



# THE UNIVERSITY *of* EDINBURGH

## Edinburgh Research Explorer

### Bayesian hierarchical random effects models in forensic science

**Citation for published version:**

Aitken, C 2018, 'Bayesian hierarchical random effects models in forensic science' *Frontiers in genetics*. DOI: 10.3389/fgene.2018.00126

**Digital Object Identifier (DOI):**

[10.3389/fgene.2018.00126](https://doi.org/10.3389/fgene.2018.00126)

**Link:**

[Link to publication record in Edinburgh Research Explorer](#)

**Document Version:**

Peer reviewed version

**Published In:**

Frontiers in genetics

**General rights**

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

**Take down policy**

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



# Bayesian hierarchical random effects models in forensic science

Colin Aitken<sup>1\*</sup>

<sup>1</sup>School of Mathematics, University of Edinburgh, United Kingdom

*Submitted to Journal:*  
Frontiers in Genetics

*Specialty Section:*  
Statistical Genetics and Methodology

*Article type:*  
Review Article

*Manuscript ID:*  
328314

*Received on:*  
27 Dec 2017

*Revised on:*  
06 Mar 2018

*Frontiers website link:*  
[www.frontiersin.org](http://www.frontiersin.org)

In review

---

### *Conflict of interest statement*

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest

### *Author contribution statement*

Sole author

### *Keywords*

Bayes Theorem, Evidence evaluation, Forensic Science, Hierarchical Models, Likelihood ratios, random effects, SAILR, statistics

### *Abstract*

Word count: 225

Statistical modelling of the evaluation of evidence with the use of the likelihood ratio has a long history. It dates from the Dreyfus case at the end of the nineteenth century through the work at Bletchley Park in the Second World War to the present day. The development received a significant boost in 1977 with a seminal work by Dennis Lindley which introduced a Bayesian hierarchical random effects model for the evaluation of evidence with an example of refractive index measurements on fragments of glass.

Many models have been developed since then. The methods have now been sufficiently well-developed and have become so widespread that it is timely to try and provide a software package to assist in their implementation. With that in mind, a project (SAILR: Software for the Analysis and Implementation of Likelihood Ratios) was funded by the European Network of Forensic Science Institutes through their Monopoly programme to develop a software package for use by forensic scientists world-wide that would assist in the statistical analysis and implementation of the approach based on likelihood ratios.

It is the purpose of this document to provide a short review of a small part of this history. The review also provides a background, or landscape, for the development of some of the models within the SAILR package and references to SAILR as made as appropriate.

### *Funding statement*

This work was supported by the European Network of Forensic Science Institutes 2015 Monopoly programme grant for Software for the Analysis and Implementation of Likelihood Ratios (SAILR), the Leverhulme Trust, grant number EM2016-027, and the Swiss National Science Foundation, grant number BSSGIO\_155809.

---

# Bayesian hierarchical random effects models in forensic science

C.G.G. Aitken<sup>1,\*</sup>

<sup>1,\*</sup> *School of Mathematics and Maxwell Institute, The University of Edinburgh, Peter Guthrie Tait Road, Edinburgh, EH9 3FD, UK*

Correspondence\*:  
C.G.G. Aitken  
cgga@ed.ac.uk

## 2 ABSTRACT

3 Statistical modelling of the evaluation of evidence with the use of the likelihood ratio has a  
4 long history. It dates from the Dreyfus case at the end of the nineteenth century through the  
5 work at Bletchley Park in the Second World War to the present day. The development received  
6 a significant boost in 1977 with a seminal work by Dennis Lindley which introduced a Bayesian  
7 hierarchical random effects model for the evaluation of evidence with an example of refractive  
8 index measurements on fragments of glass. Many models have been developed since then. The  
9 methods have now been sufficiently well-developed and have become so widespread that it is  
10 timely to try and provide a software package to assist in their implementation. With that in mind, a  
11 project (SAILR: Software for the Analysis and Implementation of Likelihood Ratios ) was funded  
12 by the European Network of Forensic Science Institutes through their Monopoly programme to  
13 develop a software package for use by forensic scientists world-wide that would assist in the  
14 statistical analysis and implementation of the approach based on likelihood ratios.

15 It is the purpose of this document to provide a short review of a small part of this history. The  
16 review also provides a background, or landscape, for the development of some of the models  
17 within the SAILR package and references to SAILR as made as appropriate.

18 **Keywords:** Bayes Theorem, Evidence evaluation, Forensic Science, Hierarchical Models, Likelihood Ratios, Random Effects, SAILR,  
19 **Statistics**

## 1 INTRODUCTION

20 Statistical analyses for the evaluation of evidence have a considerable history. It is the purpose of this  
21 document to provide a short review of a small part of this history. It brings together ideas from the last  
22 forty years for statistical models when the evidence is in the form of measurements and thus of continuous  
23 data. The data are also hierarchical with two levels. The first level is that of source, the origin of the data.  
24 The second level is of items within a source. The models used to represent the variability in the data are  
25 random effects models. The models are chosen from analyses of samples of sources from some relevant  
26 population. Finally, the analysis is Bayesian in nature with prior distributions for the parameters of the  
27 within-source distributions. The nature of the prior distributions is informed from training data based on  
28 the samples from the relevant population.

29 The remainder of the document is structured as follows. Section 2 provides a general introduction to  
 30 the likelihood ratio as a measure of the value of evidence. Section 3 provides a framework for models for  
 31 comparison and discrimination. Section 4 discusses the assessment of model performance. An Appendix  
 32 gives formulae for some of the more commonly used models.

## 2 THE VALUE OF EVIDENCE

33 Part of the role of a forensic scientist is to interpret evidence found at a crime scene in order to aid  
 34 fact-finders in a criminal case (*e.g.*, the judge or jury) in their decision making. The forensic scientist  
 35 may be asked to comment on the value of the evidence in the context of various competing statements  
 36 about the evidence, each of which may be true or false. Generally, a forensic scientist must consider two  
 37 competing statements relating to the evidence, one put forward by the prosecution in a criminal case, and  
 38 one put forward by the defence [Cook et al., 1998b]. These statements are known as *propositions*<sup>1</sup>. They  
 39 generally come in pairs that are mutually exclusive, though not necessarily exhaustive. For a debate about  
 40 the requirement, or otherwise, for the propositions to be exhaustive see Fenton et al. [2014a], Biedermann  
 41 et al. [2014], Fenton et al. [2014b].

One member of the pair is associated with the prosecution and conventionally denoted  $H_p$ . The other member of the pair is associated with the defence and conventionally denoted  $H_d$ . The evidence to be evaluated is denoted  $E$ <sup>2</sup>. The value of evidence is taken to be the relative values of the probability of the evidence if a proposition put forward by the prosecution is true and the probability of the evidence if a proposition put forward by the defence is true. However, evidence is not evaluated in isolation. There is always other information to be taken into account, including, for example, personal knowledge of the fact-finder. Denote this information by  $I$ . The value of the evidence, denoted  $V$  say, can then be written formulaically as

$$V = \frac{\Pr(E | H_p, I)}{\Pr(E | H_d, I)},$$

42 where  $\Pr$  denotes *Probability*. This ratio is known as the *likelihood ratio*.

43 The likelihood ratio is the method used by SAILR to evaluate evidence. SAILR (Software for the  
 44 Analysis and Implementation of Likelihood Ratios) is a user-friendly Graphical Interface (GUI) to calculate  
 45 numerical likelihood ratios in forensic statistics and its development under the direction of the Netherlands  
 46 Forensic Institute (NFI) was funded by the European Network of Forensic Science Institutes through their  
 47 Monopoly programme. The likelihood ratio is a generally accepted measure for the value of evidence in  
 48 much forensic case-work.

49 This representation of the value of evidence has a very good intuitive interpretation. Consider the odds  
 50 form of Bayes' Theorem in the forensic context of the evaluation of evidence. The odds form of Bayes'  
 51 Theorem then enables the prior odds (*i.e.*, prior to the presentation of  $E$ ) in favour of the prosecution  
 52 proposition  $H_p$  relative to the defence proposition  $H_d$  to be updated to posterior odds given  $E$ , the evidence  
 53 under consideration. This is done by multiplying the prior odds by the likelihood ratio. The odds form of  
 54 Bayes' Theorem may then be written as

<sup>1</sup> Other writers use the term *hypothesis* (see Section 2.7). The term *proposition* will be used except when there is an explicit need for the term *hypothesis*; see, for example, Section 3.1

<sup>2</sup> In ENFSI guidelines ENFSI [2015] 'findings' are distinguished from 'evidence'. 'Findings are the result of observations, measurements and classification that are made on items of interest.' '[E]vidence refers to outcomes of forensic examinations (findings) that, at a later point, may be used by legal decision-makers in a court of law to reach a reasoned belief about a proposition.' However, the word 'evidence' will be used in this document to refer to both situations for ease of nomenclature.

**Table 1.** Effect on prior odds in favour of  $H_p$  relative to  $H_d$  of evidence  $E$  with value  $V$  of 1,000. Reference to background information  $I$  is omitted.

Prior odds	$V$	Posterior odds
$Pr(H_p)/Pr(H_d)$		$Pr(H_p   E)/Pr(H_d   E)$
1/10,000	1,000	1/10
1/100	1,000	10
1 (evens)	1,000	1,000
100	1,000	100,000

$$\frac{Pr(H_p | E, I)}{Pr(H_d | E, I)} = \frac{Pr(E | H_p, I)}{Pr(E | H_d, I)} \times \frac{Pr(H_p | I)}{Pr(H_d | I)}. \quad (1)$$

55 The likelihood ratio (LR) is the ratio

$$\frac{Pr(H_p | E, I) / Pr(H_d | E, I)}{Pr(H_p | I) / Pr(H_d | I)} \quad (2)$$

56 of posterior odds to prior odds. It is the factor which converts the prior odds in favour of the prosecution  
 57 proposition to the posterior odds in favour of the prosecution proposition. The representation in (1) also  
 58 emphasises the dependence of the prior odds on other information  $I$ . Values of the  $LR > 1$  are supportive  
 59 of  $H_p$ , the proposition put forward by the prosecution. Values of the  $LR < 1$  are supportive of  $H_d$ , the  
 60 proposition put forward by the defence. The word ‘odds’ should be used advisedly. If  $H_p$  and  $H_d$  are not  
 61 exhaustive then the component probabilities  $Pr(H_p | E, I)$  and  $Pr(H_d | E, I)$  cannot be derived from this  
 62 ratio. All that can be said is that the posterior ratio is different from the prior ratio by a factor  $V$ .

63 An advantage of this formulation of evidence evaluation is the ease with which the effect of the addition  
 64 of new evidence can be determined. The posterior odds for one piece of evidence,  $E_1$  say, can be the prior  
 65 odds for a second piece of evidence,  $E_2$  say. Then (1) may be rewritten as

$$\frac{Pr(H_p | E_1, E_2, I)}{Pr(H_d | E_1, E_2, I)} = \frac{Pr(E_2 | H_p, E_1, I)}{Pr(E_2 | H_d, E_1, I)} \times \frac{Pr(H_p | E_1, I)}{Pr(H_d | E_1, I)}, \quad (3)$$

66 where the conditioning of the evaluation of  $E_2$  on  $E_1$  is made explicit.

67 An illustration of the effect of evidence with a value  $V$  of 1,000 on the odds in favour of  $H_p$  relative to  
 68  $H_d$  is given in Table 1.

69 The following quote is very pertinent.

70 ‘That approach does not ask the jurors to produce any number, let alone one that can qualify as a  
 71 probability. It merely shows them how a “true” prior probability would be altered, if one were in fact  
 72 available. It thus supplies the jurors with as precise and accurate an illustration of the probative force  
 73 of the quantitative data as the mathematical theory of probability can provide. Such a chart, it can be  
 74 maintained, should have pedagogical value for the juror who evaluates the entire package of evidence

75 solely by intuitive methods, and who does not himself attempt to assign a probability to the “soft”  
76 evidence.’ Kaye [1979].

77 The ‘it’ in this context is a chart depicting, in numerical terms, how much the prior odds in favour  
78 of a proposition is enhanced by the evidence being evaluated. This is a graphical equivalent of Table 1.  
79 The mathematical tool for devising such a chart is Bayes’ Theorem. These remarks of Kaye’s refer to  
80 characteristics of the general method for the evaluation of evidence that is the likelihood ratio. They do  
81 not refer to a particular case. For example, it is not possible to comment on the accuracy of a likelihood  
82 ratio estimation in a particular case because the true value of the likelihood ratio is not known nor can it be  
83 known. It is, however, possible to refer to the accuracy of a method and performance assessment in general  
84 is discussed in Section 4.

85 The use of a likelihood ratio for the evaluation of evidence is not a new idea. In the Dreyfus case  
86 [Champod et al., 1999], it was argued that

87 . . . since it is absolutely impossible for us [the experts] to know the *a priori* probability, we cannot  
88 say: this coincidence proves that the ratio of the forgery’s probability to the inverse probability is a  
89 real value. We can only say: following the observation of this coincidence, this ratio becomes  $X$  times  
90 greater than before the observation. [Darboux et al., 1908]

91 The ‘ratio’ in this quotation is the odds in favour of one proposition over another, The  $X$  refers to the  
92 likelihood ratio. The posterior odds in favour of the proposition is then  $X$  times the prior odds.

93 The ideas were also used in the work of I.J. Good and A.M. Turing as code-breakers at Bletchley Park  
94 during World War II [Good, 1979].

