ENFSI Guideline:
Route Towards Likelihood Ratio in Forensic Chemistry

ENFSI-DWG-LRC
Issue 01 – August 2026

ENFSI

Co-funded by the Internal Security Fund of the European Union

ENFSI Secretariat
Bundeskriminalamt KT-AS
65173 Wiesbaden, Germany
Phone +49 611 55 16660
E-mail secretariat@enfsi.eu
Website www.enfsi.eu

European Union’s Internal Security Fund — Police

This Guideline “Route Towards Likelihood Ratio (LR) in Forensic Chemistry” was funded by the European Union’s Internal Security Fund - Police.
The content of this Guideline represents the views of the authors only and is (his/her) sole responsibility. The European Commission does not accept any responsibility for use that may be made of the information it contains.

Acknowledgements

We acknowledge the following colleagues for substantial support to develop this guideline.

Mervi Eriksson (NBI Finland) for her general meeting and project coordination.

Elisa Kohtamaki (NBI Drug Unit) for analyzing the cannabis samples used in Example 1 of this guideline.

Martin Balmer, Rainer Albrecht and Lucca Landauer (Zurich Canton Police Switzerland,) for gathering the young cannabis plant material used in Example 1 of this guideline.

Dr. Lucas Helmecke (NRW State Bureau of Criminal Investigation, Germany) for providing plant material from CBD-rich hemp used in Example 1 of this guideline.

Dr. Martina Unterländer, Federal Criminal Police Office (BKA Germany), Section KT46-Biological Materials and Micro Traces & NBI drugs lab for the DNA-analysis of cannabis plant material used in Example 1 of this guideline.

Dr. Lars Dohmen, BKA, KT42- Inorganic Materials and Micro Traces, Coatings for providing the infrared spectra of clear coats.

EU commission for funding through the European Union’s fund 101160526 — FOR-FUTURE — ISF-2023-TF2-AG-ENFSI-IBA-2

Special thanks to the ENFSI Drugs Working Group for supporting the Subcommittee Chemometrics.

Official language

The text may be translated into other languages as required. The English language version remains the definitive version.

Copyright

Further information

For further information about this publication, contact the ENFSI Secretariat. Please check the website of ENFSI (www.enfsi.eu) for update information.

Table of Contents

1. AIMS

Forensic science is an important part of the crime investigation process that provides information in the form of forensic reports or statements to the judicial system. The role of forensic investigations has increased, and more in-depth interpretations are required to increase the impact of forensic results.

The forensic literature shows an increasing trend in the use of chemometrics. This can be seen in many forensic disciplines like drug profiling, oil spill comparison, glass-, fibre-, arson debris analysis and cartridge examinations. Enhanced, structured and reliable data processing as well as systematic interpretation of the chemometric output are increasing requirements in forensics (UNODC, 2005). These are especially needed in comparative examinations, where classification, grouping or similarity (profiling) are of interest.

Usually, the collected data requires extensive data pre-processing and analysis. Often further substantiation of the result (an opinion on the evidential value) is needed as well. The latter can be done by means of Likelihood Ratio calculation, which expresses numerical support to the forensic scientist opinion in reporting evidential value to the end users (law enforcement and court of justice) in the light of two competing hypotheses. Selection of the correct process requires chemical, chemometric and forensic knowledge. The steps are dependent on the type of the question presented as well as on the data at hand. The creation of a process including LR calculation requires representative datasets. Also, the requirements set by the quality standards must be acknowledged in the process.

All this may look like a maze to forensic scientists working with routine casework. The need to describe the situations / suitable applications / data requirements *et cetera* when chemometrics and evaluative reporting (like LR) can be used and give additional information has been recognized. However, a gap can be seen between routine forensic casework and casework for which the use of statistical software or programming (R, Python) is needed. For many forensic chemists there is a need to fill this gap.

There is a demand to demonstrate, explain and train forensic chemists how to use statistics and probabilistic reasoning in forensic casework. This can be achieved by developing the skills and knowledge in the fields of chemometrics as well as Likelihood Ratio calculations. This will help the forensic chemist to better answer the forensic questions in the required level.

This guideline aims to provide guidance and examples in the field of forensic chemistry in selection of the best route from a presented forensic question via data processing to interpretation reported to the end users (e.g. law enforcement).

Good quality datasets with known ground truth are an essential requirement when setting up a chemometric and/or LR calculation method. Some existing datasets are collected to be shared across forensic institutes to enhance method development and method validation.

This work is part of EU project: FOR FUTURE - Work package 07 (ROTOR): The route towards likelihood ratio (LR): A journey from the forensic question to its answer; 101160526 — FOR-FUTURE — ISF-2023-TF2-AG-ENFSI-IBA-2.

This is a continuation of previous projects that have addressed these topics, e.g. chemometrics in MP2016 STEFA G02 (ISF-Police: 779485-STEFA-ISFP-2016-AG-IBA-ENFSI) and likelihood ratios in the MP2013 TVEFS-2020 T6 (Home/2013/ISEC/MO/ENFSI/4000005962).

2. SCOPE

This guideline is on the quantitative evaluation of evidence in forensic chemistry. The focus is on questions about probabilistic classification and comparison, including output of chemometric analysis, and using the likelihood ratio framework.

Out of scope are questions on identification, quantification, categorical classification or on activity level.

DEFINITIONS AND TERMS

Term	Short description
Bayesian inference	*Method of statistical inference in which Bayes' theorem is used to calculate a probability of a hypothesis, given prior beliefs which are updated based on observations.*
Continuous-valued findings	*Observations or measurements that can take on any value within a given range. These values are not restricted to a set of distinct, separate numbers but instead can vary smoothly and include fractional or decimal points. Examples are time, weight, and temperature.*
Control / reference material	*Material from a known source*
ENFSI	*European Network of Forensic Science Institutes*
Factfinder	*Person or group responsible for determining the facts in a legal case based on the evidence presented, for example the court.*
Ground truth	*Reality as to a hypothesis being true or false.*
Kernel Density Estimation (KDE)	*A non-parametric method that smooths data to create a continuous estimate of the underlying probability density function, making it well-suited for capturing e.g. multi-modal distributions.*
Likelihood ratio	*Ratio of probabilities or probability densities of observations under two competing hypotheses.*
Likelihood ratio system	*Automated procedure that has analytical observations as input and an LR as output*
Non-parametric model	*A statistical approach that doesn't assume the data follow a specific distribution, like the normal shape. It does not rely on any parameters but directly on the data itself, making it useful when the underlying distribution is unknown.*
Parametric model	*A statistical approach that assumes the data follow a known distribution such as normal, defined by a fixed set of parameters (the parameters needed are mean and standard deviation for normal distribution).*
Posterior probability	*Probability (or probability density) of a hypothesis after taking account of measurements*
Prior probability	*Probability (or probability density) of a hypothesis before taking account of measurements*
Probability Density Function (PDF)	*A function that describes the relative likelihood of a continuous random variable taking on a specific value, with the total are under the curve equal to 1.*
Recovered material	*Questioned material from an unknown source.*

3. INTRODUCTION

A core element of forensic sciences is answering questions regarding seized materials (investigative samples) on behalf of investigative authorities and the judiciary. Such questions, which typically have a technical or scientific basis, are generally handled by forensic service providers. The principles of forensic sciences do not differ from those of any other scientific discipline. However, there is a particular emphasis on ensuring that the methodology and the resulting conclusions can be understood and traced by an interested layperson, such as a member of the judiciary. The range of questions can vary greatly, from simple and straightforward to very complex. Correspondingly, the answers can range from direct, clear, and unambiguous to merely indicative or evaluative. In what follows the focus will be on questions asked in the field of forensic chemistry.

It is the responsibility of investigative authorities and the judiciary to formulate and pose relevant questions concerning the seized material accurately. The forensic service provider must assess the quality of the investigative material to determine whether it is suitable for forensic analysis. The provider must then evaluate whether the posed question can be answered with the available material. Furthermore, they must examine their analytical techniques and analytical schemes to ensure they are capable of characterizing the sample's properties to answer the questions posed.

The types of questions related to material samples can essentially be divided into two categories, classification/identification and comparison:

If the question is general, e.g. "What does this material consist of?", the type of answer is an identification of substance(s) in a mixture or the discrete allocation to a class.

In such questions there is no pre-opinion about the material expressed by the commissioner. The task for the forensic service provider is then to report the findings from the analysis of the material, without any conclusions related to incriminating hypotheses.

If the question is of a different kind as such that the ultimate answer is expected to be a "yes" or "no", then the forensic service provider is tasked to report such an answer when possible, or to report a conclusion on the "yes"-side or on the "no"-side (or neutral) and with which strength the forensic results point in the respective direction. In this case, the commissioner has reasons to believe in "yes" or "no", and the forensic report is expected to either confirm/disprove or strengthen/weaken that prior opinion. Examples of such questions are "Does the material contain classified narcotic substances?", "Do the two seizures of cocaine have a common origin?", "Does the foreign paint on the hit car originate from the car suspected to have made the hit?"

In this booklet we do not focus on the first, general type of question (see section 3.1), but on the second category — questions with the possible answers "yes" or "no". These can, in turn, be divided into those that still can be answered unambiguously (see examples in section 3.2.1) and those that cannot (section 3.2.2 and 3.3).

Therefore, it is crucial for the type of question being answered to determine whether it pertains to identification/classification or individualisation. The latter is so-called source-attribution and is related to comparison (see section 3.3.).

This guideline does not address the challenges of questions related to activities.

3.1. Identification of items

Investigative samples are often complex mixtures in a matrix. The goal is to separate, isolate, and identify the substances of interest from the mixture, essentially to name them. Examples include identifying cocaine in a powder mixture or determining the presence of ethanol in a blood sample.

