279\,h_w}\sum_{i=1}^{279} K\left(\frac{d(\mathbf{x},\mathbf{y})-z_{wi}}{h_w}\right)}{\frac{1}{2496\,h_b}\sum_{i=1}^{2496} K\left(\frac{d(\mathbf{x},\mathbf{y})-z_{bi}}{h_b}\right)}. \tag{13}$$

As an example of the approach in practice, the distance between the first two sets of striation marks in the first group, items 1 and 2 in the training set of 75 items, is 11 and the likelihood ratio is 2×10^{10}.

The likelihood ratios of all possible pairwise comparisons of within-group and between-group marks were calculated. The results are presented in Table 1 in a tabular form for numbers of likelihood ratios, expressed as logarithms to base 10, within certain intervals. Logarithms of likelihood ratios less than 0 are supportive of the proposition of different sources (screwdriver) and logarithms of likelihood ratios greater than 0 are supportive of the proposition of same sources (screwdriver).

For the pairs of marks that come from the same group, more than 90% (253/279) of them result in likelihood ratios greater than 1 indicating, correctly, support for the proposition of the suspect screwdriver making the crime mark. For the pairs of marks that come from different groups,

Table 1: Likelihood ratios, expressed as logarithms to base 10, within certain intervals for 279 within-source comparisons and 2496 between-source comparisons.

Interval	$(-11,-9]$	$(-9,-7]$	$(-7,-5]$	$(-5,-3]$	$(-3,-2]$	$(-2,-1]$
Within-group comparison	0	0	0	0	0	9
Between-group comparison	2	22	70	85	54	1639
Interval	$(-1,0]$	$(0,5]$	$(5,10]$	$(10,20]$	$(20,30]$	$(30,40]$
Within-group comparison	17	113	47	81	11	1
Between-group comparison	503	121	0	0	0	0

more than 95% (2375/2496) of them result in likelihood ratios less than 1 indicating correctly, support for the proposition that the suspect screwdriver did not make the crime mark.

4. CONCLUSIONS

An approach for evidence evaluation for trace evidence in the form of hierarchical, longitudinal binary data has been described. An example of its use is given for the evaluation of evidence in the form of tool marks as measured by striation marks made by a tool. In the example described the tool is a screwdriver. An assessment of the performance of the method has shown that the support for a particular proposition as measured by a likelihood ratio is in the correct direction more than 90%; *i.e*, the likelihood ratio for support for the same source when the marks are from the same source is greater than 1 and the likelihood ratio for support for different sources when the marks are from different sources is less than 1 in more than 90% of the comparisons made. The method described is easily adaptable to other examples of hierarchical, longitudinal binary data. Another example is in forensic phonetics where the binary data would be the presence or absence of a speech characteristic with the records being several pieces of speech by each of several individuals in a sample from some population of interest.

ACKNOWLEDGEMENT

The authors are grateful to Professor N.D.K.Petraco for provision of the data on striation marks from screwdrivers.

REFERENCES

Aitken, C.G.G. and Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. In Applied Statistics. 53: 109-122; with corrigendum 665-666.

Aitken, C.G.G. and Taroni,F. (2004). Statistics and the Evaluation of Evidence for Forensic Scientists (2nd edition), John Viley & Sons Ltd., Chich- ester.

Aitken, C.G.G. and Gold, E. (2013). Evidence evaluation for discrete data. In Forensic Science International. 230: 147 - 155.

Petraco, N.D.K., Shenkin, P., Speir, J., Diaczuk, P., Pizzola, P.A., Gambino, C. and Petraco, l'\. (2012). Addressing the National Academy of Sciences' challenge: a method for statistical pattern comparison of striated tool marks. In Journal of Forensic Sciences. 57: 900 - 911.

Silverman, B.W. (1986). Density Estimation, Chapman & Hall, London.