## 95 2.1 Background Information

96 The likelihood ratio updates the prior odds, those before consideration of evidence  $E$ , to posterior odds,  
97 which take  $E$  into account. The posterior odds are the odds with which, ultimately, the fact-finder is  
98 concerned. If the likelihood ratio multiplied by the prior odds is larger than one, then the probability of  
99  $H_p$  given the evidence is larger than that of  $H_d$  given the evidence. As these propositions may not be  
100 exhaustive their explicit values, rather than their relative value, may not be known. It is the responsibility  
101 of the fact-finder to determine a value for the prior odds. The prior odds can then be combined with  
102 the likelihood ratio to obtain posterior odds. A forensic scientist is concerned only with the value of the  
103 evidence as expressed by the likelihood ratio so cannot usually comment on the value of the posterior odds.  
104 The likelihood ratio is considered as the strength of support of the evidence for one of the two propositions  
105  $H_p$  or  $H_d$ .

106 The application of this form to a specific case is crucially dependent on the background information  
107  $I$ . However, the background information available to each person is different. In part, this is because  
108 each person is different. In part it is because of professional differences. The information that a forensic  
109 scientist should use for their determination of the likelihood ratio is different from that which a fact-finder,  
110 such as judge or jury member, should use for their determination of the odds in favour of the prosecution  
111 proposition. There are differences in the background information available to these participants in the  
112 judicial process but these differences have no effect on the posterior odds in favour of the prosecution  
113 proposition

Let  $I = I_a \cup I_b$  where  $I_a$  is background information available to the forensic scientist and  $I_b$  is background  
information available to the fact-finder. There will be information available to both, the intersection  $I_a \cap I_b$

is not empty. It can then be shown [Aitken and Nordgaard, 2017] that the posterior odds may be written in the form

$$\frac{\Pr(H_p | E, I)}{\Pr(H_d | E, I)} = \frac{\Pr(E | H_p, I_b)}{\Pr(E | H_d, I_b)} \times \frac{\Pr(H_p | I_a)}{\Pr(H_d | I_a)}.$$

114 The fact-finder and the forensic scientist have to treat the common information ( $I_a \cap I_b$ ) with appropriate  
115 discretion.

## 116 2.2 Uniqueness of the Likelihood Ratio

117 The role of the likelihood ratio as the factor that updates the prior odds to the posterior odds has a very  
118 intuitive interpretation. There is also a mathematical derivation that shows it, or a function of it such as the  
119 logarithm, is the only way to update evidence. It was shown many years ago by I.J. Good in two brief notes  
120 in the *Journal of Statistical Computation and Simulation* [Good, 1989a,b] repeated in Good [1991] and  
121 in Aitken and Taroni [2004] that, with some very reasonable assumptions, the assessment of uncertainty  
122 inherent in the evaluation of evidence leads inevitably to the likelihood ratio as the only way in which this  
123 can be done.

124 Consider evidence  $E$  which it is desired to evaluate in the context of two mutually exclusive propositions  
125  $H_p$  and  $H_d$ . Denote the value of the evidence by  $V$ . As always, the value will depend on background  
126 information  $I$  but this will not be stated explicitly. There are other assumptions implicit in this approach,  
127 namely that there is a probability that can be associated with evidence and one that is dependent on  
128 propositions and only on propositions (and background information). Another assumption is that  $V$  is a  
129 function only of the probability of  $E$ , given  $H_p$  to be true, and of the probability of  $E$ , given  $H_d$  to be true.

Let  $x = \Pr(E | H_p)$  and  $y = \Pr(E | H_d)$  where  $I$  is omitted for ease of notation. The assumption that  $V$  is a function only of these probabilities can be represented mathematically as

$$V = f(x, y)$$

130 for some function  $f$ .

131 Now, consider another piece of evidence  $T$  which is irrelevant to  $E$ , to  $H_p$  and to  $H_d$ . Irrelevance is taken  
132 in the probabilistic context to be equivalent to independence so that  $T$  may be taken to be independent of  
133  $E$ , of  $H_p$  and of  $H_d$ . It is then permissible for  $\Pr(T)$  to be given notation which does not refer to any of  
134  $E$ ,  $H_p$  or  $H_d$ . Thus, let  $\Pr(T)$  be denoted by  $\theta$ . Then

$$\begin{aligned} \Pr(E, T | H_p) &= \Pr(E | H_p) \Pr(T | H_p) && \text{by the independence of } E \text{ and } T \\ &= \Pr(E | H_p) \Pr(T) && \text{by the independence of } T \text{ and } H_p \\ &= x \theta. \end{aligned}$$

Similarly,

$$\Pr(E, T | H_d) = y \theta.$$

135 The value of  $(E, T)$  is  $f(\theta x, \theta y)$  by the definition of  $f$ . However, evidence  $T$  is irrelevant and has no  
136 effect on the value of evidence  $E$ . Thus, the value of the combined evidence  $(E, T)$ ,  $f(\theta x, \theta y)$ , is equal to  
137 the value  $V$  of  $E$ ,  $f(x, y)$ , and

$$V = f(x, y) = f(\theta x, \theta y)$$



138 for all  $\theta$  in the interval  $[0,1]$  of possible values of  $\Pr(T)$ .

The only class of functions of  $(x, y)$  for which this can be said to be the case is the class which are functions of  $x/y$  or

$$\Pr(E | H_p) / \Pr(E | H_d)$$

139 which is the likelihood ratio. Hence the value  $V$  of evidence has to be a function of the likelihood ratio. It  
 140 has been argued [Lund and Iyer, 2017] that the forensic community view the likelihood ratio as only one  
 141 possible tool for communication with decision makers. The argument of Good shows that it is the only  
 142 logically admissible form of evaluation.

### 143 2.3 Weight of Evidence

144 An interesting note of terminology can be mentioned here. It is common in some legal circles to talk of  
 145 the *weight of evidence*. The concept of weight of evidence is an old idea. The term *weight of evidence*  
 146 should be used for the logarithm of the likelihood ratio. The terminology was first used by Charles Sanders  
 147 Peirce [Peirce, 1878]. The likelihood ratio is the *value* of the evidence and its logarithm is the *weight* of the  
 148 evidence. The logarithm of the likelihood ratio has the pleasingly intuitive operation of additivity when  
 149 converting the logarithm of the prior odds in favour of a proposition to the logarithm of the posterior odds  
 150 in favour of the proposition.

$$\log \left\{ \frac{\Pr(H_p | E)}{\Pr(H_d | E)} \right\} = \log \left\{ \frac{\Pr(E | H_p)}{\Pr(E | H_d)} \right\} + \log \left\{ \frac{\Pr(H_p)}{\Pr(H_d)} \right\}, \quad (4)$$

151 with  $I$  omitted. When considering the scales of justice it is the logarithm of the probabilities of the evidence  
 152 given each of the two competing propositions that should be put in the scales, not the probabilities.

### 153 2.4 Terminology for evidence

154 The evidence under consideration in this document and within the SAILR project is evidence that could  
 155 have been transferred either from the crime scene to the criminal or from the criminal to the crime scene.  
 156 Evidence that could have been so transferred is in the form of traces. Thus it has two names *transfer* or  
 157 *trace* evidence. The evidential material discussed here is in the form of individual items. Thus, there may  
 158 be a finite number of items, such as tablets or sachets of drugs or fragments of glass. Alternatively, the  
 159 evidence may be a single measurement such as that of a DNA profile.

160 Consider the situation in which a crime has been committed, there is a crime scene and the investigation  
 161 has reached the stage where a suspect has been identified. Trace evidence, denoted  $E$ , of a particular  
 162 type has been found at the crime scene and on the suspect and its value is of interest. The evidence  $E$   
 163 may be partitioned into two parts, that found at the crime scene and that found in association with the  
 164 suspect. In practice, the terminology takes a different form which depends on whether the source of the  
 165 evidence is known or not known. A distinction is also drawn between evidential material and the evidence  
 166 for evaluation. Evidence for evaluation is the observations made on the material. Only evidence which is  
 167 in the form of measurements and thus represented by continuous data is considered here. Other factors  
 168 such as the locations in which the material was found and the quantity of the material are not considered.  
 169 Evidence of a discrete nature such as binary data as in the presence or absence of striation marks is also not  
 170 considered.

171 Evidence whose source is known is called *control* evidence  $E_c$ . Evidence whose source is not known is  
 172 called *recovered* evidence  $E_r$ . Measurements on  $E_c$  are conventionally denoted  $\mathbf{x}$  where  $\mathbf{x} = (x_1, \dots, x_m)$

173 are  $m$  sets of measurements and where  $x_i, i = 1, \dots, m$  may be univariate or multivariate. Measurements  
 174 on  $E_r$  are conventionally denoted  $\mathbf{y}$  where  $\mathbf{y} = (y_1, \dots, y_n)$  are  $n$  sets of measurements and where  
 175  $y_j, j = 1, \dots, n$  may be univariate or multivariate<sup>3</sup>.

176 For an evaluative comparison of  $\mathbf{x}$  and  $\mathbf{y}$ , background data  $\mathbf{z}$  are needed. These background data should  
 177 be a representative sample of all possible sources from the population of interest, known as the *relevant*  
 178 population. Ideally, the sample should be a random sample but this is rarely possible for practical reasons.  
 179 The sample is often what might be called a *convenience* sample. If the convenience sample can be  
 180 demonstrated to be composed of sources chosen in a manner independent of the case under investigation  
 181 then the inference based on the comparison of  $\mathbf{x}$  and  $\mathbf{y}$  informed on  $\mathbf{z}$  should be valid. Computation of the  
 182 likelihood ratio requires data files from  $\mathbf{x}$ ,  $\mathbf{y}$  and  $\mathbf{z}$

183 One example of evidence in the form of multivariate data relates to glass elemental content. Such data  
 184 are often subjected to a logarithmic transformation after taking the ratios of a particular elemental content  
 185 to the oxygen content, for example,  $\log_{10}(NaO) = \log_{10}(Na/O)$ . These measurements can be for each  
 186 of  $m$  fragments of control evidence and for each of  $n$  fragments of recovered evidence [Zadora et al.,  
 187 2014]. This evidence can be multivariate as there can be several ratios measured for each fragment, e.g.,  
 188  $\log_{10}(NaO), \log_{10}(MgO)$  and  $\log_{10}(AlO)$ . The control evidence is the measurements from a number  
 189  $m$  of fragments of glass from a broken window at a crime scene; the source of the fragments is known  
 190 to be the window, items within source are the fragments. The recovered evidence is the measurements  
 191 from a number  $n$  of fragments of glass found in association with a suspect, for example on clothing  
 192 identified as theirs. The source of the fragments of glass from the suspect is unknown. It may or may not  
 193 have come from the window at the crime scene. A second example could be the measurements of colour  
 194 chromaticity coordinates on fibres and the evidence is bivariate [Martyna et al., 2013]. There are three  
 195 colour chromaticity coordinates. The sum of their values is fixed so given the values of any two, the third  
 196 is known. Control evidence is the measurements of colour chromaticity coordinates from a number  $m$  of  
 197 fibres from an article of clothing belonging to a suspect; the source is the article, the items are the fibres.  
 198 Recovered evidence is the measurements of colour chromaticity coordinates from a number  $n$  of fibres  
 199 found at a crime scene. Thus control evidence may be found at a crime scene or in association with a  
 200 suspect. Similarly, recovered evidence may be found at a crime scene or in association with a suspect.

201 Often the number  $m$  of control items can be chosen by the investigator. The number  $n$  of recovered items  
 202 may be determined by what is available and the investigator has little choice in the selection of this number.  
 203 If the number of recovered items is large, in some sense, and perhaps so large as for it to be impractical to  
 204 count or analyse them, then the investigator may decide to select  $n$  items where  $n$  is less than the number  
 205 available. Procedures for the choice of  $n$  and the manner of selection of the items are not discussed in  
 206 this document or SAILR other than to note that the evidence selected should be representative of the total  
 207 evidence available as far as is possible. Further information is available in Aitken and Taroni [2004] and  
 208 references therein.

209 The likelihood ratio  $V$  for the comparison of  $\{\mathbf{x}, \mathbf{y}\}$  where  $E$  is replaced by  $\{\mathbf{x}, \mathbf{y}\}$  is then

$$V = \frac{\Pr(\mathbf{x}, \mathbf{y} | H_p)}{\Pr(\mathbf{x}, \mathbf{y} | H_d)}, \quad (5)$$

210 where again the conditioning on  $I$ , the background information, has been omitted for clarity of notation.

<sup>3</sup> The use of  $\mathbf{x}$  and  $\mathbf{y}$  here is not to be confused with the use of  $x = \Pr(E | H_p)$  and  $y = \Pr(E | H_d)$  in Section 2.2.

211 Often, the propositions being considered are  $H_p$  that the control and recovered evidence are from the same  
 212 source and  $H_d$  that the control and recovered evidence are from different sources. In such a circumstance,  
 213  $\mathbf{x}$  and  $\mathbf{y}$  may be assumed independent if  $H_d$  is true as they come from different sources. Then (5) may be  
 214 written as

$$V = \frac{\Pr(\mathbf{x}, \mathbf{y} \mid H_p)}{\Pr(\mathbf{x} \mid H_d) \Pr(\mathbf{y} \mid H_d)}. \quad (6)$$

215 If  $\mathbf{x}$  and  $\mathbf{y}$  are continuous data, as is the case when the evidence is in the form of measurements rather than  
 216 counts, the probabilities in the numerator and denominator are replaced by probability density functions,  
 217 denoted say  $f(\mathbf{x}, \mathbf{y})$  for the joint density and  $f(\mathbf{x})$  and  $f(\mathbf{y})$  for the marginal distributions. The continuous  
 218 analogue of (6) can then be written as

$$V = \frac{f(\mathbf{x}, \mathbf{y} \mid H_p)}{f(\mathbf{x} \mid H_d) f(\mathbf{y} \mid H_d)}. \quad (7)$$

219 In most cases, the full specification of the probability density function is unknown. The form of the  
 220 distribution may be known or a reasonable assumption of its form may be made. For example, it may  
 221 be known or can be assumed that the appropriate distribution is a Normal distribution. This assumption  
 222 may be based on the unimodal, symmetric nature of the distribution. If the distribution has a positive  
 223 skew then a transformation to normality with a logarithmic transformation of the data may be possible  
 224 before consideration of the likelihood ratio. However, the parameters may neither be known nor able to be  
 225 assumed known.