In **substance identification**, as explained above, the characteristics of the molecule to be analysed are known. Capturing characteristics is objective and sufficient to reach an unequivocal identification, provided the method of analysis is powerful enough and the sample material is of sufficient quality. An identification of a chemical substance can be done unambiguously by following the best practice manuals e.g. of ENFSI-DWG (ENFSI BPM DWG-CDA-001-001, 2020) or SWGDRUG (SWGDRUG Recommendations for the Analysis of Seized Drugs, 2024).

After identification, if necessary, the concentration of a substance can be quantified through a content determination. Quantitative analysis always includes a statistical error that can be determined during the method validation.

Material identification can also involve brand identification, such as determining the manufacturer and model of a firearm (e.g. SIG P226), the calibre of a bullet (e.g. 9 mm Parabellum), or the species of a plant (e.g. *Cannabis sativa L.*). In material identification, the task generally involves classifying or identifying the sample. Nature is replete with classes and species that can be distinctly recognized and identified simply by sorting them against their unique characteristics. Occasionally, the identification of classes or species is also of forensic interest in solving a criminal case.

3.2. Classification

3.2.1. Categorical classification

The term classification is well-known within analytical chemistry and refers to the attribution of a material to a group of substances, like opioids, benzodiazepines etc. Like identification, this attribution is considered unambiguous, since it is solely based on known observed characteristics from chemical analyses using standard operating procedures. This unambiguous classification is referred to as categorical.

Deviations from standard operating procedures, like insufficient amount of material, bad quality etc. would imply that no conclusion about which class the material belongs to is given, or that a probabilistic classification (see below) is made.

3.2.2. Probabilistic classification

Forensic classification is wider than classification to a group of substances on molecular level (subsection 3.2.1). As an example, standard float glass can be processed further to produce toughened glass, frosted glass, laminated safety glass and soundproof glass. A questioned material (a fragment) can be unambiguously identified as glass, but the type of glass requires more investigations that may not provide unambiguous conclusions. For categorical classification the conclusion is completely determined by the chemical results, for probabilistic classification possible sources of uncertainty must be considered.

3.3. Comparison of items

In the forensic inquiry on comparison of items, the focus is not on the definitive identification of the material itself but rather on determining whether the materials are related — whether both samples originate from the same source or not. Examples for comparison include attributing a fingerprint to a finger, determining whether two drug samples come from the same batch, assigning a bullet to a firearm, comparing a glass fragment to a windowpane, or allotting a fibre found on a person's clothing to the driver’s seat of a stolen vehicle. Such examples illustrate that the focus is on the source, making this category of questions correspond to individualisation. These investigations may be particularly challenging due to the presence of very similar characteristics among different individuals and the fact that the entirety of all individuals is typically neither known nor analytically accessible, making it difficult to prove individualisation beyond doubt.

3.4. Uncertainties

In forensic science, a precise understanding and clear distinction of terms such as *uncertainty, measurement uncertainty, variance, limitation, error* and *likelihood ratio* are absolutely essential to ensure accurate interpretations and avoid confusion.

Too often, the natural sciences and their applications, such as forensic chemistry, are mistakenly perceived as inherently imperfect or incapable of providing definitive results, compared with the certainty of mathematics. This perspective needs to be corrected and nuanced. On the one hand, nature offers many systems, classes and species that can be characterised, stabilised and identified objectively and unambiguously, independently of any subjective expert judgement. On the other hand, if uncertainty is inherent in any measurement of the real world, it is precisely the role of statistics – a fundamental branch of mathematics – to provide the rigorous frameworks and tools to quantify, model and manage it transparently. Therefore, the forensic chemist must be constantly aware of the meaning and implications of these terms, not to deny the existence of uncertainty, but to control it to arrive at objective and reliable conclusions regarding the identification or comparison of materials.

3.4.1. Variability

A measuring instrument captures certain physical or chemical characteristics of a sample. It is inherent to any measurement that, when repeated on the same sample material, slight variations will occur, with each measurement result differing slightly but generally being close together. The degree of dispersion of these individual measurements around their mean is called the variance of the measurements (or the standard deviation, its square root). A low variance indicates a good proximity between repeated measurements.

The term 'measurement error', sometimes used, can be misleading. It does not refer to an incorrect result but to variance and the term 'measurement uncertainty' is preferable to communicate this, see section 3.4.2. Despite minor variances in individual characteristics of a sample, the overall characteristics (e.g., an IR spectrum, mass spectrum, or NMR spectrum) can be unequivocally recognized, allowing a substance in a mixture to be distinctly identified (classification).

In classification, the variability of individual measurements, quantified and managed by statistical methods, needs to be considered because it determines the reliability of the final classification. When the variability is considered to be negligible with respect to the separation between classes i.e. it does not interfere with the characterisation or distinction of the properties under study, then the classification is referred to as unambiguous, also called a categorical classification.

If this is not the case, i.e. the variability is large enough to affect the distinction between different sets of characteristics, the classification is probabilistic. This makes it difficult to classify unambiguously and can lead to misclassification, a genuine error.

A definitive classification requires a combination of known, stable characteristics (chemical and/or physical properties). Additionally, the analytical method or techniques must be specific and selective enough to represent the characteristics adequately. The instrumental contribution to the overall variability can be considered as 'internal'.

There are also ‘external’ contributions to the variability. An important one is heterogeneity of the sample. The heterogeneity is relevant for quantitative analysis for a multi-substance sample (usually called a mixture) when the amount of target analyte differs in the subunits (like MDMA in Ecstasy tablets). For further information on sample heterogeneity and influence of quantitative analysis, see (ENFSI, 2016).

Another ‘external’ contribution to variability and therefore to the measured variance may be caused by contamination or impurities on the surface of the item to be analysed (identified or compared). The relevance of this contribution is dependent on the question, on the item and on the chosen analytical scheme. It could be absolutely irrelevant for e.g. classification, when the target substances can be separated from impurities, or very relevant when the impurities are considered relevant for a question on comparison, like in drug profiling with natural impurities and diluting agents.

3.4.2. Measurement uncertainty

Several components like internal variability (explained in the previous subsection), bias and drift contribute to measurement uncertainty. Practically, measurement uncertainty especially matters in quantitative measurements (determining the concentration of an analyte in a sample), also see (ENFSI, 2024). The measurement uncertainty of a procedure is determined during its validation and following the ISO 17025 norm should be a known quantity. Measurement uncertainty as a whole does not play a role in the determination of likelihood ratios. Thus, the current guideline will not go much deeper into this concept.

3.4.3. Limitation

Every chemical-physical examination method or analytical technique has limits. It can pertain certain characteristics according to capabilities of the technique, but not other characteristics. Moreover, every analytical technique is characterized by its specificity and selectivity, both of which are essential elements of method validation. Specificity involves amongst others the limit of detection or limit of quantitation. If a sample is only present at trace levels, an otherwise straightforward identification or classification can be restricted, as the desired characteristics may not be sufficiently represented against background noise.

Another limitation in the area of molecular recognition (material identification) concerns isomers. An analytical technique may, for instance, be unable to differentiate between meta- and para-positions in an aromatic ring. If relevant to the forensic question, these positions must be represented by other, more selective techniques. This limitation concerning positional isomers should not be regarded as an error but as a limitation of the method. It simply cannot distinguish between two positions of substituents on the aromatic ring. However, the substituent on the aromatic ring itself and the rest of the molecular structure remain undisputed. A reasonable forensic presentation of results should therefore use a discrete ‘either-or’ approach and should not present the result as a density distribution or a continuum between meta and para. The substituent cannot be between positions.

A further limitation performing classifications involves systems whose complete characteristics are not known yet, or where the existing datasets do not yet encompass the full characteristics. Such systems are still open (in the sense of not complete) and can be expanded. Statistical methods can indeed identify some of the characteristics and commonalities within a dataset, allowing samples with the same partial characteristics to be grouped. This is known as clustering. A solid classification (closed system) is not possible because the full characteristics are not yet known or available in the dataset. The group characteristics can still change, expand, or branch within clusters. In classifications, the system is fixed in advance.

3.4.4. Error

In this guideline we use the term error as a misclassification, either in material identification or in comparison studies (individualisation). Sometimes measurement uncertainty is referred to as measurement error. Here, we aim to use the terms as clearly and consistently as possible, avoiding the term 'error' when referring to variation or limitation.

An error in classification is for example possible when an insufficient number of characteristics are used. Two similar classes may share several common characteristics. Therefore, method validation must ensure that sufficient selectivity is considered to enable a correct classification.

The highest risk for errors lies in performing individualisation. Partly because the entire population of individuals is not available (limited dataset), and the intra-variability cannot be adequately differentiated from the nearest similar (unknown) individual.

3.4.5. Human error

Human error such as contamination, mixing-up of specimens or erroneous reading-off instruments may occur. Such errors are not inherent properties of an instrumental-analytical technique and are not an integral part of the performance characteristic to be validated. Errors of this kind must be handled by the quality control procedures applied at the laboratory.

Human errors should not be confused with other sources of uncertainty as described in sections 3.4.1 to 3.4.4.

3.5. Likelihood Ratio

The forensic service provider can take steps to reduce some types of the variability described above e.g. by raising the sensitivity and/or increasing the sample size.

However, there are sources of variation that are out of control for the forensic service provider. Reporting a conclusion about whether two seizures of cocaine have a common origin is affected not only by how similar the two seizures are from a chemical point-of-view, but also by how common the chemical characteristics are among all potential origins of the seizures. The fact that two seizures show almost identical chemical characteristics is not sufficient evidence that they have a common origin. Hence, the uncertainty on the interpretation of the evidence induced by the possibility that seizures from different origins may share the same chemical characteristics requires a quantification of uncertainty included in the report. This quantification may be made through a likelihood ratio.

Though a likelihood ratio in this sense expresses uncertainty in answering a forensic question, it is more related to express a degree of certainty, a degree of ‘yes’ to the question as mentioned in section 3, i.e. how probable an observation (like an analytical result) is to be found under a hypothesis 1 (e.g. common origin of two seizures of cocaine) compared to find the same observation under an alternative hypothesis 2 (e.g. different origins of two seizures of cocaine).

Identification and classification are well-established terms in forensic chemistry and are seldom associated with uncertainty that could have any impact on the conclusions drawn. Such categorical classification is particularly the case for chemical analysis of drugs. In cases like these it makes no sense to use likelihood ratios to quantify whether the result is correct. However, some classification goes beyond unambiguous identification of the class of substances to which the identified substance belongs and are as such a probabilistic classification. As an example, the forensic question can be whether fire debris sent to the laboratory contain traces of ignitable liquids. The analysis may show a chemical pattern that is known to be found in motor gasoline, which belongs to the class of ignitable liquids. However, it cannot be excluded that the same pattern can be found in another product, e.g. in the matrix that was burnt, that does not belong to the class of ignitable liquids.

Assigning likelihood ratios answering such types of forensic questions has gained importance and popularity in forensic sciences in recent years, especially for source attribution or when comparing different hypotheses regarding the course of events in criminal cases. In the example with fire debris, a commissioner may have reasons to believe that ignitable liquids were used to set the burnt object on fire, but the output from chemical analysis cannot be reported as an unambiguous confirmation (or disproval) of such an opinion. Instead, the forensic expert must provide a likelihood ratio for the hypothesis that there are traces of ignitable liquids in the fire debris. This likelihood ratio depends on how common the chemical pattern is deemed to be among products that are classified as ignitable liquids and also among products that are not classified as ignitable liquids.

3.5.1. Likelihood ratios versus categorical conclusions

An alternative way of reporting conclusions is in a categorical form (“The fire debris sent to the laboratory contains traces of ignitable liquids”). A problem with this way of reporting is

that an error rate will be involved, varying depending on the sample's complexity or the expert's level of expertise. Although principally speaking the error rate of an expert (or expert team) can be estimated, for example when datasets of casework with known true conclusions are available, a disadvantage of reporting categorically is that even if the error rate is known, it only shows how often someone is right on average, not how strongly the current evidence supports the conclusion. The likelihood ratio is devised precisely to handle the latter question, which is one of the reasons that a case specific likelihood ratio by a forensic expert is considered to be more appropriate and transparent.

3.5.2. Likelihood ratios and scales of conclusion

Likelihood ratios are often reported using a verbal scale of conclusions to aid interpretation, e.g. as described in Nordgaard et al. (2012) or European Network of Forensic Science Institutes (2015). However, such conclusion scales are not unique to the use of LRs. Historically—and in many countries still today—the forensic community has provided factfinders with conclusion scales that consider the main hypothesis only. In source attribution, for example, these scales focus solely on how well two objects match, without accounting for alternative explanations. Although such match-based scales may appear similar to LR-based conclusion scales, they overlook a critical aspect of forensic reasoning: the uncertainty associated with the alternative hypothesis.

In contrast, likelihood ratios explicitly evaluate the probability of the observed evidence under both the main and alternative propositions. This dual consideration provides a more comprehensive and logically consistent approach for evaluating and communicating the evidential value.

3.5.3. Guidance on the use of likelihood ratios

In a preceding project, cf. (Ahrens et al, 2021), the processing of datasets using chemometric methods was already described. The chemometric result, which, for example, represents the degree of similarity with a reference sample or dataset through a numerical value, was shown to require evaluation by the expert in two ways: operationally (chemometrically) and chemically.

In this guideline, we aim to provide guidance to the forensic chemist on how to use datasets related to specific sample materials (e.g., drug samples, fire debris samples, vehicle paints) with statistical methods to derive forensic results and where it is appropriate to use a likelihood ratio for answering the question (Huhtala et al, 2023). In sections 4 and 5 we will thoroughly explain the concept of likelihood ratio, and how and why assessment of the LR-model is to be performed.

4. BACKGROUND

4.1. Project background

In the previous project ‘Steps Towards European Forensic Area - STEFA, WP7’ (Ahrens et al, 2021) the usual forensic workflow and relation between different phases in forensic casework involving illicit drugs were explained. The process starting from crime scene and ending by the reporting of results in court including the questions asked, for example by investigative police, prosecutor, defence attorney or court of law, were covered.

Applications using chemometrics in a more or less standardized manner were demonstrated by three case examples involving three types of illicit drugs: ecstasy tablets, amphetamine and cocaine. (Salonen et al, 2020) The examples included low-and high-dimensional data, so-called Type 1 and Type 2 data which were used for identification, classification, comparison and quantification purposes. (Bovens et al, 2019)

However, the chemometric results as such should never be reported alone. Before reporting, quality assessment steps, which may consist of operational, chemical, and forensic assessments are required. In each case a forensic chemist needs to consider the suitability of chemometric methods, based on their strengths, weaknesses, opportunities and threats (SWOT). This is because while chemometric methods are powerful tools managing complex data, they are to some extent chemically blind. (Huhtala et al, 2023)

Depending on the question and the context in which the results are used, a strength of the results could be reported. This can be done by appointing a likelihood ratio value when there are two different hypotheses presented.

4.2. Theoretical background

The likelihood ratio (LR) is a classic concept in statistical inference. One might think that this concept is specific to forensic interpretation, as it is often mentioned in modern literature on the subject. However, the concept is purely statistical and is not specifically related to forensic science. As its name suggests, it is a ratio of two likelihoods, but what is a likelihood? In statistics, the likelihood of a hypothesis (or model) corresponds to the probability (or probability density) of observing a specific set of data if that hypothesis is true. The key difference with the concept of probability lies in the perspective: probability is concerned with the likelihood of different outcomes for a given model, while likelihood is concerned with how 'plausible' different models (or hypotheses) are for a set of observed and fixed data. Although likelihood is calculated as a probability, it is used to evaluate the strength of evidence in favour of different hypotheses rather than the future frequency of events.

4.2.1. Conditional probabilities

To understand likelihoods, it is necessary to understand the concept of a *conditional probability*. Conditional probabilities are assigned to events when there is information (or information is assumed) about events, scenarios etc. related to the event of interest. For instance, if parts of an unknown plant (crushed leaves, stems and buds) are to be analysed for their chemical compounds, and the event of interest is whether the plant contains illegal amounts of classified narcotic substances, the probabilities on forehand of this event are

different should the plant be hemp taken from an established rope producer compared to if it was seized from a private person.

In this example we would therefore differ between

- the probability that the plant contains illegal amounts of narcotic substances
- the probability that the plant contains illegal amounts of narcotic substances given it was taken from a rope producer
- the probability that the plant contains illegal amount of narcotic substances given it was taken from a private person

The last two probabilities are examples of conditional probabilities, while the first is a general probability, not assuming any specific origin of the plant (no scenario behind). Generally, all probabilities are conditional since it is not possible to avoid assumptions completely. As a simple example, consider the tossing of a coin, e.g. a one-euro coin. The coin has two sides, and some intuition tells us that for each side the probability of landing upwards should be 50%. However, such probabilities require that the coin is symmetric, that it is impossible for it to land sideways, and that the person tossing the coin is not tampering with the tossing experiment in any way (e.g. holding it in their hands with one of the sides upwards and just dropping it smoothly on the surface where it is supposed to land).

In forensic interpretation it is common to compare *conditional* probabilities with each other. In the above example it may be of interest to compare the probability that the plant contains illegal amounts of narcotic substances given it was taken from a rope producer with the probability that the plant contains illegal amounts of narcotic substances given it was taken from a private person. This is particularly interesting once we have obtained the results from the chemical analysis. Note that at that stage nothing is uncertain about the results, they are already given, so it may seem unnecessary to consider the probability of obtaining them (cf. subsection 3.1. Identification and subsection 3.4.2. Measurement Uncertainty). However, the two conditional probabilities to be compared still say something about the two different conditions, i.e. that the plant was taken from a rope producer and that the plant was taken from a private person respectively.

When an event is observed, the conditional probability of that event given a specific scenario is interpreted as the likelihood of that scenario. It measures how well the scenario explains the observed event. Note that this is not the same as the probability of having the scenario. The conditional probability is rather a retrospective measure of the likeliness of the scenario, i.e. observing the event and making inference about its cause[1].

4.2.2. A more formal description

Using symbolic terms, we write E for the observed event and H for the specific scenario. The probability of the event given the scenario (or expressed differently: conditioning on that the specific scenario is true) is written as

$$P(E|\mathrm{H})$$

[1] This is referred to as *inference to the best explanation*. See e.g. Taroni et al (2014).

where *P* stands for probability. The bar (|) separates what is addressed by the probability, here *E* and on what condition it is addressed, here H.

P(*E*|H) has two possible meanings. If H is known but nothing has yet been observed (no analysis has been made), then *P*(*E*|H) expresses how probable it is to obtain *E* (from the analysis to be made). If *E* is obtained beforehand from the analysis, but it is not known whether H is true (or if some other scenario is true), then *P*(*E*|H) expresses how well H explains *E*.