The numerator of (7) may be written as  $f(\mathbf{x}, \mathbf{y} \mid H_p) = f(\mathbf{y} \mid \mathbf{x}) f(\mathbf{x} \mid H_p)$ . Since the distribution of  $\mathbf{x}$  is independent of whether  $H_p$  or  $H_d$  is true,  $f(\mathbf{x} \mid H_p) = f(\mathbf{x} \mid H_d)$  and (7) may be written as

$$f(\mathbf{y} \mid \mathbf{x}, H_p) / f(\mathbf{y} \mid H_d).$$

226 See (18) for an example.

## 227 2.5 Training data

228 When parameters are not known, information about their possible values may be obtained from data  
 229 independent of the crime but thought to be relevant for consideration of the variability in the measurements  
 230 of the data comprising the evidence. These data are the *training data* or *background data* and are  
 231 conventionally denoted  $\mathbf{z}$ . These data are considered to be a sample from a population, known as a  
 232 *relevant* population. There is considerable continuing debate as to how to choose a population that is  
 233 relevant for a particular crime and, once chosen, how a sample may be chosen from it to be a representative  
 234 sample of the population. See, for example, *R. v. T* [2010] EWCA 2439, where the debate related to the  
 235 choice of populations of shoes relevant for the consideration of evidence of shoeprints. Often the sample is  
 236 a convenience sample; see Section 2.4.

237 An alternative procedure would be to sample anew each time from a population deemed relevant to the  
 238 case under investigation. A relatively early example of this is the investigation of a murder in Biggar, a  
 239 town near Edinburgh, in 1967. A bite mark found on the breast of a young girl who had been murdered had  
 240 certain characteristic marks, indicative of the conformation of the teeth of the person who had bitten her.  
 241 A 17-year-old boy was found with this conformation and he became a suspect. Examination of 90 other  
 242 boys of the suspect's age showed that the particular conformation was not at all common. The 90 other

243 boys could be considered as a sample from a relevant population. Further details are available in Harvey  
244 et al. [1968]. However, in most individual investigations it is not practical to obtain such a bespoke relevant  
245 population.

## 246 2.6 Hierarchy of evidence

247 Often, with measurements, the training data can be thought of as a set of sources of items. Measurements  
248 are made of one or more characteristics of the items. For example, consider again the composition of the  
249 elemental ratio of various elements of glass to oxygen for glass fragments from a set of windows. The  
250 items are glass fragments. A source would be a window. The training set is a set of windows. The set of  
251 windows is a sample from some population of windows, deemed relevant for crimes involving windows.  
252 The measurements are said to be *hierarchical* with two levels. One level is the fragment of glass within  
253 a window. Variation amongst measurements of fragments within a window is known as *within-group*  
254 or *within-source variation*. The second level is the window. Variation amongst measurements between  
255 windows is known as between-group or between-source variation. Measurements are taken from an item  
256 (fragments of glass) within a source (window). Notationally, the training data  $\mathbf{z}$  has two indices, one for each  
257 level and may be represented as  $\mathbf{z} = \{z_{k\ell}; k = 1, \dots, g, \ell = 1, \dots, h\}$  where  $g$  is the number of sources  
258 in the training set and  $h$  is the number of measurements within sources. The number of measurements  
259 within sources need not necessarily be constant though it is computationally convenient if this can be  
260 arranged during the compilation of the training set. Occasionally there may be further levels, for example  
261 measurement error.

## 262 2.7 Propositions

263 As well as evidence ( $E$ ) and background information  $I$ , evidence evaluation depends on propositions  $H_p$   
264 and  $H_d$ . There are different types of propositions, also known as *levels*. Both propositions ( $H_p$  and  $H_d$ ) in  
265 any particular situation for the evaluation of evidence are at the same level. There are four different levels  
266 of propositions, known respectively as *offence* level, *activity* level, *source* level and *sub-source* level (Cook  
267 et al. [1998a]; Evett et al. [2000]).

268 The levels, with examples, are described as follows.

- 269 • *Offence level*: the propositions may be that the defendant is guilty of an offence (truly guilty, not just  
270 declared guilty) and that the defendant is innocent (truly innocent, not just declared not guilty).
- 271 • *Activity level*: the propositions concern an activity by the defendant which may or may not be a criminal  
272 act. An example of a pair of activity level propositions could be that the defendant hit the victim and  
273 that the defendant did not hit the victim.
- 274 • *Source level*: the propositions concern the source of evidential material. There is no consideration  
275 of the activity that may have led to the material being where it was found. An example of a pair of  
276 source level propositions could be that blood found at the scene of a crime came from the defendant  
277 and that the blood found at the scene of the crime came from some other source, unrelated to the  
278 defendant. Note that this example is one in which the two propositions are not exhaustive; relatives of  
279 the defendant are not included. SAILR can only be used for likelihood ratio computation on source  
280 level.
- 281 • *Sub-source level*: the propositions concern material for which it is not possible to identify a source. An  
282 example of a pair of sub-source level propositions could be that DNA found at a crime scene came  
283 from the defendant and that DNA found at the crime scene came from some other source, unrelated to

284 the defendant. The quantity of material found is insufficient to identify its source, *e.g.*, whether it came  
285 from blood or semen.

### 3 FRAMEWORK FOR MODELS

286 The likelihood ratio may be used in the context of forensic science in two different ways, that of comparison  
287 and that of discrimination. For comparison, two pieces of evidence found in different places are compared  
288 to see if they had a common source. For discrimination, one piece of evidence is compared with several  
289 sets of training or background data from different sources to see from which source the evidence may have  
290 come.

291 Most of the models described here are so-called *feature-based models*. These are models developed from  
292 the measurements (features) on the evidential material. Other models described are so-called *score-based*  
293 *models*. There may be occasions with multivariate data when a feature-based model is not tractable, *e.g.*,  
294 multidimensional binary data where the number of possible models is unmanageable. On such occasions,  
295 the distance, denoted  $d(\mathbf{x}, \mathbf{y})$ , between control ( $\mathbf{x}$ ) and recovered ( $\mathbf{y}$ ) data can be used instead.

#### 296 3.1 Comparison for feature-based models

297 3.1.1 The likelihood ratio approach for continuous univariate evidential data with Normal  
298 distributions for the means and known variances

299 A common problem occurs in forensic science when the prosecution and defence propositions concern  
300 whether two objects are from the same source or from different sources. For example, if a glass fragment is  
301 found on a suspect and there is a broken window at the crime scene, one proposition might be that the glass  
302 fragment found on the suspect came from the window at the crime scene, and the other proposition might  
303 be that the glass fragment came from some other window. The evidence is given by a set of measurements  
304 from the glass fragment found on the suspect (the recovered sample) and a set of measurements from one  
305 or more glass fragments from the crime scene (the control sample). The problem is one of comparison.

306 The structure of these models reflects the hierarchical nature of the underlying data (measurements and  
307 variation within a source and then variation between sources). Using a distribution for the means  $\theta_1$  and  $\theta_2$   
308 in this way accounts for variance within source ( $\sigma^2$ ) and variance between sources ( $\tau^2$ ).

309 The problem for the fact-finder is to determine which of the two propositions ( $H_p$  or  $H_d$ ) is more likely,  
310 given all of the evidence in the case. Denote the other evidence and background information by  $I$  as before.  
311 The fact-finder can consider which proposition is more likely by considering the relative size of the two  
312 probabilities  $\Pr(H_p | \bar{x}, \bar{y}, I)$  and  $\Pr(H_d | \bar{x}, \bar{y}, I)$  (technically, in cases where the statistical assumptions  
313 include knowledge of the variances  $\sigma^2$  and  $\tau^2$  and of a Normal distribution for the measurements, the means  
314 of the control and recovered samples are sufficient statistics so can be used in place of the measurements  $\mathbf{x}$   
315 and  $\mathbf{y}$ ). Let  $f(\bar{x}, \bar{y} | H_p, I)$  be the joint probability density function of  $\bar{x}$  and  $\bar{y}$ , given proposition  $H_p$  and  $I$   
316 and let  $f(\bar{x}, \bar{y} | H_d, I)$  be the joint probability density function of  $\bar{x}$  and  $\bar{y}$  given proposition  $H_d$  and  $I$ . In  
317 this context (1) may be represented as

$$\frac{P(H_p | \bar{x}, \bar{y}, I)}{P(H_d | \bar{x}, \bar{y}, I)} = \frac{f(\bar{x}, \bar{y} | H_p, I)}{f(\bar{x}, \bar{y} | H_d, I)} \times \frac{P(H_p | I)}{P(H_d | I)}, \quad (8)$$

318 where  $E$  is replaced by  $(\bar{x}, \bar{y})$ . For examples where the within-source variance is not known, the sample  
319 variances of  $\mathbf{x}$  and  $\mathbf{y}$  will also be included in the representation.

320 Denote the common mean of the measurements under the prosecution proposition by  $\theta_1 = \theta_2 = \theta$ . The  
 321 likelihood ratio  $V$  is given by (7). This may be rewritten as

$$V = \frac{\int f(\bar{x} | \theta)f(\bar{y} | \theta)f(\theta)d\theta}{\int f(\bar{x} | \theta_1)f(\theta_1)d\theta_1 \int f(\bar{y} | \theta_2)f(\theta_2)d\theta_2}, \quad (9)$$

322 where the dependence on  $I$  has been suppressed for ease of notation. The analytical form of this likelihood  
 323 ratio, given the independence and Normality assumptions detailed above, is given by Lindley [1977]. The  
 324 density functions  $f(\bar{x} | \theta)$  and  $f(\bar{y} | \theta)$  are taken to be density functions of a Normal distribution. Note that  
 325 when the prosecution proposition is chosen the random variables  $\bar{X}$  and  $\bar{Y}$ , of which  $\bar{x}$  and  $\bar{y}$  are realisations,  
 326 are conditionally independent, conditional on  $\theta$ . They are independent if it is known they are from the same  
 327 source. The distributions associated with these density functions are termed the within-source distributions,  
 328 because they account for the within-source variability. The distribution associated with the density function  
 329  $f(\theta)$  is termed the between-source distribution because it accounts for between-source variability, and it is  
 330 a prior distribution for  $\theta$ . The use of a between-source distribution allows the rarity of the data  $\mathbf{x}$  and  $\mathbf{y}$  to  
 331 be taken into account when assessing the strength of the evidence; see (13) for an example. Information  
 332 to assist with the estimation of the prior distribution is contained in the training set. If the control and  
 333 recovered samples have similar means, and the mean is unusual, then the strength of evidence supporting  
 334 the proposition that the samples are from the same source should be stronger than if the mean is relatively  
 335 common.

336 A solution to this problem of the comparison of sources in the case where the measurements are univariate  
 337 and are assumed to be independent and Normally distributed was developed by Lindley [1977]. Some  
 338 details are given in the Appendix; see (12) and (13). Denote the  $m$  measurements on the control sample by  
 339  $\mathbf{x} = (x_1, \dots, x_m)$  and the  $n$  measurements on the recovered sample by  $\mathbf{y} = (y_1, \dots, y_n)$ . The corresponding  
 340 means of each of these samples are denoted  $\bar{x}$  and  $\bar{y}$ . The two propositions to be considered are at the  
 341 source level and are:

- 342 •  $H_p$ : the control and recovered sample are from the same source.
- 343 •  $H_d$ : the control and recovered sample are from different sources.

344 Lindley's solution assumes that the means  $\bar{x}$  and  $\bar{y}$  of the control and recovered samples are sample  
 345 means of data, whose corresponding random variables have Normal distributions with means  $\theta_1$  (control)  
 346 and  $\theta_2$  (recovered), respectively, and variances  $\sigma^2/m$  (control) and  $\sigma^2/n$  (recovered). The variance  $\sigma^2$  is a  
 347 within-group (e.g. within window) variance. The means  $\theta_1$  and  $\theta_2$  are the means of the groups associated  
 348 with  $\mathbf{x}$  and  $\mathbf{y}$  in the terminology of hierarchical data. Variability between groups has also to be considered.  
 349 This is done with consideration of the variation in the group means. The two means  $\theta_1$  and  $\theta_2$  are also  
 350 assumed to be realisations of a random variable which is Normally distributed, this time with mean  $\mu$  and  
 351 variance  $\tau^2$ . At present the variances  $\sigma^2$  and  $\tau^2$  are assumed known. Also, the within-group variance  $\sigma^2$  is  
 352 assumed constant within groups. An expression for the likelihood ratio if the between-group distribution is  
 353 not Normal but is represented with a general distribution  $p(\cdot)$ , with second derivative  $p''(\cdot)$  is given by (14).

354 An extension using kernel density estimation has been derived to allow for a general non-Normal between-  
 355 group distribution (15). Checks of the distributional assumptions and estimation of hyperparameters are  
 356 made using a training set of groups which are assumed to be a random sample of groups (sources) from  
 357 some relevant population. Later work (e.g., Bozza et al. [2008] with an extension to multivariate data, (24))  
 358 relaxes the assumption that  $\sigma^2$  and  $\tau^2$  are known.