Suppose now we have obtained *E* from the analysis beforehand and there are several possible scenarios, which we can denote H_1, H_2, ... Then there are as many likelihoods as there are scenarios, i.e. *P*(*E*|H_1), *P*(*E*|H_2), ... Since each of these likelihoods are expressed as conditional probabilities of *E*, they do not have to sum to 100%. It is very well possible that every likelihood is close to 100% or close to 0%. For instance, if the event is that the outcome of the chemical analysis of plant material from an unknown plant is that it contains 10% THC, and the scenarios are all about the origin of the plant, then several of these origins would give a substantial probability of obtaining such a concentration of THC. The sum of these probabilities over a large set of origins (for instance over 100 private persons suspicious to grow cannabis sativa for drug production) would certainly exceed 100%.

4.2.3. In terms of a case

Now, assume we have two scenarios that we denote H_1 and H_2 respectively. H is here short for *hypothesis* indicating that we don't know whether this scenario is true or not. H_1 usually stands for the *main hypothesis*, which is the hypothesis we beforehand have reasons to suspect is the true one. H_2 then stands for the *alternative hypothesis*. As an example, assume we have a seizure of (a plastic bag of) cocaine powder, taken from an arrested person, and we seek to answer whether the seizure originates from a powder mixture seized from distributor X or from a powder mixture seized from distributor Y. H_1 and H_2 would then be

H_1: The seizure originates from the powder mixture with distributor X (PM X)

H_2: The seizure originates from the powder mixture with distributor Y (PM Y)

Besides cocaine (i.e. cocaine hydrochloride) the chemical analysis also reveals a low concentration of lidocaine hydrochloride in the seized powder. This is known to be a common cutting agent in cocaine powder. An analysis of PM X shows a quite substantial concentration of lidocaine hydrochloride, while an analysis of the powder PM Y shows no amounts of it.

These findings do not tell us with 100% certainty that cocaine seized from the arrested person originates from PM X, even if we should assume that PM X and PM Y are the only possible origins of the seized powder. This is so since we have only observed a low concentration of lidocaine hydrochloride in the seized powder, and we don't know how homogenized PM Y was at the time when the seized bag was filled with cocaine powder. Moreover, since the concentration of lidocaine hydrochloride in PM X is substantial, while it is low in the seized powder, there is lack of conformity between the two.

What we can do is to assess how probable it is that powder originating from PM X has a low concentration of lidocaine hydrochloride given the concentration is substantial in PM X. Furthermore, we can assess how probable it is that powder originating from PM Y has a low concentration of lidocaine hydrochloride given that the analysis shows no lidocaine hydrochloride in PM Y.

Such assessments depend on how representative the analysis samples are from PM X and PM Y of the true concentrations in these powder mixtures, both at the time of analysis but also at the time point when the seized bag was filled.

Let E denote the findings regarding lidocaine hydrochloride in the powder seized from the arrested (low concentration), in PM X (substantial concentration) and in PM Y (none). The probabilities to be assigned can be written

$$P(E|H_1,I) \text{ and } P(E|H_2,I) \tag{4.2.1}$$

where | stands for the information available for these probability assignments, i.e. what we know about the representability of the samples taken from PM X and PM Y respectively, and what we know about the time point when the seized bag was filled in relation to the time point when PM X and PM Y were confiscated. Moreover, I also contains knowledge about the analysis instrument, the laboratory conditions etc. Note that generally no probability of I is considered and it is not a matter of interest. But I is important since different amounts of information would generally imply differences in assigned probabilities. Scarce information leads to less precision in the assessment. As an example, a forensic scientist may assess a probability of an event to be about 33-35% when there is lots of information available, while in a situation with much less information the probability of the same event is assigned to be about 20-40%.

Here $P(E|H_1,I)$ is understood as the likelihood of H_1, i.e. the likelihood of the origin of the seizure being PM X, and $P(E|H_2,I)$ is understood as the likelihood of the origin being PM Y, both in light of the forensic findings E. . Now, the *likelihood ratio* of H_1 versus H_2 is simply the ratio of these two likelihoods, i.e.

$$LR_{H_1 \, vs \, H_2} = \frac{P(E|H_1,I)}{P(E|H_2,I)} \tag{4.2.2}$$

Correspondingly, the likelihood ratio of H_2 versus H_1 is

$$LR_{H_2 \, vs \, H_1} = \frac{P(E|H_2,I)}{P(E|H_1,I)} \tag{4.2.3}$$

What do these likelihood ratios mean? To explain this, it is convenient to have numerical values entered. Assume we have assessed $P(E|H_1,I)$ to be fairly low, say 10% (i.e. 0.10). This can be due to that we consider our sample from PM X to be representative, but we have difficulties in deeming whether PM X was homogeneous or not when the seized bag was filled from it. Note that when conditioning on H_1 H,we take the standpoint that the powder in the bag originates from PM X, which is why conditioning on PM Y becomes irrelevant for $P(E|H_1,I)$. Assume further that we have assessed $P(E|H_2,I)$ to be very low, say 1% (i.e. 0.01). This can be because we consider our sample from PM Y to be representative and that we have few reasons to believe that PM Y was substantially heterogeneous when the seized bag was filled from it.

The likelihood ratio $LR_{\mathrm{H}_1 vs\, \mathrm{H}_2}$ then becomes 0.10/0.01 = 10. This means that our findings *E* are ten times more probable if the powder in the bag originates from PM X compared to if it originates from PM Y. The likelihood ratio $LR_{\mathrm{H}_2 vs\, \mathrm{H}_1}$ becomes 0.1, but the interpretation is the same. It is recommended to interpret likelihood ratios in terms of how much higher the highest likelihood is compared to the lowest. Other ways of verbal interpretations of the likelihood ratio become mathematically inconvenient[2].

Now, $LR_{\mathrm{H}_1 vs\, \mathrm{H}_2}$ = 10, means there is support for H_1 over H_2 and the support is of magnitude 10. Any value of $LR_{\mathrm{H}_1 vs\, \mathrm{H}_2}$ above one means support for H_1 over H_2 and any value below one means support for H_2 over H_1. Different levels of support are given by the very value of $LR_{\mathrm{H}_1 vs\, \mathrm{H}_2}$.

4.2.4. Likelihood ratios with continuous-valued findings

In most casework within forensic chemistry continuous-valued measurements are taken (e.g. peak areas, concentrations). For such observations it is meaningless to assign probabilities of individual values. This is so since the probabilities of all possible outcomes of an experiment must sum to one. However, the possible outcomes of measuring a quantity of interest are uncountably many, so the only possible probability to assign to a specific value is zero. It may be possible to divide the possible range of measurements into intervals and assign probabilities to these intervals. However, when the intervals are made smaller and smaller to approach the resolution of the measurement equipment, the probabilities become ridiculously small.

As an example, suppose we are comparing glass objects measuring the refractive index with a resolution of five decimals. Typical values are around 1.47-1.58, meaning that measurement values such as 1.51085 and 1.54884 can be obtained. A value of 1.54884 is then representative for all values included in the interval (1.548835,1.548845). However, probabilities of intervals of that length are very small. In the range from 1.47000 to 1.58000 there are 11 000 values with five decimals' resolution, and hence 11 000 intervals. Many of these must have very small probabilities for all 11 000 to sum to one. Uniformly distributed probabilities would give the probability 1/11 000 for each interval, but since we can expect some ranges of refractive indices to be much more common than others, a uniform distribution is not realistic. Instead, there must be some ranges where the intervals have fairly high probabilities to conform with reality, which may in turn give other intervals far too low probabilities compared with reality.

Assume now that the hypothesis H_1 is that a recovered glass fragment originates from a window smashed in the course of a burglary, and the hypothesis H_2 is that the fragment originates from another smashed window. The refractive index of the fragment and the smashed window where the burglary took place are both measured as 1.54884. Using intervals, we would then express the findings *E* as the measured refractive indices of the glass fragment and the smashed windows both being in the interval (1.548835,1.548845). If glass having refractive indices in that interval is rare, the probability of the interval may be

[2] As an example, if the likelihood ratio is 0.1, you should avoid interpreting it as the findings being ten times *less* probable if *H* is true compared to if *A* is true. "ten times *higher*" is a mathematically proper expression, while "ten times *less*" is not.

unrealistically small and consequently the likelihood ratio calculated for H_1 against H_2 will overstate the support for H_1.

From here on, we refer to a measurement originating from a suspected source (here a smashed window) as measurement on *control* or *reference* material, and on questioned material (from an unknown source) as a measurement on *recovered* material. Note that the distinction between the two types may not be completely clear-cut in case both measurements are from an unknown source, and that the word recovered might be interpreted more broadly than for crime scene material. Because of the use of the terms in the literature, we adhere to the terms control and recovered material.

The valid probative measure of continuous data is a so-called probability density function. This function is continuous and has a value for each finite set of measurements[3] (assuming an infinite resolution). A set of measurements on rare glass would have a low density of refractive index, while a set of measurements on common glass would have a high density. The expression for the likelihood ratio in light of a measurement on recovered glass (here the fragment) and a control measurement (here the smashed window) will then be

$$LR = \frac{f(x, y|\mathrm{H}_1, I)}{f(x, y|\mathrm{H}_2, I)} \tag{4.2.4}$$

where f is the probability density function and x and y are the measurement values. It can be that H_1 and H_2 imply different probability density functions. An illustration of a density function of two measurements is given in Figure 4-1. What is different compared to assigning probabilities to intervals is that each pair of measurements has a density that reflects its relative commonness/rareness in the population of possible pairs of measurements without any restriction that these relative values should sum to one (1). The restriction on the probability density function is that it is always non-negative and the integral over its entire support must be equal to one.

From a probability density function it is possible to compute probabilities of intervals by integrating the function over these intervals. This is however not the same as distributing probabilities that sum to one (100%) over several intervals. A particular measurement can be seen as a representative value for all its neighbouring values (cf. the interval description for refractive indices above) that cannot be precisely measured due to their higher resolution, but it is better to use the probability density function evaluated at this representative value, than assuming any value in its neighbourhood to be the 'true' measurement with uniform distribution (could the resolution be increased).

It should be said that the equation (4.2.4) for the likelihood ratio is generic and can be rewritten using probability calculus to obtain the more useful expression

$$LR = \frac{f(y|x, \mathrm{H}_1, I)}{f(y|\mathrm{H}_2, I)} \tag{4.2.5}$$

where y is the measurement on the recovered material and x is the measurement on the control material (see e.g. Aitken et al, 2020). For one-dimensional measurements the density functions in the numerator and denominator are functions of one variable only,

[3] By set we mean that measurements are taken both on the recovered glass and on the control glass.

which simplifies the deduction of such functions (cf. the function of two variables in Figure 4-1). Equation (4.2.5) is the basis for doing so-called feature-based evaluation of forensic evidence (see further section 4.2.5).

An analogous reformulation can be done for LR-expressions with probabilities. We then separate the findings on the recovered material E_R from the findings on the control material E_C (hence $E = (E_R, E_C)$ and write the likelihood ratio as

$$LR = \frac{P(E_R|E_C, \mathrm{H}_1, I)}{P(E_R|\mathrm{H}_2, I)} \tag{4.2.6}$$

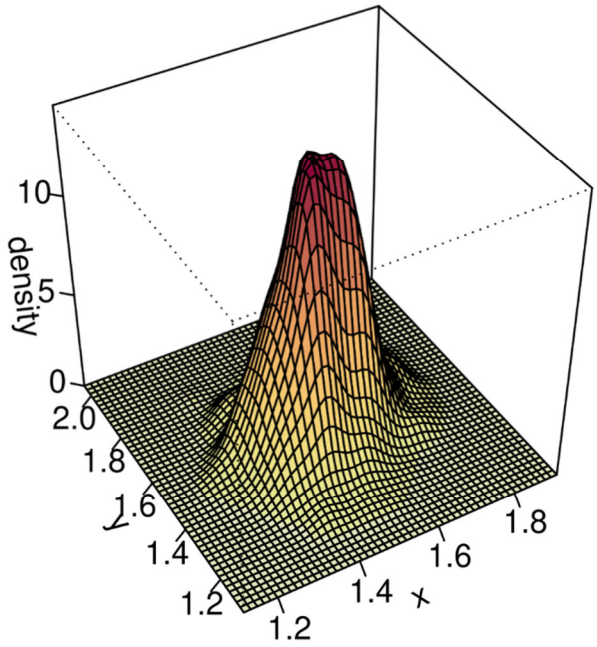

Figure 4-1. A probability density function for a set of two features (x and y)

4.2.5. Feature based and score-based LR's

In section 4.2.4 the likelihoods and the likelihood ratios were assigned using probabilities or probability densities of the observations/measurements directly. The probability distributions used are then univariate or multivariate depending on the dimensions of the observations/measurements. For instance, if a measurement is a vector of 14 monitored peak areas in a chromatogram, the probability distribution is 14-dimensional. For the comparison of two materials, a conditional probability distribution of the measurements on material 1 given the measurements on material 2 and given hypothesis H_1 and a conditional probability distribution of the measurements on material 1 given hypothesis H_2 are used (cf. Equation 4.2.5). This way of assigning likelihood ratios is referred to as *feature-based*

evaluation. By feature is meant a particular variable or a function of one or several variables that has an important interpretation regarding the type of material studied. It could be a single peak area, but also a ratio of two peak areas, a mathematical transform (like the logarithm) of a peak area etc.

Feature-based evaluation uses all information there is in a measurement with respect to the probability distribution ruling the possible outcomes of the measurement. However, this distribution needs to be estimated from background data and the higher the dimension the more data is needed for the estimation to be robust.

When the amount of background data is small and the dimension of the measurement is fairly high, feature-based evaluation is not reliable even if it is possible to fit a probability distribution. It may be possible to perform dimensionality reduction, like principal component analysis, but if correlations between the dimensions are not very high or very low, the number of principal components needed to capture a high proportion of the total variance of the measurements may still be too many to succeed well with the estimation of a probability distribution.

An alternative to feature-based evaluation is to change the perspective from evaluating the very features obtained (joint probability of the features of the two materials in a comparison) to evaluating the pairwise similarity (or distance) between the materials. A measure of similarity (or dissimilarity) is then used, like a multivariate distance function (Euclidean distance, Canberra distance etc.) or a score obtained from a machine learning procedure. This measure is referred to as the *score* from the comparison. Probability distributions of this score valid under hypothesis H_1 and hypothesis H_2 respectively are then used to assign the likelihood ratio, and this is referred to as *score-based* evaluation. Technically, this can be described as

$$LR = \frac{f(s|\mathrm{H}_1)}{f(s|\mathrm{H}_2)} \tag{4.2.7}$$

where *s* is the score. This modelling step is sometimes referred to as 'calibration'. There are more ways to perform calibration, for example by using logistic regression.

The benefit of using score-based evaluation is that it reduces (normally) the dimension down to one and hence the amount of background data needed to estimate the probability distributions is much less than for the feature-based evaluation. The drawback is that a score does not capture all information there is from the measurements involved in the comparison. In particular, a similarity score says nothing about the rarity of the measurements in the population of interest.

4.3. Bayesian reasoning

Bayesian reasoning is an approach to statistical inference where probabilities are commonly interpreted as degrees of beliefs. To understand it, let´s compare it to another probabilistic approach to statistics, the frequentist reasoning.

The frequentist view focuses on the repetition of an experiment under similar conditions, and probability is interpreted as the long-run relative frequency of occurrence. The frequentist view is recognized by tools as t-tests, confidence intervals and p-values. These valuable tools are widely recognized, as the frequentist approach is the primary method

taught to students, particularly at the introductory level. In fact, in natural sciences students often only encounter the frequentist view in probability courses.

The personal belief of a rational individual is preferably based on data, relevant to the question. In this regard the Bayesian does not differ from the frequentist view. However, data may be lacking for a specific question because some questions are unusual. Other situations may be non-repeatable, and data therefore unobtainable. It is in such circumstances that a fundamental difference between the two approaches becomes apparent.

Since a probability to a Bayesian approach is a measure of *degree of belief,* rather than a *long-term frequency,* it is entirely possible to assign a probability to a non-recurring situation. For example, a game of football is played only once under similar conditions. But that doesn't stop us from having a belief about the outcome of the game based on personal judgment, experience, and intuition. From a frequentist perspective, assigning a probability to a specific outcome requires reference to data from repeated games played under sufficiently similar conditions, whereas the Bayesian framework permits probability statements to be made even in the absence of such repeatable data.

Another important difference is that a frequentist makes binary decisions by accepting or rejecting a null hypothesis, based on the idea of statistical significance. A Bayesian does not make binary decisions on hypotheses based on statistical significance but revises their degree of personal belief by adjusting the odds of one hypothesis being true compared to another. Bayesian reasoning centres on evaluating the probability of hypotheses given observed data. In contrast, the frequentist framework treats probability as a long-run frequency of observed data under a fixed hypothesis and thus does not focus on probabilities to hypotheses themselves.

As an example of the different views, consider the probability of rolling a six on a die. The frequentist does not pronounce on the probability of getting a six until this particular die has been rolled a number of times. The Bayesian on the other hand might first state that the probability of a six is expected to be one in six, because prior knowledge says that dice are usually fair. After the Bayesian has rolled the die a few times, the degree of personal belief may change to something else for that particular die. But both the initial and the revised expectations are considered rational probabilities by the Bayesian, and as such they both follow the laws of probability.

4.3.1. Why Bayesian reasoning is appropriate in the evaluative framework of forensic science

Bayesian inference is appropriate in the evaluative framework of forensic science because it provides the factfinder with a structured, probabilistic framework for updating beliefs based on new evidence. Forensic scientists assist the factfinder in inferring the most probable cause of an observed outcome by applying logic, probability theory, and systematic methodology to the available evidence (Berger, 2010).

Logic: Forensic scientists should not answer questions about the probability of propositions, e.g. "what is the probability that this has happened" or "the probability that X is guilty". Assigning probabilities to propositions, prior and posterior to the analysis of the forensic scientist, is the domain of the court. The forensic scientist does not have all information in

the case, which is necessary for an assessment of the prior odds, nor does the expert have that role in the legal system.

But scientists have indeed tried to answer questions about propositions in the past, probably not realizing they are trespassing into the domains of the court (Evett, 1998). Fortunately, Bayes' theorem enables us to clearly identify the legal question appropriate for science:

"what are the probabilities of *the evidence* given the hypotheses?".

which are exactly the verbal analogues of the expressions in (equation 4.2.1)

$$P(E|H_1,I) \text{ and } P(E|H_2,I)$$

To understand the difference between this question and the question suited for the court, and the logic behind this division, is essential if one is to claim the evaluation of evidence to be scientific.

In Bayes theorem, the likelihood ratio is a separate entity in the mathematical equation. The likelihood ratio should be multiplied to the prior odds to give the posterior odds. Clearly pointing out the likelihood ratio as the forensic domain, and the prior odds and posterior odds as 'no-go-zones' for the forensic scientists is perhaps the greatest benefit of the Bayesian approach.

A conceptual illustration of likelihood ratios can be visualized by imagining the factfinder in a room, where two opposing walls represent one of two competing hypotheses: the prosecution proposition (H1) and the defence proposition (H2). Prior to evaluating the forensic evidence, the factfinder occupies a position within this space that reflects his or her initial belief relative to the two hypotheses, i.e., the prior odds. These prior odds are based on information that may not be accessible to the expert, such as testimonies or evidence presented by other practitioners, and personal beliefs. As such, each person's prior state is likely to be different and unknown.