359 The likelihood ratio can be used to assess evidence in a criminal trial and hence is a solution to the  
360 comparison of sources problem; Lindley [1977].

361 This approach for evidence evaluation based on the likelihood ratio is different from an approach based  
362 on hypothesis testing. The likelihood ratio approach has many advantages; a discussion of these can be  
363 seen in Aitken and Stoney [1991] and Aitken and Taroni [2004]. One such advantage is that the likelihood  
364 ratio has no dependence on an arbitrary cut off point (*e.g.*, 5% significance). Another advantage is that the  
365 use of a likelihood ratio reduces the risk that a transposition of the conditional probabilities (also known  
366 as the prosecutor's fallacy) occurs, a transposition which confuses the probability of finding the evidence  
367 on an innocent person with the probability of the innocence of a person on whom the evidence has been  
368 found. In addition, the likelihood ratio provides a method of comparing the likelihood of the evidence  
369 under the propositions of both the prosecution and the defence. This guards against potentially misleading  
370 situations when the likelihood under only one of these propositions is considered. Finally, an approach  
371 based on the likelihood ratio ensures equality of treatment of both propositions. In a procedure based on  
372 hypothesis testing, a null hypothesis is assumed true unless sufficient evidence is found to reject it at a  
373 pre-specified significance level. Often, the null hypothesis is that of a common source,  $\theta_1 = \theta_2$  in Lindley's  
374 example. This is the prosecution proposition. Thus the burden of proof is placed on the defence to put  
375 forward sufficient evidence to enable rejection of the prosecution proposition, contrary to the dictum of  
376 'proof beyond reasonable doubt'. The prosecution need prove nothing.

### 377 3.1.2 The likelihood ratio approach for other forms of continuous evidential data, including 378 multivariate data

379 Later work on evidence evaluation has extended the work done in Lindley [1977] to cover other data  
380 types, allowing for different forms of the within and between source distributions (Aitken and Lucy [2004];  
381 Aitken et al. [2006]; Aitken et al. [2007a]). In Bozza et al. [2008] and Alberink et al. [2013], extensions are  
382 given so that the between-source distribution in (9) becomes a function of both the mean and the variance.  
383 This allows for variation in the variance of samples from different sources. All of these extensions assume  
384 that the  $m$  measurements  $\mathbf{x}$  are independent and that the  $n$  measurements  $\mathbf{y}$  are independent. Methods for  
385 autocorrelated data types, such as measurements associated with drug traces on banknotes are described in  
386 Wilson et al. [2014, 2015].

387 For multivariate measurements which are independent and which have a multivariate Normal distribution  
388 the analytical form is derived in Aitken and Lucy [2004]. The likelihood ratio is given for two forms of the  
389 distribution of the mean between sources. The first form assumes multivariate Normality, and the second  
390 form uses nonparametric kernel density estimation. The within-source variance is assumed constant over  
391 all sources.

392 When there are several variables graphical models may be used to reduce the number of parameters  
393 needing to be estimated. The kernel density approach given in Aitken and Lucy [2004] can then be  
394 used to calculate likelihood ratios for the subsets of variables as indicated by the graphical models. The  
395 graphical model considers partial correlations amongst the variables and partitions these variables into  
396 overlapping subsets known as *cliques*. The overall distribution may then be represented as a function of the  
397 distributions over the cliques. These clique distributions have very few variables each (*e.g.*, one, two or  
398 three; and the overall likelihood ratio is then a product of likelihood ratios which are based on one-, two- or  
399 three-dimensional data (Aitken et al. [2007]). Such a process for the reduction of dimension is necessary to  
400 avoid the curse of dimensionality whereby very large data sets are needed for the estimation of parameters  
401 in a multi-dimensional parameter set.

402 In Aitken et al. [2006] the multivariate methods used in Aitken and Lucy [2004] assuming Normality  
 403 are extended further to allow for another level of variance (e.g., measurement error) to be taken into  
 404 account, giving a three-level model. A model assuming an exponential distribution for between-sources in  
 405 a three-level model is assumed in Aitken et al. [2007a] and the analytical form of the likelihood ratio is  
 406 derived. Variation between the means of samples from different sources, variation between the means of  
 407 different samples taken from the same source and variation within repeated measurements on the same  
 408 sample are taken into account.

409 Relaxation of the assumption that samples from different sources will have the same variance means that  
 410 an analytical solution is not available. Measurements are assumed multivariate and independently Normally  
 411 distributed as before but the between-source (prior) distribution is taken to be the product of a multivariate  
 412 Normal distribution (for the mean of the between-source distribution) and an inverse Wishart distribution  
 413 (for the covariance of the between-source distribution). In this way, variation of covariances, as well as  
 414 means, between different sources is taken into account. An analytical form of the likelihood ratio is not  
 415 available so Markov chain Monte Carlo (MCMC) methods are used to estimate it [Bozza et al., 2008] (24.

416 A similar approach to Bozza et al. [2008] for the evaluation of the likelihood ratio for the comparison  
 417 of sources problem is used by Alberink et al. [2013] in that variation in the variance parameter between  
 418 sources is modelled as well as variation in the mean parameter, although in Alberink et al. [2013] the data  
 419 are univariate. As with all of the other approaches discussed, the within-source distribution is Normal,  
 420 and the data are assumed independent. There are two main extensions seen in Alberink et al. [2013]. The  
 421 first is that three different distributions are used for the between-source distribution. One is the univariate  
 422 equivalent of the between-source distribution used in Bozza et al. [2008] (a semi-conjugate prior), one is a  
 423 non-informative prior, proportional to the inverse of the variance, and one is the conjugate prior distribution  
 424 seen on p. 74 of Gelman et al. [2004]. This conjugate prior distribution gives a between-source distribution  
 425 for the parameter  $(\mu, \sigma^2)$ , denoting group mean and variance, of

$$\begin{aligned}\mu &\sim N(\mu_0, \sigma^2/\kappa_0) \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)\end{aligned}$$

426 where  $\mu_0, \kappa_0, \nu_0$  and  $\sigma_0^2$  are hyperparameters to be estimated and the notation  $\text{Inv-}\chi^2$  corresponds to  
 427 a scaled inverse chi-squared distribution. The difference between this and the univariate equivalent of  
 428 the between-source distribution used in Bozza et al. [2008] is that the variance of the parameter  $\mu$  is  
 429 proportional to  $\sigma^2$ . An analytical form of the likelihood ratio for the two cases when the between-source  
 430 distribution is given by the non-informative prior and when the between-source distribution is given by the  
 431 conjugate prior [Alberink et al., 2013] who also show that no analytic solution exists if a semi-conjugate  
 432 prior is used (16, 17).

433 As in Bozza et al. [2008], Alberink et al. [2013] use MCMC methods to evaluate the likelihood ratio  
 434 when the between-source distribution is given by the semi-conjugate prior, although there are differences  
 435 in the implementation, leading to the second main extension. Alberink et al. [2013] use prior distributions  
 436 on the hyperparameters of the between-source distribution and then combine these prior distributions with  
 437 training data to obtain a posterior distribution for the hyperparameters, conditional on the training data.  
 438 All of the other methods discussed estimate the parameters of the between-source distribution directly  
 439 from the training data using summary statistics. The methods used in Alberink et al. [2013] allow for a  
 440 Bayesian approach for the estimation of the between-source distribution. One disadvantage of this approach



441 is that the method for estimating the likelihood ratio used in Bozza et al. [2008] is no longer feasible  
442 because, instead of having a known analytic form for the between-source density function, draws from  
443 the between-source distribution are obtained using MCMC methods. Monte Carlo integration is used by  
444 Alberink et al. [2013] to estimate the likelihood ratio.

445 All of the literature discussed in Sections 3.1.1 and 3.1.2 evaluates likelihood ratios for continuous  
446 evidential data. There are some common assumptions. All assume that measurements are independent  
447 and that the within-source distribution is Normal (univariate or multivariate). Constant variation between  
448 sources of the within-source distribution is assumed by Lindley [1977], Aitken and Lucy [2004], Aitken  
449 et al. [2007] and Aitken et al. [2006]. This assumption is relaxed by Bozza et al. [2008] and Alberink et al.  
450 [2013], allowing the variance to vary between sources. A Bayesian approach is used by Alberink et al.  
451 [2013] to obtain the parameters of the between-source distribution.

452 Methods for the evaluation of continuous, autocorrelated data are described in Wilson et al. [2014] and  
453 Wilson et al. [2015]. The data used for illustration are the quantities of drugs on banknotes where quantities  
454 on adjacent notes cannot be considered independent. Some work has also been done on the evaluation of  
455 evidence for discrete data, particularly in the field of DNA profiling [Buckleton et al., 2005] and more  
456 recently on data relating to clicks in speech [Aitken and Gold, 2013] and the presence or absence (binary  
457 data) of striation marks for screwdrivers [Aitken and Huang, 2017].

### 458 3.2 Discrimination

459 Forensic scientists are not only interested in comparisons of two pieces of evidence, such as control  
460 and recovered evidence, under different propositions, that of same source *versus* that of different source,  
461 without attention being paid to the identity of the source. There is also interest in the source of one piece of  
462 evidence. The support of the evidence for a proposition of source is of interest. The problem concerns the  
463 determination of whether a sample of data is more likely to be from one population (source) or another. Of  
464 course, such a determination is the concern of the fact-finder. The scientist is concerned with the probability  
465 of the measurements on the evidential material if the material came from one source or if it came from  
466 another. If there are more than two possible sources, then prior probabilities, that is, probabilities for each  
467 source under consideration before the material is examined, are needed in order to obtain a likelihood ratio.  
468 In this problem there is only one set of evidential data compared with the two sets (control and recovered)  
469 in the comparison problem. The aim is to assist the decision-maker as to the population of origin of the  
470 evidential data. This is a problem of *discrimination*, as distinct from a problem of *comparison*.

471 An example of the use of likelihood ratios in a problem of this sort can be seen in Zadora et al. [2010]  
472 which looks at the discrimination of glass samples and in Wilson et al. [2014, 2015] which considers  
473 discrimination between banknotes associated with a person associated with criminal activity and banknotes  
474 associated with a person not associated with criminal activity. As with the problem of comparison of  
475 sources, the likelihood ratio alone cannot determine whether a set of data is more likely from one population  
476 or another; it must be considered in conjunction with the prior odds. The derivation of the likelihood ratio  
477 for such discrimination problems is discussed in Taroni et al. [2010] (Chapter 8). The likelihood ratio for a  
478 set of evidence consisting of  $n$  measurements,  $\mathbf{z} = (z_1, \dots, z_n)$ , under two propositions,  $H_p$  and  $H_d$ , is  
479 considered.<sup>4</sup> The two propositions are given by

- 480 •  $H_p$  : data  $\mathbf{z}$  are from population 1, and
- 481 •  $H_d$  : data  $\mathbf{z}$  are from population 2.

---

<sup>4</sup> Note the change of use of notation. In this Section,  $\mathbf{z}$  refers to evidential data and not to training data.

482 The likelihood ratio  $V$  for the discrimination problem, where  $I$  is the background information as usual, is  
 483 given in Taroni et al. [2010] by

$$V = \frac{f(\mathbf{z} | H_p, I)}{f(\mathbf{z} | H_d, I)}. \quad (10)$$

484 This expression can be compared with (7) and the comparison problem. In the comparison context, the  
 485 joint density function of control and recovered data is considered. In the discrimination problem, two (or  
 486 more) possible sources (populations) are identified.

487 Assume as for the comparison problem that the data are hierarchical and that there are two possible  
 488 sources. The probability density function of groups of data from source  $i$  is parameterised by  $\theta_i$ ,  $i = 1, 2$   
 489 (possibly multivariate). If the value of  $\theta_i$  (for  $i \in \{1, 2\}$ ) varies between different groups in population  $i$   
 490 then by conditioning on  $\theta_1$  in the numerator and  $\theta_2$  in the denominator, the likelihood ratio  $V$  can be written

$$V = \frac{\int f(\mathbf{z} | \theta_1) f(\theta_1) d\theta_1}{\int f(\mathbf{z} | \theta_2) f(\theta_2) d\theta_2}. \quad (11)$$

491 The probability density function  $f(\theta_i)$  models the variability of the parameter  $\theta_i$  between groups in  
 492 population  $i$ , and is termed the between-group density function (the associated distribution function will be  
 493 termed the between-group distribution function). This is analogous to the between-source distribution used  
 494 to model variability between sources in the comparison of sources problem. Similarly, the density function  
 495  $f(\mathbf{z} | \theta_i)$  is termed the within-group density function (with the associated distribution function termed the  
 496 within-group distribution function).

497 Using this formulation for the likelihood ratio, the methods discussed previously for the evaluation of the  
 498 likelihood ratio for the comparison of sources problem can be adapted to evaluate the value of evidence for  
 499 discrimination problems. The limitations and assumptions of these methods still apply.

500 In the context of discrimination, training data are a random sample of groups from each or both of the  
 501 sources. Variation is between groups within each of the sources. There is an abuse of terminology here.  
 502 In the comparison problem with the proposition of common source, the control and recovered evidence  
 503 are deemed to be from the same source but without specification of the source. The source is a member  
 504 of a population of sources. In the discrimination problem, support for a particular source is assessed.  
 505 The distinction between comparison and discrimination problems is emphasised in Zadora et al. [2014]  
 506 where the two problems are discussed in different chapters (and note that discrimination is there noted as  
 507 classification).

### 508 3.3 Score-based models

Return now to consideration of the problem of comparison of sources with a  $p$ -dimensional control  
 measurement  $\mathbf{x} = (x_1, \dots, x_p)$  and a  $p$ -dimensional recovered measurement  $\mathbf{y} = (y_1, \dots, y_p)$ . For those  
 occasions when a feature-based model is not tractable (e.g., multidimensional binary data), the distance  
 $d(\mathbf{x}, \mathbf{y})$ , known as a *score* can be used instead. The value of the evidence is then

$$V = \frac{f(d(\mathbf{x}, \mathbf{y}) | H_p, I)}{f(d(\mathbf{x}, \mathbf{y}) | H_d, I)}.$$

509 Rarity is not considered. Inference may then continue as before but using the score, which is univariate, as  
 510 the statistic of interest. Score-based approaches estimate the probability distribution function of a calculated  
 511 score. Score-based approaches have been used for handwriting (Hepler et al. [2012]) and speech recognition  
 512 (Gonzalez-Rodriguez et al. [2006], Brümmer and Du Preez [2006], Morrison [2011]). Score-based methods  
 513 do not require the distributional assumptions (such as within-source Normality) needed to fit the models  
 514 described above but do still require a function to be chosen to model the probability distribution function  
 515 of the score.

516 There are various distance measures that may be used. Three examples are

- 517 • Euclidean:  $d = \sqrt{\sum_{i=1}^p (x_i - y_i)^2}$ ;
- 518 • Manhattan:  $d = \sum_{i=1}^p |x_i - y_i|$ ;
- Pearson correlation distance:  $100(1 - r)/2$  with

$$r = \frac{\sum_{i=1}^p (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^p (x_i - \bar{x})^2 \sum_{i=1}^p (y_i - \bar{y})^2}}.$$

519 Other examples are available in SAILR. For multiple control and recovered data  $\mathbf{x}_i, i = 1, \dots, m$  and  
 520  $\mathbf{y}_i, i = 1, \dots, n$ , respectively, pairwise score measurements or means can be used.

521 For the calculation of score-based likelihood ratios, distributions of scores of same-source comparisons  
 522 and of different-source comparisons are required. Determination of the same-source distribution can  
 523 be made by comparing every measurement in a training set  $\mathbf{z}$  with every other measurement within its  
 524 own source except with itself for which the distance is zero. For the different-source distribution, every  
 525 measurement is compared with all measurements from other sources. These results may then be used  
 526 to estimate the distributions of same-source and between-source comparisons. The distributions can be  
 527 represented initially by histograms. They may then be smoothed with a kernel density estimation or an  
 528 appropriate parametric distribution. The current choice of parametric distribution in SAILR is a Gamma  
 529 distribution or a Weibull distribution. The chosen distribution functions, one for same-source comparisons  
 530 and one for different-source comparisons, then can be used to determine the density calculation of the  
 531 evidence score for both distributions and hence calculate a likelihood ratio.

### 532 3.4 Comparison of feature-based and score-based models

533 Models for discrimination and for comparison that use the original data are feature-based models. The  
 534 models discussed in Sections 3.1 and 3.2 are all feature-based. Feature-based multivariate Normal models  
 535 compare the probability of observing the evidence given that the evidential samples (control and recovered)  
 536 measured, and compared, come from the same source or come from different sources. In contrast, the  
 537 score-based model compares the probability of observing the pairwise similarity between two samples  
 538 (control and recovered) given that they come from the same source with the probability of the pairwise  
 539 similarity given that the samples come from different sources. A comparison of the performances of  
 540 score-based and frequency-based likelihood ratios for forensic MDMA comparisons is given in Bolck et al.  
 541 [2015].

542 The benefits and shortcomings of both methods are given by Bolck et al. [2015] as:

- 543 • Feature-based benefits:
- 544 • Original data dimensionality preserved; no information loss.

- 545 • Rarity and similarity of the features relate directly to the magnitude of the likelihood ratio.
- 546 • Feature-based shortcomings:
  - 547 • Covariance estimation is difficult when limited data are available relative to the dimensionality of
  - 548 the variables.
  - 549 • The feature-based method is often less robust than the score-based model when there are limited
  - 550 population samples.
- 551 • Score-based benefits:
  - 552 • Covariance estimation between sources is possible with few samples available.
  - 553 • The method is robust and able to be generalised to new samples.
- 554 • Score-based shortcomings:
  - 555 • There is a loss of information because of a reduction of dimensionality.
  - 556 • The value of the likelihood ratio is based on the similarity of pairwise scores rather than the similarity
  - 557 and rarity of features.

### 558 3.5 Summary of feature-based models

559 References for details of a selection of feature-based two-level models with within-group measurements  
 560 independent and Normally distributed are listed here. Equation numbers are given for models for which  
 561 further details are given in the Appendix.

- 562 • Univariate:
  - 563 • Within-group Normal,
  - 564 Between-group Normal for between-group mean (assume within-group variance known)
  - 565 (Lindley [1977], (12), (13)).
  - 566 • Within-group Normal,
  - 567 Between-group Taylor expansion for between-group mean (assume within-group variance known)
  - 568 (Lindley [1977], (14)).
  - 569 • Within-group Normal,
  - 570 Between-group kernel for between-group mean (assume within-group variance known),
  - 571 (Aitken and Taroni [2004], (15)).
  - 572 • Within-group Normal,
  - 573 Between-group distribution:
    - 574 (a) Normal distribution - semi-conjugate prior,
    - 575 (b) Non-informative prior, proportional to the inverse of the variance,
    - 576 (c) Conjugate prior - Normal, scaled inverse chi-squared
    - 577 (Alberink et al. [2013], (16), (17))
- 578 • Bivariate:
  - 579 Numerator (predictive distribution) [Bernardo and Smith, 1994],
  - 580 Denominator (kernel),
  - 581 (Evetts et al. [1987], (18)).
- 582 • Multivariate, within-group measurements independent and Normally distributed
  - 583 • Within-group Normal,
  - 584 Between-group kernel for distribution of group means,

- 585 Within-group variance assumed common and estimated from training data. (Aitken and Lucy [2004],  
586 (4.1), (19)).
- 587 • Within-group Normal
  - 588 Between-group Normal for distribution of group means,  
589 Inverse Wishart for the covariance of within-source distribution, (Bozza et al. [2008], (24)).
  - 590 • With graphical models:  
591 See Section 3.1.1; Aitken et al. [2007].
  - 592 • In the presence of zeros, that is when no measurement of a specific characteristic has been made on  
593 certain members of the control data set, the recovered data set or the training data set: both Normal  
594 and kernel between-group distributions considered. Estimation of covariance matrices by imputation  
595 and by available cases (Zadora et al. [2010]).
  - 596 • In addition, when within-group measurements are autocorrelated and Normally distributed see  
597 Wilson et al. [2014, 2015].

#### 4 MODEL PERFORMANCE

598 Model performance for the comparison problem is assessed with a training set and associated data  $\mathbf{z}$  as  
599 discussed in Section 2.6. If possible, another set, known as a *validation* set could be used. The training set  
600 and validation set should both comprise several sources of data from a relevant population. Within each  
601 source, measurements are taken on each of several items. The source of each member of the two sets is  
602 known. Models and parameters can be fitted using the training set. The performance can be assessed using  
603 the validation set. Thus when a method for comparison or discrimination is tested using members of the  
604 data set it is known if the correct answer is given. In the absence of a validation set, the performance can be  
605 assessed through a second use of the training set (e.g., with a leaving-one-out method). Validation enables  
606 the provision of measures of performance based on calculated likelihood ratios.

607 For a comparison of two members of the validation (or training) set a likelihood ratio is calculated. There  
608 are two conclusions that may be drawn by the fact-finder: they are from the same source or they are not  
609 from the same source. If the likelihood ratio is greater than 1, then this is support for the proposition of a  
610 common source for the two members of the validation (training) set being compared. If they are truly from  
611 the same source then this is counted as a correct result. Similarly, if its value is less than 1, then this is  
612 support for the proposition of different sources for the two members of the validation (training) set being  
613 compared. If they are truly from different sources then this is counted as a correct result. However, if the  
614 two members have a value for the likelihood ratio of greater than 1 when they are from different sources,  
615 this is an incorrect result and the result is known as a *false positive*. Similarly, if the two members have a  
616 value for the likelihood ratio of less than 1 when they are from the same source, this is an incorrect result  
617 and the result is known as a *false negative*.

618 For discrimination with two groups, say  $A$  and  $B$ , the member of the data set may be classified by the  
619 fact-finder as belonging to group  $A$  or to group  $B$ . False positives and false negatives can be defined in  
620 a manner analogous to that of the comparison procedure. A likelihood ratio is calculated. If its value is  
621 greater than 1, then this is support for the proposition that the member of the training set belongs to group  
622  $A$ , say. If the member is truly from group  $A$  then this is counted as a correct result. Similarly, if its value  
623 is less than 1, then this is support for the proposition that the member is from group  $B$ . If it is truly from  
624 group  $B$ , then this is counted as a correct result. However, if the member has a value for the likelihood  
625 ratio of greater than 1 when it is from group  $B$ , this is an incorrect result and the result is a false positive,

626 say. Similarly, if the member has a value for the likelihood ratio of less than 1 when it is from group  $A$ , this  
627 is an incorrect result and the result is a false negative.

628 For both comparison and discrimination problems, the strength of the support is measured by the value  
629 of the likelihood ratio. As noted in Section 2.3 if the logarithm is taken this is known as the weight of  
630 evidence. Given the existence of a validation (training) set it is possible to measure the performance of a  
631 method for comparison or discrimination as the correct answer is known. It is not possible to assess the  
632 result in an individual case; the correct answer in an individual case is not known.

633 The likelihood ratio, or a function of it such as the logarithm, has been shown by [Good, 1989a,b]  
634 (Section 2.2) to provide the best (only) value of the evidence. Attempts to express the uncertainty associated  
635 with this assessment (e.g. with a confidence interval) are attempts to put a probability on a probability and  
636 should not be done [Taroni et al., 2016]. This view is not universally agreed, see discussion issues of *Law,*  
637 *Probability and Risk* (2016, volume 15, issue 1) and *Science and Justice* (2017, volume 56). Note also the  
638 quote from Kaye [1979] in Section 2: 'It thus supplies the jurors with as precise and accurate an illustration  
639 of the probative force of the quantitative data as the mathematical theory of probability can provide'. It is  
640 not necessary to provide an interval estimate.

641 There are several measures of performance.

642 • *The percentage of false positives and of false negatives amongst all the comparisons or discriminations*  
643 *tested.* Often, in a criminal case, one of the propositions is associated with the prosecution, hence the  
644 notation  $H_p$ , and other is associated with the defence, with the notation  $H_d$ . In such a circumstance,  
645 the burden of proof lies with the prosecution. It is a more serious error to support the prosecution  
646 proposition wrongly than to support the defence proposition wrongly. Let support for the prosecution  
647 proposition be known as a positive result. Thus, when considering the performance of a test, it is better  
648 to choose a test in which there is a low false positive rate and a high false negative rate rather than one  
649 in which there is a high false positive rate and low false negative rate. Ideally, zero false positive and  
650 zero false negative results are best but such an ideal is rarely achieved.

651 • *A Tippett plot.* See Evett and Buckleton [1996] and Tippett et al. [1968]. This is a graphical measure of  
652 rates of misleading evidence for comparisons. It is the complement of empirical cumulative distribution  
653 functions for same-source and different-source comparisons. The plots come in pairs, one for same-  
654 source comparisons and one for different-source comparisons. The  $\log(LR)$  is plotted on the  $x$ -axis  
655 and, for a particular value  $x_0$  of the  $\log(LR)$ , the  $y$ -axis is the relative frequency of the number of  
656 comparisons greater than  $x_0$ . For same-source comparisons, it is to be hoped that all  $\log(LR)$  values  
657 are greater than 0. Thus for  $x < 0$ , it is hoped the corresponding value on the  $y$ -axis will be 1 (or  
658 100%). Similarly, for different-source comparisons, it is to be hoped that all  $\log(LR)$  values are less  
659 than 0. Thus for  $x > 0$ , it is hoped the corresponding value on the  $y$ -axis will be 0 (or 0%).

660 The vertical distance from the intersection of the same-source plot with the line  $\log(LR) = 0$  and the  
661 line  $y = 1(100\%)$  is the rate of misleading evidence for same-source comparisons, the proportion of  
662 same-source comparisons that have a value of  $\log(LR) < 0$  ( $LR = 1$ ). The vertical distance from the  
663 intersection of the different-source plot with the line  $\log(LR) = 0$  and the line  $y = 0(0\%)$  is the rate of  
664 misleading evidence for different-source comparisons, the proportion of different-source comparisons  
665 that have a value of  $\log(LR) > 0$  ( $LR = 1$ ).

666 • *Detection error trade-off (DET) curves.* See Meuwly et al. [2017]. A detection error trade-off (DET)  
667 plot is a 2-dimensional graphical representation in which the proportion of false positives is plotted  
668 as a function of the proportion of false negatives. The closer the curves to the coordinate origin, the

669 better are the discriminating capabilities of the method. The intersection of a DET curve with the main  
 670 diagonal of the DET plot marks the Equal Error Rate (EER) which is the point when the proportions  
 671 of false positives and false negatives are equal.