In Figure 4-2, this initial position is depicted as being closer to the wall representing H_2 for factfinder Jim and near the centre of the room for factfinder Jane, i.e. Jim believes H_2 is more probable, and Jane believes it to be fifty-fifty. Upon receiving the forensic evidence, presented as a likelihood ratio, the factfinder metaphorically 'throws a ball' across the room, and the distance of the throw corresponds to the strength of the likelihood ratio. The starting point represents the prior odds, the length of the throw is determined by the likelihood ratio, and the ball's landing point represents the updated belief, or the posterior odds.

In Figure 4-2, the factfinders receive the same likelihood ratio from the forensic scientist. The forensic report finds the results more probable under H_1 than under H_2, and both factfinders throw the ball an equal distance in the direction of the H_1-wall. Since the forensic scientist does not know where in the room the factfinder starts it is not possible to tell where the ball lands. It is not enough to know the length of the throw. In other words, because the forensic scientists do not possess knowledge of the prior odds, they are limited to reporting the likelihood ratio alone and must avoid making statements about the posterior odds.

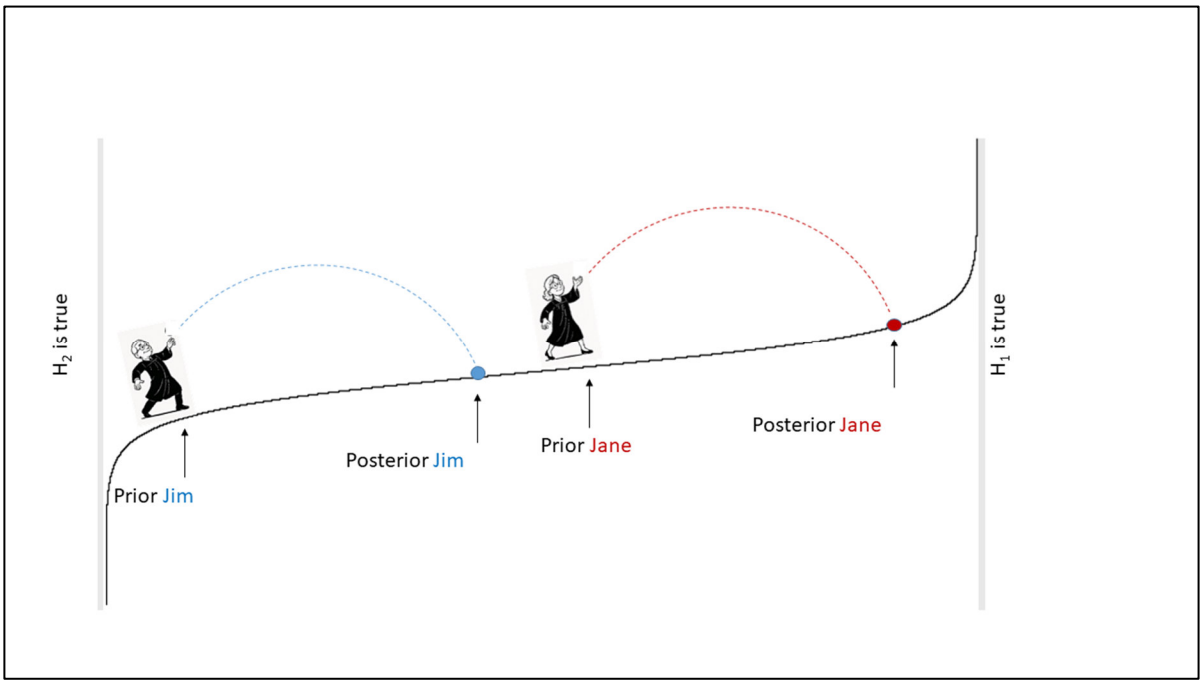

Figure 4-2. A graphical representation of likelihood ratios with respect to prior odds and posterior odds of the factfinder.

Like the likelihood ratio, both the prior and posterior odds are expressed as ratios. In the case of odds, the numerator and denominator represent the probabilities of two mutually exclusive and exhaustive hypotheses, and thus their sum equals one. Consequently, knowing the probability of one hypothesis immediately determines the probability of the other. In contrast, the numerator and denominator of the likelihood ratio do not typically sum to one, as they are conditional probabilities under two competing hypotheses. This distinction is akin to knowing your precise location within a room when using odds, whereas the likelihood ratio alone provides only directional information without specifying absolute position.

Probability theory: In a trial, evidence themes of various kinds occur. The overarching theme of proof in a criminal trial is the prosecutor's description of the crime, and it is referred to as the main theme of the trial (Dahlman, 2018). But in a criminal trial there are also sub-themes. A sub-theme can be a certain prop, e.g. intent, or it can be to prove that a suspect was at the scene.

Since Bayesian probability is a measure of a '*degree of belief*', rather than a '*long-run relative frequency*', it is entirely possible to assign a probability to (Taroni et al, 2023):

(1) a non-recurring situation, such as a crime (main theme).

(2) the typicality of a feature, such as a randomly acquired characteristic of a shoe mark, with no or limited background data (sub-theme).

In a criminal trial, the court has to ponder on both the main theme and any sub-themes. Note that both types of themes can be, and usually are, referring to non-recurring situations. To decide whether the hypotheses presented by the two sides are sufficiently supported by evidence, the court must therefore in both cases rely on the use of personal judgment, experience and intuition, i.e. assign Bayesian probabilities.

Forensic sciences are usually involved to evaluate propositions related to sub-themes. Because there are few fully quantitative forensic evidence evaluations, the situation of the forensic scientist is no different than that of the court. Personal judgment, experience and intuition come into play. The possibility to assign probabilities to non-recurring situations is crucial for the legal system as a whole, and forensics are no exception.

Methodology: The judiciary constantly needs to update the assessment of the probability of an event given new data. However, to update beliefs dependent on at least two different propositions a probabilistic methodology is needed. Because Bayes theorem is an inevitable consequence of probability theory it is a method for revising one's belief in the light of new information. Bayes theorem states exactly how the different propositions can be pitted against each other. Thus, Bayes theorem is the fundamental formula of forensic science interpretation, and its importance cannot be over-emphasized (Evett, 1998).

4.4. Literature survey

Since the turn of the century there has been an increasing number of publications on determination of Likelihood Ratios in forensic chemistry. A systematic literature search of these is described in (Van Lierop et al, 2024), using keywords in the Scopus database. In this publication an overview is found of separate areas of chemical forensic expertise in which Likelihood ratio systems have been described:

- Explosions and Explosives – Explosive Individualisation
- Food Forensics – Fraud
- Forensic Ballistics – Cartridge Case Comparison
- Forensic Ballistics – Pattern Matching
- Forensic Drug Analysis – Drug Individualisation
- Forensic Drug Analysis – Activity Level Drug Analysis
- Forensic Fire Analysis – Fire Debris Classification
- Marks and Impressions – Shoemark Comparison
- Marks and Impressions – Tape Comparison
- Materials and Microtraces – Car Fuel Individualisation
- Materials and Microtraces – Car Paint Individualisation
- Materials and Microtraces – Car Plastics Individualisation
- Materials and Microtraces – Gasoline Individualisation
- Materials and Microtraces – Geolocation
- Materials and Microtraces – Glass Classification
- Materials and Microtraces – Glass Individualisation
- Materials and Microtraces – Gunshot Residue Analysis
- Materials and Microtraces – Ink Individualisation
- Materials and Microtraces – Paint Individualisation
- Toxicology – Alcohol Abuse Prediction

Typically based on either type 1 or type 2 data, systems are developed that either quantify Likelihood ratios for classification or individualisation (comparison) questions. The list of areas is not exhaustive. Example publications for all the areas described above can be found directly in the reference list of (Van Lierop et al, 2024).

5. HOW TO CONSTRUCT AND VALIDATE LR SYSTEMS?

5.1. LR systems

Likelihood ratios are a useful tool to express evidential strength. Given observations in forensic casework, and two competing hypotheses, they quantify in how far the observations support the one hypothesis over the other. When an automated system is devised that expresses LRs in casework with the same type of questions and analysis method, one may construct an 'LR system'. An LR system is an automated procedure that has analytical observations as input and evidential strength in the form of an LR as output. An advantage of such a procedure is that it may be objectively tested, which is problematic for human experts, and that the procedure will always give the same answer given the same input.

5.2. Construction

Likelihood ratio systems may be devised in diverse ways. Typically, the system will start from observations on items which are here denoted as x and y, which before they are observed are random variables X and Y. These observations can be, for example, absorbance intensities at specific wavelengths from IR-spectroscopy or peak areas at certain retention times from chromatography. The observations contain a fixed number of elements. Following a general set-up in which given H1, x and y are from the same source and given H2 they are from different sources, the likelihood ratio under consideration may be defined as

$$LR(x,y) = \frac{P(X=x,Y=y \mid H_1)}{P(X=x,Y=y \mid H_2)}. \qquad (5.2.1)$$

Here the letter *P* may either denote a discrete probability (in the case that there are only finite possible outcomes) or a continuous probability density function (in the case that outcomes are on a continuous range of values, which are all possible), depending on the type of measurements. If statistical models are determined to estimate the numerator and denominator of equation 5.2.1, this is a *feature-based* likelihood ratio (see section 4.2.5). The statistical models may be for example multidimensional normal distributions.

In an alternative setting comparison scores $s(x,y)$ are used based on *comparisons* between pairs of items/vectors as above, leading to an LR definition of

$$LR(s(x,y)) = \frac{P(s(X,Y)=s(x,y) \mid H_1)}{P(s(X,Y)=s(x,y) \mid H_2)}. \qquad (5.2.2)$$

The latter is a *score-based* likelihood ratio (see section 4.2.5). Estimation of the probabilities or probability density functions involved may take place in any number of ways. In subsection 5.2.1 is given an account for kernel density estimation.