672 • *Empirical cross-entropy*. See Meuwly et al. [2017], Ramos et al. [2013] and Ramos and Gonzalez-  
 673 Rodriguez [2013]

674 The performance of probabilistic assessments has been addressed by *strictly proper scoring rules*  
 675 (SPSR). Consider two propositions about a parameter  $\theta$ , one that  $\theta = \theta_p$  and one that  $\theta = \theta_d$ , with  
 676  $\Pr(\theta = \theta_p) = 1 - \Pr(\theta = \theta_d)$  For evidence evaluation, the *logarithmic* SPSR is used and defined as

$$\begin{aligned} C(\Pr(\theta_p | I), \theta) &= -\log_2(\Pr(\theta_p | I)) \text{ if } \theta = \theta_p, \\ &= -\log_2(1 - \Pr(\theta_d | I)) \text{ if } \theta = \theta_d, \end{aligned}$$

677 The measure of accuracy for evidence evaluation based on SPSR is a weighted average value of the  
 678 logarithmic scoring rule, and is known as the *empirical cross-entropy* (ECE):

$$\begin{aligned} ECE &= -\frac{\Pr(\theta_p | I)}{N_p} \sum_{\theta_{(i)}=\theta_p} \log_2 \Pr(\theta_p | E_i, I) \\ &\quad - \frac{\Pr(\theta_d | I)}{N_d} \sum_{\theta_{(j)}=\theta_d} \log_2 \Pr(\theta_d | E_j, I) \\ &= \frac{\Pr(\theta_p | I)}{N_p} \sum_{\theta_{(i)}=\theta_p} \log_2 \left( 1 + \frac{1}{LR_i \times O(\theta_p)} \right) \\ &\quad + \frac{\Pr(\theta_d | I)}{N_d} \sum_{\theta_{(j)}=\theta_d} \log_2 \left( 1 + LR_j \times O(\theta_p) \right), \end{aligned}$$

679 where  $LR_i(LR_j)$  is the likelihood ratio for the  $i$ -th ( $j$ -th)  $E_i$  ( $E_j$ ) piece of evidence where  $\theta = \theta_i(\theta_j)$ ,  
 680 respectively, and  $O(\theta_p)$  denotes the prior odds  $\Pr(H_p)/\Pr(H_d)$ . For the discrimination problem with  
 681 two sources, the parameters  $\theta_p$  and  $\theta_d$  represent the parameters of the two sources. For the comparison  
 682 problem  $\theta_p$  represents same-source comparisons and  $\theta_d$  represents different-source comparisons in the  
 683 validation dataset.

684 This measure tends to indicate better performance when the likelihood ratio leads to the correct  
 685 decision. The numerical value will be lower as the performance increases. The ECE can be represented  
 686 as an ECE-plot, showing its value for a certain range of priors.

#### 687 4.1 Conclusion

688 The development of methods for the evaluation of evidence for frequency-based continuous two-level  
 689 models is described from the hierarchical model for univariate continuous data developed by Lindley [1977]  
 690 to multivariate models with unknown means and covariances [Bozza et al., 2008]. This development is of  
 691 interest in its own right as a compilation of some thirty years of development. However, it also provides a

692 background to the development of the SAILR package, a package which extends these ideas to include  
693 score-based models.

694 Formulae for many of these are given in the Appendix and may also be found in many books on the  
695 subject (e.g. Aitken and Taroni [2004]; Zadora et al. [2014]).

696 There is much more that can be reviewed. References for some of the omissions of this paper are given  
697 here. It is hoped they are useful. There have been few papers on models for discrete data; see Aitken and  
698 Gold [2013] for an example. Score-based models have received a lot of attention recently and are included  
699 in SAILR; see Bolck et al. [2015] for examples. Graphical models provide an approach for a reduction in  
700 the dimensionality of multivariate problems; see Zadora et al. [2014] for examples.

## 701 Appendix: Inventory of frequency-based continuous two-level models

702 The purpose of the inventory is to illustrate the development of methods for the evaluation of evidence  
703 for frequency-based continuous two-level models. The formulae that are given are for the likelihood ratio,  
704 denoted  $V$  (9), with the purpose of illustrating how rarity and similarity are assessed within the same  
705 formula and how uncertainty in means and variances is considered.

706 There are no derivations of formulae. Source references where the derivations may be found are given  
707 in association with each model. Whilst SAILR provides a software package for evaluation of evidence,  
708 appropriate R code is also available elsewhere, e.g., Zadora et al. [2014].

### 709 General notation for univariate models (Lindley [1977])

710 Measurements are Normally distributed about the true values with a known, constant variance  $\sigma^2$ . For  
711  $m$  measurements  $(x_1, \dots, x_m)$  of a control item, the mean  $\bar{X}$  (the random variable corresponding to  
712 observation  $\bar{x}$  is Normally distributed with mean  $\theta_1$  and variance  $\sigma^2/m$ . For  $n$  measurements  $(y_1, \dots, y_m)$   
713 of a recovered item, the mean  $\bar{Y}$  is Normally distributed with mean  $\theta_2$  and variance  $\sigma^2/n$ . If the control and  
714 recovered items come from the same source, the prosecution proposition  $H_p$ , then  $\theta_1 = \theta_2$ . If the control  
715 and recovered items come from different sources, the defence proposition  $H_d$ , then  $\theta_1 \neq \theta_2$ .

Assume  $\theta \sim N(\mu, \tau^2)$  and let

$$a^2 = \frac{1}{m} + \frac{1}{n}, \quad \sigma_1^2 = \tau^2 + \sigma^2/m, \quad \sigma_2^2 = \tau^2 + \sigma^2/n, \quad \sigma_3^2 = \tau^2 + \sigma^2/(m+n),$$

and

$$W = (m\bar{X} + n\bar{Y})/(m+n), \quad Z = (\sigma_2^2\bar{X} + \sigma_1^2\bar{Y})/(\sigma_1^2 + \sigma_2^2).$$

716 Formulae are given below for realisations of these random variables: thus  $\bar{X}, \bar{Y}, W$  and  $Z$  are replaced  
717 by  $\bar{x}, \bar{y}, w$  and  $z$

### 718 Value of evidence for univariate models

719 • The distribution of the true values  $\theta$  is Normal, mean  $\mu$  and variance  $\tau^2$ , where  $\tau^2$  is assumed known.

$$V = \frac{\sigma_1\sigma_2}{a\sigma\sigma_3} \exp \left\{ -\frac{(\bar{x} - \bar{y})^2\tau^2}{a^2\sigma^2(\sigma_1^2 + \sigma_2^2)} \right\} \exp \left\{ -\frac{(w - \mu)^2}{2\sigma_3^2} + \frac{(z - \mu)^2(\sigma_1^2 + \sigma_2^2)}{2\sigma_1^2\sigma_2^2} \right\}. \quad (12)$$



- 720 • The between-group standard deviation  $\tau \gg \sigma$  such that  $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \tau^2$  and  $z = (\bar{x} + \bar{y})/2$  and  
 721  $m = n = 1$  without loss of generality then

$$V = \frac{\tau}{\sigma\sqrt{2}} \exp\left\{-\frac{(\bar{x} - \bar{y})^2}{4\sigma^2}\right\} \exp\left\{\frac{(z - \mu)^2}{2\tau^2}\right\}. \quad (13)$$

722 The term  $(\bar{x} - \bar{y})^2/4\sigma^2$  is a measure of similarity. The more similar (closer together)  $\bar{x}$  and  $\bar{y}$  are, the  
 723 smaller  $(\bar{x} - \bar{y})^2/4\sigma^2$  is and hence the larger the term  $\exp\left\{-\frac{(\bar{x} - \bar{y})^2}{4\sigma^2}\right\}$  is (note the negative  
 724 sign) and hence the larger  $V$  is. The term  $(z - \mu)^2/2\tau^2$  is a measure of rarity. The overall mean of the  
 725 population from which the measurements are assumed to have come is  $\mu$ . The mean  $z$  of the control  
 726 mean  $\bar{x}$  and recovered mean  $\bar{y}$ , weighted by their variances so that the mean with the smaller variance  
 727 is given the larger weight is compared with the overall mean. The further  $z$  is from  $\mu$ , the larger the  
 728 term  $(z - \mu)^2/2\tau^2$  is and hence the larger the term  $\exp(z - \mu)^2/2\tau^2$  is (note the implicit positive sign)  
 729 and hence the larger  $V$  is.

- 730 • The between-group distribution is not Normal but is represented with a general distribution  $p(\cdot)$ , with  
 731 second derivative  $p''(\cdot)$  then

$$V = \frac{1}{a\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\bar{x} - \bar{y})^2}{2a^2\sigma^2}\right\} \frac{p(w) + \frac{1}{2}p''(w)\sigma^2/(m+n)}{\{p(x) + \frac{1}{2}p''(x)\sigma^2/m\}\{p(\bar{y}) + \frac{1}{2}p''(y)\sigma^2/n\}}. \quad (14)$$

- The between-group distribution is represented by a kernel density estimate (Aitken and Taroni [2004],  
 p. 338). Consider background data of the form  $\{z_{ij}, i = 1, \dots, k; j = 1, \dots, l\}$  where  $k$  is the number  
 of groups and  $l$  is the number of members of each group, assumed constant amongst groups. Let  $\bar{z}_i$   
 denote the mean of the  $i$ -th group and  $\bar{z}$  the overall mean. The within-group variance is then estimated  
 by

$$\hat{\sigma}^2 = \sum_{i=1}^k \sum_{j=1}^l (z_{ij} - \bar{z}_i)^2 / (kl - k)$$

732 and the between-group variance  $\tau^2$  by

$$s^2 = \sum_{i=1}^k (\bar{z}_i - \bar{z})^2 / (k - 1) - \hat{\sigma}^2 / l.$$

$$V = \frac{K \exp\left\{-\frac{(\bar{x} - \bar{y})^2}{2a^2\sigma^2}\right\} \sum_{i=1}^k \exp\left\{-\frac{(m+n)(w - z_i)^2}{2[\sigma^2 + (m+n)s^2\lambda^2]}\right\}}{\sum_{i=1}^k \exp\left\{-\frac{m(\bar{x} - z_i)^2}{2(\sigma^2 + ms^2\lambda^2)}\right\} \sum_{i=1}^k \exp\left\{-\frac{n(\bar{y} - z_i)^2}{2(\sigma^2 + ns^2\lambda^2)}\right\}} \quad (15)$$

733 where

$$K = \frac{k\sqrt{(m+n)}\sqrt{(\sigma^2 + ms^2\lambda^2)}\sqrt{(\sigma^2 + ns^2\lambda^2)}}{a\sigma\sqrt{(mn)}\sqrt{\{\sigma^2 + (m+n)s^2\lambda^2\}}}.$$

- The distribution of the true values  $\theta$  is Normal, mean  $\mu$  and variance  $\tau^2$ , where  $\tau^2$  is not assumed  
 known (Alberink et al. [2013]). Conjugate priors are chosen for  $\theta$  and  $\sigma^2$ . The prior distribution for  
 $\theta$ , or more rigorously,  $\theta \mid \tau^2$  is  $N(\mu, \tau^2/\kappa_0)$ , for parameters  $\mu$  and  $\kappa_0$ . In this situation, a prior is  
 introduced for  $\tau$ , which is such that  $\nu_0\tau_0^2/\tau^2 \sim \chi^2(\nu_0)$  for parameters  $\nu_0$  and  $\tau_0$ . Formulaically, the

joint prior is

$$p_2(\theta, \sigma^2) = c_2^{-1}(\sigma^2)^{-(\nu_0+3)/2} \exp\left(-\frac{1}{2}\sigma^{-2}(\nu_0\tau_0^2 + \kappa_0(\tau - \tau_0)^2)\right),$$

with  $c_2$  a normalising constant. Let

$$\rho_0 = \nu_0\tau_0^2, \rho_k = \nu_0\tau_0^2 + n_k s_k^2 + \frac{\kappa_0 n_k}{k_0 + n_k}(\bar{x}_k - \mu_0)^2,$$

with  $k = 1, 2$ , and

$$\rho_{1,2} = \nu_0\tau_0^2 + \sum_{k=1}^2 n_k s_k^2 + \sum_{k=1}^2 \frac{\kappa_0 n_k}{k_0 + n_k}(\bar{x}_k - \mu_0)^2 + \frac{n_1 n_2}{k_0 + n}(\bar{x}_1 - \bar{x}_2)^2.$$

734 The likelihood ratio is then

$$LR = \frac{\Gamma(\nu_0/2)\Gamma((\nu_0 + n)/2)}{\Gamma((\nu_0 + n_1)/2)\Gamma((\nu_0 + n_2)/2)} \left(\frac{(\kappa_0 + n_1)(\kappa_0 + n_2)}{\kappa_0(\kappa_0 + n)}\right) \times \left(\frac{\rho_1}{\rho_{1,2}}\right)^{n_1/2} \left(\frac{\rho_2}{\rho_{1,2}}\right)^{n_2/2} \left(\frac{\rho_1 \rho_2}{\rho_0 \rho_{1,2}}\right)^{\nu_0/2}, \quad (16)$$

735 [Alberink et al., 2013].