For score-based LRs, a general step-plan was described in (Leegwater et al, 2024) that might be followed, including Python software that may be used for this. Following an example where 10-dimensional elemental data on glass fragments are compared using different types of score functions (such as the Euclidian or Manhattan distance between the 10-dimensional outcomes), here the separate steps described are

- data exploration
- splitting of data in training and validation subsets
- data pre-processing
- calculating scores for comparisons and
- calculating LRs from scores.

The data used to construct an LR system always requires measurements on multiple samples of (most of) the sources present in the training data. Otherwise, it is impossible to model the common source hypothesis (H_1). At least two samples are required, but in practice more samples can be used.

On each of these samples, multiple repeated measurements can be done. The repeated measurements are typically summarised (or aggregated) to a single value per sample. It is important to do this aggregating in a consistent way. It should be the same when modelling both of the hypotheses, and also when evaluating a case. For example: if two repeats per sample are available in the training data, they could be aggregated by their average. An LR system can then be constructed based on these averages over two repeats. If in a case three repeats are available per sample, still an average over two (randomly selected) repeats must be used. It would be wrong to, for example, select the highest or the lowest of the three repeats, or to use the average over all three repeats or over the highest (or lowest) two repeats.

When more than two samples are available for the sources in the training data, they could also be aggregated (especially in score-based systems). This should again be done in a consistent way. It is also important that all sources are equally represented when modelling the two hypotheses. If different numbers of samples are available for the sources, a straightforward way of achieving this is by randomly selecting a chosen number of samples and discarding the others.

When collecting training data to construct an LR system, it is recommended to take into account the number of samples and repeats that could be expected in typical cases, to prevent mismatches between the numbers of measurements.

After construction of an LR system, it needs to be validated. The process of validation includes splitting the data and often performing cross validation. This process will produce metrics for the LR system. We describe possible ways to do this in subsection 5.3. Examples are given in chapter 6. For a more detailed example including programming code, see (Leegwater et al, 2024).

5.2.1. Kernel density estimation

A technique that is often used for this is Gaussian kernel density estimation, or KDE, cf. (Rice, 2007). The method uses normal kernels around observed outcomes, with a certain bandwidth, and uses this to approximate the probability density for any comparison score. The bandwidth used for the KDE is often according to Silverman's rule-of-thumb (Silverman, 1998). Silverman's rule provides optimal bandwidth for normal distributions. However, for forensic chemistry data, which is often multimodal (such as THC concentrations), strict application of this rule can result in over-smoothing of the data. In such cases, there are other ways to find the bandwidth.

Kernel density estimation is as its name reveals used for estimating/approximating probability density functions, i.e. probative measures for continuous-valued measurements. A dataset of such measurements is often illustrated as a histogram, which is done by dividing the total range of the measurements into a number of intervals with (usually) equal lengths. The number (frequency) of measurements in each interval is represented by a bar, and the resulting bars are drawn adjacent to each other, weaved together (hence the prefix *histo-* which is Latin for tissue). In Figure 5-1 a histogram drawn from measured THC concentrations in 50 seized cannabis resin materials for one year.

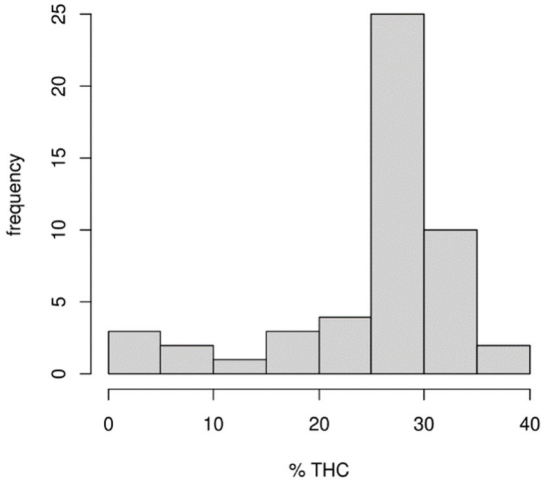

Figure 5-1. Histogram of measured THC concentrations in 50 seized cannabis resin materials.

The concentration of THC is a typical continuous-valued measurement. The histogram in Figure 5-1 indicates that values between 20% and 35% are more common (the distribution is *denser* in that range) and that there is a smaller peak in density for values between 0% and 10%. However, since the histogram is based on 50 measurements the shape of the distribution should be considered uncertain.

The 50 measurements were sampled from a larger set of 822 measurements and in Figure 5-2 is shown a histogram based on all these 822 values.

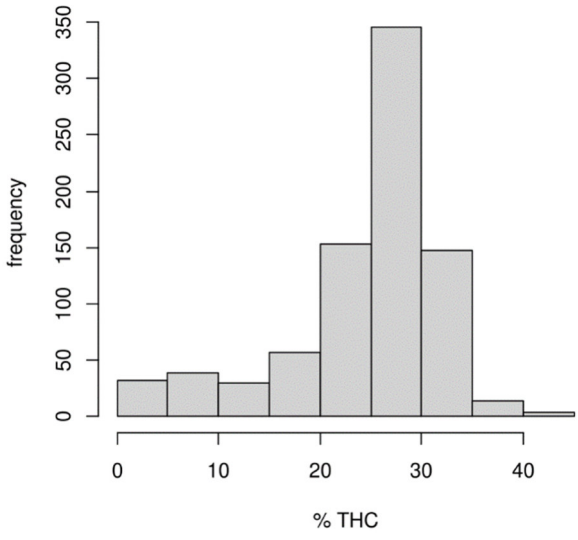

Figure 5-2. Histogram of measured THC concentrations in 822 seized cannabis resin materials.

The shape of the histogram in Figure 5-2 is much more reliable than the corresponding shape of the histogram in Figure 5-1. Here it is quite clear that the range between 20% and 35% is the densest range of values and that the distribution has a longer tail to the left (towards 0%) and a shorter to the right. The distribution is thus left-skewed.

Several chemometrics methods are based on the assumption that the analytical measurements are normally distributed. Therefore, if the shape of a histogram does not fit well with the typical bell-shape attempts are made to transform data to approach this shape (see e.g. Ahrens et al, 2021). Typically, the shape is right-skewed and transforms like the fourth-root or the logarithm can be efficient. However, there must be reasons to assume normal distribution of the transformed data for such an approach to be valid.

Should there be reasons to assume normal distribution (of non-transformed or transformed measurement values), the probability density function can be easily estimated by calculating the mean and standard deviation of the data assumed to be normally distributed and then plug in to the mathematical expression for the normal distribution[4].

When there are no reasons to assume a normal distribution or any other family of parametric distributions (like gamma distributions or beta distributions), it is preferred to fit a distribution directly to the data points (i.e. to the shape obtained from the histogram). Today's most common way of doing this is to use kernel density estimation. Simply expressed, this means that for any point, x, in the range covered by the available data, a smoothing function (a kernel) is for each available observation, x_i, in the dataset applied to the difference between x and x_i, and the resulting values are averaged. The kernel is a symmetric function and very often the probability density function of a normal distribution. The kernel can be used with different bandwidths, where a small bandwidth gives a less smoothed density function and a large bandwidth gives a much smoother density function. Silverman (1998) suggested an optimal bandwidth calculated from the data when the kernel is Gaussian.

Figure 5-3 shows the histogram of Figure 5-1 together with kernel densities with a Gaussian kernel computed from the 50 measured THC concentrations using three different bandwidths: b_1 = Silverman's optimal bandwidth, $b_2 = b_1/2$, and $b_3 = 2b_1$.

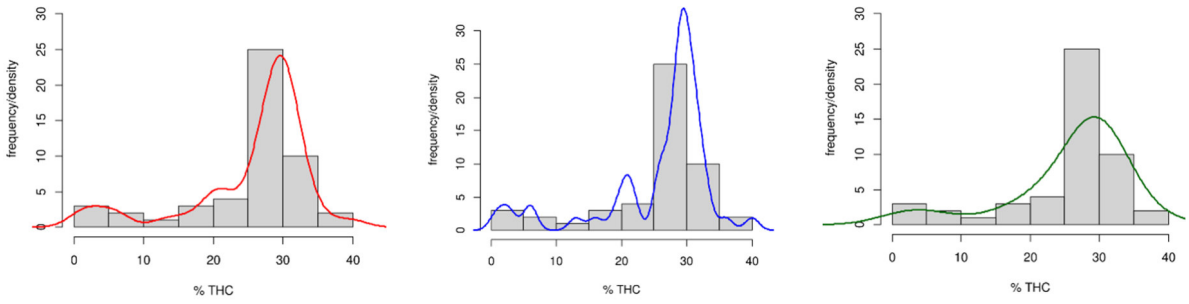

Figure 5-3. Kernel densities with a Gaussian kernel computed from 50 measured THC concentrations using Silverman's optimal bandwidth (left, b_1), half of Silverman's optimal bandwidth (middle, $b_2=b_1/2$), and double Silverman's optimal bandwidth (right, $b_3=2b_1$).

[4] For μ being the mean and σ the standard deviation, the probability density function of a normal distribution is

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

As can be seen in Figure 5-3, a lower bandwidth than the optimal gives a kernel density that follows too much the bars of the histogram, i.e. less smoothed. Such an irregular density function is not realistic. The optimal bandwidth is smooth but follows naturally the increased density for concentrations closer to 0%. As discussed before, using just 50 values to deem on the shape of the distribution is uncertain. Using a higher bandwidth than the optimal gives a much more smoothed kernel density, but this shape may also be questionable due to the fairly low number of values.