736 • A semi-conjugate prior can be chosen for  $\theta$  and  $\sigma^2$  (Alberink et al. [2013]) such that  $\theta \sim N(\mu_0, \tau_0^2)$   
 737 and  $\sigma^2$  has an inverse chi-squared distribution with parameters  $(\nu_0, \sigma_0^2)$  such that  $\nu_0, \sigma_0, \mu_0$  and  $\tau_0$  and  
 738 the mean and variance are statistically independent. Then

$$p_3(\mu, \sigma^2) = c_3^{-1}(\sigma^2)^{-(\nu_0+2)/2} \exp\left(-\frac{1}{2}(\nu_0\sigma_0^2\sigma^{-2} + \tau_0^{-2}(\mu_0 - \mu)^2)\right) \quad (17)$$

739 with  $c_3$  the normalising constant, [Alberink et al., 2013]

740 Value of evidence for multivariate models

741 • An early approach to the estimation of the likelihood ratio for multivariate data was used in the case  
 742 of bivariate colour chromaticity co-ordinates for fibres (Evetts et al. [1987]). Let  $\mathbf{y}$  denote a bivariate  
 743 vector of complementary chromaticity co-ordinates measured from a fibre found at the crime scene and  
 744 assumed to come from an article of clothing worn by the criminal. Let  $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_m)$  denote a set  
 745 of bivariate vectors of complementary chromaticity co-ordinates measured from  $m$  fibres taken to be a  
 746 representative sample from a garment belonging to a suspect. The propositions are  $H_p$ : the recovered  
 747 fibre came from the suspect's garment, and  $H_d$ : the recovered fibre came from some other source.

748 The numerator of the likelihood ratio is taken to be  $f(\mathbf{y} | H_p, \mathbf{x})$  and the denominator to be  $f(\mathbf{y} | H_d)$ ;  
 749 see (2.4).

The measurements were assumed to have distributions  $f(\mathbf{y} | \mu, \Sigma)$  and  $f(\mathbf{x}_i | \mu, \Sigma)$ ,  $i = 1, \dots, m$  that were bivariate Normal with mean  $\mu$  and covariance matrix  $\Sigma$ . Vague priors are chosen for  $\mu$  and  $\Sigma$ :

$$f(\mu | \Sigma) \propto c \text{ for } \mu \text{ and } f(\Sigma) \propto |\Sigma|^{-3/2} \text{ for } \Sigma,$$

750 where  $c$  is a constant, independent of  $\mu$ . The probability density function for  $f(\mathbf{y} \mid \mathbf{x}, H_p)$  is then a  
751 bivariate Student density function of the form:

$$\frac{\Gamma(m/2)}{\pi\Gamma((m-2)/2)} \left/ \left\{ \left| \frac{(m-1)(m+1)}{m} S_x \right|^{1/2} \left[ 1 + (\mathbf{y} - \bar{\mathbf{x}})' \frac{(m-1)(m+1)}{m} S_x^{-1} (\mathbf{y} - \bar{\mathbf{x}}) \right]^{m/2} \right\} \right. \quad (18)$$

752 where  $\bar{\mathbf{x}}$  and  $S_x$  are the sample mean and covariance matrix, respectively, for the measurements  
753 (Aitchison and Dunsmore [1975], Aitchison et al. [1977]). The denominator  $f(\mathbf{y} \mid H_d)$  is taken as  
754 a kernel density estimate. Further work on likelihood ratios for fibre evidence of complementary  
755 chromaticity co-ordinates is described in Wakefield et al. [1991].

756 • Likelihood ratio with the assumption of constant within-source variation and between-source normality;  
757 see (Aitken and Lucy [2004]).

Let  $\Omega$  denote a population of  $p$  characteristics of items of a particular evidential type. Background data are measurements of these characteristics on a random sample of  $m$  members from  $\Omega$  with  $n(\geq 2)$  replicate measurements on each of the  $m$  members. The background data are denoted as  $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijp})^T, i = 1, \dots, m, j = 1, \dots, n$  with

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

The control and recovered measurements are denoted by  $\{\mathbf{y}_l\} = (\mathbf{y}_{lj}, j = 1, \dots, n_l, l = 1, 2)$  where  $\mathbf{y}_{lj} = (y_{lj1}, \dots, y_{ljp})^T$ , with

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

758 For within-source variation, the mean vector within source  $i$  is denoted by  $\theta_i$  and the within-source  
759 covariance matrix by  $U$  and  $(\mathbf{X}_{ij} \mid \theta_i, U) \sim N(\theta_i, U), i = 1, \dots, m, j = 1, \dots, n$ .

760 For between-source variation, the mean vector between sources  $i$  is denoted by  $\mu$  and the between-  
761 source covariance matrix by  $C$  and  $(\theta_i \mid \mu, C) \sim N(\mu, C), i = 1, \dots, m$ .

762 The means  $(\mathbf{Y}_l \mid \theta_l, D_l) \sim N(\theta_l, D_l)$  where  $D_l = n_l^{-1}U$  and for between-source normality,  
763  $(\mathbf{Y}_l \mid \mu, C, D_l) \sim N(\mu, C + D_l), l = 1, 2$ .

The value of the evidence is the ratio of

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

764 to

$$|2\pi C|^{-1/2} |2\pi\{n_1 U^{-1} + C^{-1}\}^{-1}|^{1/2} |2\pi\{n_2 U^{-1} + C^{-1}\}^{-1}|^{1/2} \times \exp\left\{-\frac{1}{2}(H_4 + H_5)\right\} \quad (19)$$

765 where

$$\begin{aligned}
 H_2 &= (\mathbf{y}^* - \mu)^T \left( \frac{U}{(n_1 + n_2)} + C \right)^{-1} (\mathbf{y}^* - \mu), \\
 H_3 &= (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)^T (D_1 + D_2)^{-1} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2), \\
 H_4 &= (\mu - \mu^*)^T \{ (D_1 + C)^{-1} + (D_2 + C)^{-1} \} (\mu - \mu^*), \\
 H_5 &= (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)^T (D_1 + D_2 + 2C)^{-1} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2), \\
 \mathbf{y}^* &= \frac{n_1 \bar{\mathbf{y}}_1 + n_2 \bar{\mathbf{y}}_2}{n_1 + n_2}, \\
 \mu^* &= \{ (D_1 + C)^{-1} + (D_2 + C)^{-1} \}^{-1} \{ (D_1 + C)^{-1} \bar{\mathbf{y}}_1 + (D_2 + C)^{-1} \bar{\mathbf{y}}_2 \}.
 \end{aligned}$$

766 The notation is chosen to match that in Aitken and Lucy [2004]<sup>5</sup>

767 The form of presentation is also chosen to be comparable with the univariate case described in Lindley  
 768 [1977]. This emphasises the factors for rarity and similarity. The terms  $H_2$  and  $H_4$  are measures of  
 769 rarity of means of the control and recovered measurements, first weighted by sample sizes and second  
 770 weighted by covariances. The terms  $H_3$  and  $H_5$  are measures of similarity of the control and recovered  
 771 measurements.

- 772 • Likelihood ratio with the assumption of constant within-source variation and kernel density estimation  
 773 of between-source variation ; the formula for the likelihood ratio is not given here, for reasons of space,  
 774 but is available in Aitken and Lucy [2004].
- 775 • Likelihood ratio when the assumption of the constant within-source variability is relaxed; see (Bozza  
 776 et al. [2008]).

777 Consider background data of  $p$ -variables, with  $m$  groups and  $n_i$  measurements  $\{z_{ij} =$   
 778  $(x_{ij1}, \dots, x_{ijp}, i = 1, \dots, m; j = 1, \dots, n_i)\}$  in each group. Denote the mean vector within-group  
 779  $i$  by  $\theta_i$  and the matrix of within-group variances and covariances by  $W_i$  and let  $\psi = (\theta, W)$  with  
 780  $\theta = (\theta_1, \theta_2)$  and  $W = (W_1, W_2)$ . Given  $\theta_i$  and  $W_i$ , the distribution of  $Z_{ij}$  is taken to be Normal  
 781 with  $Z_{ij} \sim N(\theta_i, W_i)$ . The distribution of the within-group mean  $\theta$  is taken to be Normal, such that  
 782  $\theta_i \sim N(\mu, B)$ ,  $i = 1, \dots, m$ . The distribution of the within-group matrix  $W$  is taken to be an inverted  
 783 Wishart distribution, such that  $W_i \sim IW(U, n_w)$ ,  $i = 1, \dots, m$  where the number of degrees of  
 784 freedom  $n_w$  is chosen to reduce the variability of the Wishart distribution.

785 A number  $n$  of measurements are available:  $n_1$  measurements  $\mathbf{y}_1 = (y_{11}, \dots, y_{1n_1})$  from a recovered  
 786 source and  $n_2$  measurements  $\mathbf{y}_2 = (y_{21}, \dots, y_{2n_2})$  from a control source;  $n_1 + n_2 = n$  and let  $\mathbf{y}$   
 787 denote  $(\mathbf{y}_1, \mathbf{y}_2)$ <sup>6</sup>.

Consider the proposition that the control and recovered measurements have the same source. Then  
 $\theta_1 = \theta_2$  and  $W_1 = W_2$ . The density function of the data is

$$f(\mathbf{y} \mid \psi, H_1) = \prod_{l=1}^2 \prod_{j=1}^{n_l} (2\pi)^{-p/2} |W|^{1/2} \exp \left\{ -\frac{1}{2} (\mathbf{y}_{lj} - \theta)' W^{-1} (\mathbf{y}_{lj} - \theta) \right\}.$$

788 The prior density of  $\psi$  is

<sup>5</sup> The notation of  $\mathbf{x}$  for training data and  $\mathbf{y}_1$  and  $\mathbf{y}_2$  for control and recovered data is used here for consistency with Aitken and Lucy [2004] in contrast to  $\mathbf{z}$ ,  $\mathbf{x}$  and  $\mathbf{y}$  in the rest of the paper. Also,  $H_1$  denotes  $\sum_{i=1}^2 \text{trace}(S_i U^{-1})$  where  $S_i = \sum_{j=1}^{n_i} (\mathbf{y}_{lj} - \bar{\mathbf{y}}_i)(\mathbf{y}_{lj} - \bar{\mathbf{y}}_i)^T$ , an expression used in intermediate calculations but not in the final result.

<sup>6</sup> For notational convenience, both control and recovered data are denoted with  $\mathbf{y}$ ; often  $\mathbf{x}$  denotes control data and  $\mathbf{y}$  denotes recovered data.

$$\pi(\psi | H_1) = (2\pi)^{-p/2} |B|^{-1/2} \exp \left\{ -\frac{1}{2}(\theta - \mu)'B^{-1}(\theta - \mu) \right\} \times \frac{c |U|^{(n_w-p-1)/2}}{|W|^{n_w/2}} \exp \left\{ -\frac{1}{2}tr(W^{-1}U) \right\}.$$

789 The complete conditional density of  $\theta$  is then

$$\pi(\theta | W, \mathbf{y}) \propto \exp \left[ -\frac{1}{2} \left\{ \sum_{l=1}^2 \sum_{j=1}^{n_l} (\mathbf{y}_{lj} - \theta)'W^{-1}(\mathbf{y}_{lj} - \theta) + (\theta - \mu)'B^{-1}(\theta - \mu) \right\} \right]. \quad (20)$$

790 The complete conditional density of  $W$  is

$$\pi(W | \theta, \mathbf{y}) \propto |W|^{-n/2} \exp \left\{ -\frac{1}{2} \sum_{l=1}^2 \sum_{j=1}^{n_l} n_l (\mathbf{y}_{lj} - \theta)'W^{-1}(\mathbf{y}_{lj} - \theta) \right\} \times |W|^{-n_w/2} \exp \left\{ -\frac{1}{2}tr(W^{-1}U) \right\}. \quad (21)$$

791 The function  $\pi(\psi | \mathbf{y}, H_k)$  is obtained from (20) and (21) with the use of Gibbs sampling. Consider the proposition  $H_2$  that the control and recovered measurements have different sources. The density function of the data  $\mathbf{y}$  is then

$$f(\mathbf{y} | \psi, H_2) = \prod_{l=1}^2 \left[ \prod_{j=1}^{n_l} (2\pi)^{-p/2} |W_l|^{-1/2} \exp \left\{ -\frac{1}{2}(\mathbf{y}_{lj} - \theta_l)'W_l^{-1}(\mathbf{y}_{lj} - \theta_l) \right\} \right].$$

792 The complete conditional densities of  $\theta$  and  $W$  are, for  $l = 1, 2$ ,

$$\pi(\theta_l | W_l, \mathbf{y}) \propto \exp \left\{ -\frac{1}{2}(\theta_l - \mu_l^*)'B_l^{*-1}(\theta_l - \mu_l^*) \right\}. \quad (22)$$

793 and

$$\pi(W_l | \theta_l, \mathbf{y}) \propto |W_l|^{-(n_l+n_w)/2} \exp \left\{ -\frac{1}{2} \exp \left( -\frac{1}{2}tr \left[ W_l^{-1} \{ n_l(\theta_l - \bar{\mathbf{y}}_l)(\theta_l - \bar{\mathbf{y}}_l)' + S_l U \} \right] \right) \right\}. \quad (23)$$

794 with

$$\begin{aligned}
 B_l^* &= (B^{-1} + n_l W_l^{-1})^{-1}, \\
 \mu_l^* &= B_l^* (B^{-1} \mu + n_l W_l^{-1} \bar{y}_l), \\
 S_l &= \sum_{j=1}^{n_l} (\bar{y}_{lj} - \bar{y}_l)' (\bar{y}_{lj} - \bar{y}_l),
 \end{aligned}$$

795 where  $\bar{y}_l = \sum_{j=1}^{n_l} y_{lj} / n_l$ .  
 796 The function  $\pi(\psi | \mathbf{y}, H_2)$  is obtained from (22) and (23) with the use of Gibbs sampling.  
 797 The marginal likelihood is then given from the equation

$$m(\mathbf{y} | H_k) = \frac{f(\mathbf{y} | \psi, H_k) \pi(\psi | H_k)}{\pi(\psi | \mathbf{y}, H_k)}. \quad (24)$$

798 Further details are available in Bozza et al. [2008]

## CONFLICT OF INTEREST STATEMENT

799 The author declares that the research was conducted in the absence of any commercial or financial  
 800 relationships that could be construed as a potential conflict of interest.

## FUNDING

801 This work was supported by the European Network of Forensic Science Institutes 2015 Monopoly  
 802 programme grant for Software for the Analysis and Implementation of Likelihood Ratios (SAiLR), the  
 803 Leverhulme Trust, grant number EM2016-027, and the Swiss National Science Foundation, grant number  
 804 BSSGI0\_155809.

## ACKNOWLEDGEMENTS

805 The author acknowledges very helpful contributions from Annabel Bolck and all other members of the  
 806 SAILR group including Leon Aronson, David Lucy, Jonas Malmborg, Petter Mostad, Tereza Neocleous,  
 807 Anders Nordgaard, Jane Palmberg, Amy Wilson and Grzegorz Zadora.

## THE SAILR PACKAGE

808 An early version of this document was written as an internal landscape document for the SAILR project.  
 809 Further information about the project is available from Dr. Jeannette Leegwater at the Netherlands  
 810 Forensic Institute (jleegwater@nfi.minvenj.nl). Details of the software are available on-line from  
 811 <https://downloads.holmes.nl/sailr/sailr>. Operation of SAILR requires at least Java 8 to be installed. Java 8  
 812 can be downloaded from [http://www.oracle.com/technetwork/pt/java/javase/downloads/jre8-downloads-](http://www.oracle.com/technetwork/pt/java/javase/downloads/jre8-downloads-2133155.html)  
 813 2133155.html.

## REFERENCES

814 J. Aitchison and I. Dunsmore. *Statistical Prediction Analysis*. Cambridge University Press, 1975.

- 815 J. Aitchison, J.D.F. Habbema, and J.W. Kay. A critical comparison of two methods of statistical  
816 discrimination. *Applied Statistics, Journal of the Royal Statistical Society, Series C*, 26:15–25, 1977.
- 817 C. G. G. Aitken and E. Gold. Evidence evaluation for discrete data. *Forensic Science International*, 230:  
818 147 – 155, 2013. doi: <http://dx.doi.org/10.1016/j.forsciint.2013.02.042>.
- 819 C. G. G. Aitken and D. Lucy. Evaluation of trace evidence in the form of multivariate data. *Journal of the*  
820 *Royal Statistical Society. Series C (Applied Statistics)*, 53:109–122, 2004.
- 821 C. G. G. Aitken, D Lucy, G Zadora, and J. M. Curran. Evaluation of transfer evidence for three-level  
822 multivariate data with the use of graphical models. *Computational Statistics & Data Analysis*, 50(10):  
823 2571–2588, 2006.
- 824 C.G.G. Aitken and C. Huang. Evidence evaluation for hierarchical, longitudinal, binary data using a  
825 distance measure. *Statistica Applicata - Italian Journal of Applied Statistics*, 27:213–223, 2017.
- 826 C.G.G. Aitken and A. Nordgaard. Letter to the editor – the roles of participants’ differing  
827 background information in the evaluation of evidence. *Journal of Forensic Sciences*, 2017.  
828 <https://doi.org/10.1111/1556-4029.13712>.
- 829 C.G.G. Aitken and D.A Stoney. *The Use of Statistics in Forensic Science*. Ellis Horwood Limited,  
830 Chichester, 1991.
- 831 C.G.G. Aitken and F. Taroni. *Statistics and the Evaluation of Evidence for Forensic Scientists*. Wiley,  
832 Chichester, 2 edition, 2004.
- 833 C.G.G. Aitken, G. Zadora, and D. Lucy. A two-level model for evidence evaluation. *Journal of Forensic*  
834 *Sciences*, 52:412–419, 2007. doi: 10.1111/j.1556-4029.2006.00358.x.
- 835 C.G.G. Aitken, Q. Shen, R. Jensen, and B. Hayes. The evaluation of evidence for exponentially distributed  
836 data. *Computational Statistics & Data Analysis*, 51(12):5682–5693, 2007a.
- 837 I. Alberink, A. Bolck, and S. Menges. Posterior likelihood ratios for evaluation of forensic trace evidence  
838 given a two-level model on the data. *Journal of Applied Statistics*, 40:2579–2600, 2013. doi: 10.1080/  
839 02664763.2013.822056.
- 840 J.M. Bernardo and A.F.M. Smith. *Bayesian theory*. John Wiley and Sons, Chichester, 1994.
- 841 A. Biedermann, T. Hicks, D. Taroni, C. Champod, and C.G.G. Aitken. On the use of the likelihood ratio  
842 for forensic evaluation: response to Fenton et al. [2014a]. *Science and Justice*, 54:316–318, 2014.
- 843 A. Bolck, H. Ni, and M. Lopatka. Evaluating score- and feature-based likelihood ratio models for  
844 multivariate continuous data: applied to forensic MDMA comparison. *Law, Probability and Risk*, 14:  
845 243–266, 2015.
- 846 S. Bozza, F. Taroni, R. Marquis, and M. Schmittbühl. Probabilistic evaluation of handwriting evidence:  
847 likelihood ratio for authorship. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57:  
848 329–341, 2008. doi: 10.1111/j.1467-9876.2007.00616.x.
- 849 N. Brümmer and J. Du Preez. Application-independent evaluation of speaker detection. *Computer Speech*  
850 *and Language*, 20:230–275, 2006.
- 851 J.S. Buckleton, C.M. Triggs, and S.J. Walsh. *Forensic DNA Evidence Interpretation*. CRC Press, Boca  
852 Raton, 2005.
- 853 C. Champod, F. Taroni, and P. Margot. The Dreyfus case - an early debate on experts’ conclusions (an  
854 early and controversial case on questioned document examination). *International Journal of Forensic*  
855 *Document Examiners*, 5:446–459, 1999.
- 856 R. Cook, G. Evett, I.W. and Jackson, P. J. Jones, and J. A. Lambert. A hierarchy of propositions: deciding  
857 which level to address in casework. *Science & Justice*, 38(4):231–239, 1998a.
- 858 R. Cook, I.W. Evett, G. Jackson, P. J. Jones, and J. A. Lambert. A model for case assessment and  
859 interpretation. *Science & Justice*, 38(3):151–156, 1998b.

- 860 J.G. Darboux, P.E. Appell, and J.H. Poincaré. Examen critique des divers systèmes ou études graphiques  
861 auxquels a donné lieu le bordereau. In *L'affaire DREFUS - la révision du procès de Rennes - enquête de*  
862 *la chambre criminelle de la Cour de Cassation*, pages 499–600. Ligue française des droits de l'homme  
863 et du citoyen., Paris, 1908.
- 864 ENFSI. *Guideline for evaluative reporting in forensic science*, 2015. URL [http://enfsi.eu/](http://enfsi.eu/documents/forensic-guidelines/)  
865 [documents/forensic-guidelines/](http://enfsi.eu/documents/forensic-guidelines/).
- 866 I.W. Evett and J.S. Buckleton. Statistical analysis of str data. In A. Carracedo, B. Brinkmann, and W. Bär,  
867 editors, *Advances in Forensic Haemogenetics 6*. Springer Verlag, 1996.
- 868 I.W. Evett, P.E. Cage, and C.G.G. Aitken. Evaluation of the likelihood ratio for fibre transfer evidence in  
869 criminal cases. *Applied Statistics*, 36:174–180, 1987.
- 870 I.W. Evett, G. Jackson, and J. A. Lambert. More on the hierarchy of propositions: exploring the distinction  
871 between explanations and propositions. *Science & Justice*, 40(1):3–10, 2000.
- 872 N. Fenton, D. Berger, D. Lagnado, M. Neil, and A. Hsu. When 'neutral' evidence still has probative value  
873 (with implications from the Barry George case). *Science and Justice*, 54:274 – 287, 2014a.
- 874 N. Fenton, D. Lagnado, D. Berger, M. Neil, and A. Hsu. Response to 'On the use of the likelihood ratio for  
875 forensic evaluation: response to Fenton et al.'. *Science and Justice*, 54:319 – 320, 2014b.
- 876 A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, London,  
877 2 edition, 2004.
- 878 J. Gonzalez-Rodriguez, A. Drygajlo, D. Ramos-Castro, M. Garcia-Gomar, and J. Ortega-Garcia. Robust  
879 estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer*  
880 *Speech & Language*, 20(2-3):331–355, 2006.
- 881 I. J. Good. C312: Yet another argument for the explication of weight of evidence. *Journal of Statistical*  
882 *Computation and Simulation*, 31:58–59, 1989a.
- 883 I. J. Good. C319: Weight of evidence and a compelling metaprinciple. *Journal of Statistical Computation*  
884 *and Simulation*, 31:121–123, 1989b.
- 885 I. J. Good. Weight of evidence and the Bayesian likelihood ratio. In C.G.G. Aitken and D.A. Stoney,  
886 editors, *The Use of Statistics in Forensic Science*, pages 85–106. Ellis Horwood, Chichester, 1991.
- 887 I.J. Good. Studies in the history of probability and statistics. XXXVIII A. M. Turing's statistical work in  
888 World War II. *Biometrika*, 66:393–396, 1979.
- 889 W. Harvey, O. Butler, J. Furness, and R. Laird. The Biggar murder: dental, medical, police and legal  
890 aspects. *Journal of the Forensic Science Society*, 8:1568–219, 1968.
- 891 A. B. Hepler, C. P. Saunders, L. J. Davis, and J. Buscaglia. Score-based likelihood ratios for handwriting  
892 evidence. *Forensic Science International*, 219:129–140, 2012.
- 893 D.H. Kaye. The laws of probability and the law of the land. *The University of Chicago Law Review*, 47:  
894 34–56, 1979.
- 895 D. V. Lindley. A problem in forensic science. *Biometrika*, 64(2):207–213, 1977. doi: 10.1093/biomet/64.2.  
896 207. URL <http://biomet.oxfordjournals.org/content/64/2/207.abstract>.
- 897 S. P. Lund and H. Iyer. Likelihood ratio as weight of forensic evidence: a closer look. *Journal of Research*  
898 *of National Institute of Standards and Technology*, 122:27, 2017. <https://doi.org/10.6028/jres.122.027>.
- 899 A. Martyna, D. Lucy, G. Zadora, B.M. Trzcinska, D. Ramos, and A. Parczewski. The evidential value of  
900 microspectrophotometry measurements made for pen inks. *Analytical Methods*, 5:6788–6795, 2013.  
901 doi: 10.1039/c3ay41622d.
- 902 D. Meuwly, D. Ramos, and R. Haraksim. A guideline for the validation of likelihood ratio methods  
903 used for forensic evidence evaluation. *Forensic Science International*, 276:142–153, 2017. doi:  
904 10.1016/j.forsciint.2016.03.048.



- 905 G.S. Morrison. A comparison of procedures for the calculation of forensic likelihood ratios from acoustic  
906 phonetic data multivariate kernel density (mkvd) versus Gaussian mixture model-universal background  
907 model (gmm-ubm). *Speech Communication*, 53:91–98, 2011.
- 908 C.S. Peirce. The probability of induction. In J.R. Newman, editor, *The World of Mathematics, 1956*,  
909 volume 2, New York, 1878. Simon Schuster.
- 910 D. Ramos and J. Gonzalez-Rodriguez. Reliable support: measuring calibration of likelihood ratios. *Forensic  
911 Science International*, 230:156–169, 2013.
- 912 D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, and C.G.G. Aitken. Information-theoretical assessment of  
913 the performance of likelihood ratio computation methods. *Journal of Forensic Sciences*, 58:1503–1518,  
914 2013.
- 915 F. Taroni, S. Bozza, A. Biedermann, P. Garbolino, and C. G. G Aitken. *Data Analysis in Forensic Science:  
916 a Bayesian Decision Perspective*. Wiley, Chichester, 2010.
- 917 F. Taroni, S. Bozza, A. Biedermann, and C.G.G. Aitken. Dismissal of the illusion of uncertainty in the  
918 assessment of a likelihood ratio. *Law, Probability and Risk*, 15:1–16, 2016.
- 919 C.F. Tippett, V.J. Emerson, M.J. Fereday, F. Lawton, and S.M. Lampert. The evidential value of the  
920 comparison of paint flakes from sources other than vehicles. *Journal of the Forensic Science Society*, 8:  
921 61–65, 1968.
- 922 J.C. Wakefield, A.M. Skene, A.F.M. Smith, and I.W. Evett. The evaluation of fibre transfer evidence in  
923 forensic science: a case study in statistical modelling. *Applied Statistics*, 40:461–476, 1991.
- 924 A. Wilson, C.G.G. Aitken, R. Sleeman, and R. Carter. The evaluation of evidence relating to traces of  
925 cocaine on banknotes. *Forensic Science International*, 236:67–76, 2014.
- 926 A. Wilson, C.G.G. Aitken, R. Sleeman, and R. Carter. The evaluation of evidence for autocorrelated data  
927 in relation to traces of cocaine on banknotes. *Applied Statistics*, 64:275–298, 2015.
- 928 G. Zadora, T. Neocleous, and C. G. G. Aitken. A two-level model for evidence evaluation in the presence  
929 of zeros. *Journal of Forensic Sciences*, 55(2):371–384, 2010. doi: 10.1111/j.1556-4029.2009.01316.x.
- 930 G. Zadora, A. Martyna, D. Ramos, and C.G.G. Aitken. *Statistical analysis in forensic science: evidential  
931 value of multivariate physicochemical data*. John Wiley and Sons Ltd., Chichester, 2014.