Figure 5-4 shows the histogram of Figure 5-2 also with kernel densities using the same bandwidths choices as in Figure 5-3 but now computed from 822 measures of THC concentration.

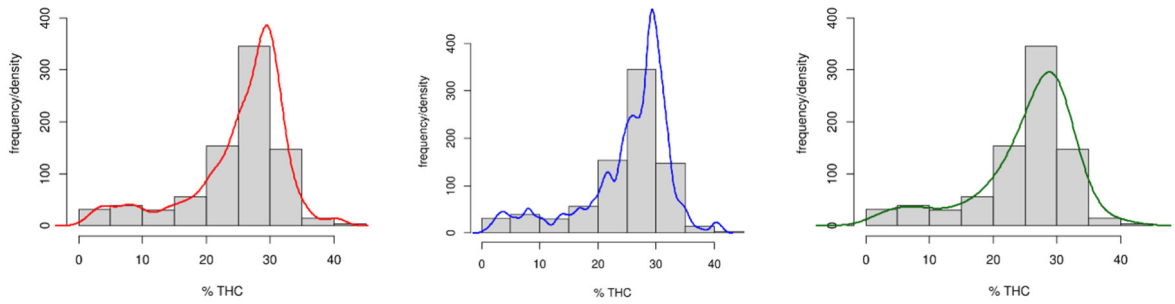

Figure 5-4. Kernel densities with a Gaussian kernel computed from 822 measured THC concentrations using Silverman's optimal bandwidth (left), half of Silverman's optimal bandwidth (middle), and double Silverman's optimal bandwidth (right).

The differences in smoothing can also clearly be seen in Figure 5-4, but the differences between the three kernel densities is smaller than in Figure 5-3. The more data used for a kernel density estimation, the less influence of the choice of bandwidth on the general shape.

It is very important to only use the kernel density for points that are within the range of the data used for the estimation. Outside that range the estimation is not reliable.

5.2.2. Elicitation

An alternative approach to building an LR system using empirical data to produce likelihood ratios is to rely on expert knowledge and gauge the relevant probabilities based on expert judgement. This process is known as elicitation. In essence, elicitation involves asking an expert or group of experts to give assessments on relevant probabilities that can be used to obtain likelihood ratios for evaluative reporting.

This approach is especially useful when obtaining empirical data is challenging due to cost or complexity of the problem, but experts still have substantial special knowledge e.g. through practical and theoretical understanding of chemistry or physical processes.

An example for this is described in section 6.3.

5.3. Validation

An advantage of likelihood ratio systems that are based on structured empirical data is the possibility to validate the methodology, i.e. to check if the system performs as intended and produces reliable, accurate, and reproducible results. In contrast, to validate likelihood ratio

models driven by a manual effort presents significant challenges due to the labour-intensive nature of the process. A meaningful validation procedure requires conducting at least hundreds of classifications or comparisons on items with known ground truth, which is a cumbersome manual task. However, a computerized method can efficiently perform hundreds, thousands and even more classifications or comparisons when presented to enough ground truth data. This chapter is intended to assist the forensic practitioner in evaluating the validity of computerized likelihood ratio models developed.

Likelihood ratio systems are based on empirical data ('data-driven') and are constructed and optimized by using some training dataset, see section 5.2. Choices, such as which distribution model or score function to employ, are usually based on the model performance on that training data. Therefore, there is a risk of overestimating how the model would perform in casework situations, in particular if the training dataset is limited. The overestimation arises from the underlying population being only partially characterized by the training data. Later on, the casework data represents a new set of data from the underlying population, which may not align with the training set. If this is the case, it is expected to have a worse performance than on the training data. Therefore, the performance of the selected LR model should be evaluated using a validation set of data. The two datasets should be separate, and preferably independent. If the training data is used also for validation, one is destined to overfit the model and end up with the risk of misleading results on new cases.

The validation set used to evaluate a likelihood ratio system should be independent from the training set—ideally not only in terms of the data it contains, but also in how and by whom it was produced. The validation data could come from different instruments, operators, or laboratory conditions than those used to generate the training data.

This is important because instruments and technicians can vary over time or between laboratories. These variations may influence the measurements or features extracted from the data, which in turn can affect the performance of the LR system. When a new instrument is used, chemists are used to validating it. However, re-validating the LR method may be cumbersome, and not necessary if this was taken into account in the initial validation.

In cases where fully independent validation data is not available, it is still critical to partition the existing data into separate sets: one used for training the model, and the other reserved for validation. Data can be assigned to either set by a random selection in suitable software. However, it is difficult to completely avoid bias in the sets. For instance, an outlier could end up in either set and influence the performance. Therefore, we will give the general advice to use cross-validation to avoid bias by assigning data to either set. Instead of training and validating a model on a single split of the dataset, cross-validation involves using each split as a validation set while using the combination of the remaining splits as the training set. In this way, each split is used as validation set exactly once.

It is necessary to allocate all replicates between the training and validation sets in such a way that all measurements pertaining to all samples from a single source are confined within either the training or the validation set.

5.3.1. Performance characteristics

Various validation strategies for likelihood ratio systems on comparison issues have been proposed in the literature, see e.g. (Meuwly et al, 2017), (Ramos et al, 2020) and (Leegwater et al, 2024). These strategies generally focus on evaluating model performance through illustration of LRs observed under H_1 and H_2 and a range of metrics. The performance metrics are numerical values that can be compared to pre-set acceptance criteria during the validation. A brief overview of performance metrics is provided below; for a more detailed discussion, the reader is encouraged to consult the relevant literature.

Density plots and histograms are produced to visualize the distributions associated with the competing hypotheses.

Sensitivity (True Positive Rate) refers to the probability that the LR system predicts a value greater than 1, given that the main proposition is true. This metric assesses the system's ability to correctly identify cases in which the evidence supports the main proposition. The performance of the system is better if this probability is higher.

Specificity (True Negative Rate) refers to the probability that the LR system predicts a value less than 1, given that the alternate proposition is true. This metric assesses the system's ability to correctly identify cases in which the evidence supports the alternate proposition. The performance of the system is better if this probability is higher.

Empirical Cross Entropy (ECE) is based on a 'penalty function' that penalizes misleading LR predictions, discouraging models that produce inaccurate or uncertain likelihood ratios as described by (Ramos and Gonzalez-Rodriguez, 2013). The ECE plot can be visualized as a function of prior odds, as shown in Figure 6-9. The ECE value is preferably lower than that of a non-informative system, i.e. a system that always outputs an LR of 1, which is represented by the black line in the plot. For the ECE values, the performance of the system is better if the values are lower.

There is software in R (package Comparison) or Python (package LIR) that can be used to produce an ECE plot from a set of likelihood ratios.

The Cost of the Log Likelihood Ratio (Cllr) provides a single value that quantifies model performance (Brümmer, 2006). It is a global metric that evaluates both the discriminative power and calibration of an LR system on a validation set and is by definition the ECE value for prior odds of 1. A perfect LR system, which always makes correct predictions with no uncertainty, has a Cllr of zero. A non-informative system, which consistently outputs an LR of one, has a Cllr of one. As such, we expect LR systems to produce Cllr values ranging from zero to one, with lower values indicating better performance. A low Cllr indicates that the system provides reliable and well-calibrated LRs on average, but it is not a measure of the uncertainty specific to the LR calculated for a single sample case. While the absolute value of Cllr is challenging to interpret, it serves as a useful tool for comparing LR systems on the same dataset (Lierop, 2024).

If for *k*=1,2, the term $LR_{\mathrm{k,i}}$ is the *i*th observed LR value given hypothesis $\mathrm{H_k}$, the Cllr is by definition

$$Cllr = \frac{1}{2n_1}\sum_{i=1}^{n_1}\log_2\left(1+\frac{1}{LR_{1,i}}\right)+\frac{1}{2n_2}\sum_{i=1}^{n_2}\log_2\left(1+LR_{2,i}\right) \qquad (5.3.1)$$

Equal Error Rate (EER): The EER represents the error percentage at which the false acceptance rate (FAR) equals the false rejection rate (FRR). A model with better discrimination minimizes the overlap between these two error rates and, as a result, will exhibit a lower EER. The Detection Error Trade-off (DET) plot visually represents the discriminating ability of the LR system. The calculation for the DET plot entails looking at each likelihood ratio across the range of LR values—from the smallest to the largest in the validation dataset—as a threshold. The FAR and FRR are calculated using each LR in that range as the threshold, producing a series of FARs and FRRs corresponding to the different thresholds. The DET plot presents the FAR versus FRR along the series, using a logarithmic scale to better highlight differences across the error range. The EER is identified as the point where the DET curve intersects the main diagonal, which is shown as a black line in Figure 6-15.

Extrapolation boundaries: It is crucial to consider the validity of LRs, especially those estimated using KDE, as the tail values of the distribution are difficult to model stably. If the training data is sparse or lacking in the region of interest, LRs may be grossly overestimated, leading to potentially misleading conclusions. Due to data sparsity in these tails, LR systems often reach a point where it becomes uncertain whether the results are still supported by the available data. In the literature methods are described that lead to the use of lower and upper boundaries for LRs (Vergeer et al, 2016), (Alberink et al, 2026).

Herein, we have opted for the method proposed by (Alberink et al, 2026). This method relies on a mathematical criterion to determine the lower and upper bounds of acceptable LR values. This gives the user of the system information on the range in which it is sensible to use the system. The above is easily illustrated visually, and software is available to do so.

Sensitivity analysis: The performance of an LR system typically depends on modelling choices, parameter settings, and the available training data. When there is doubt about (some of) them or when more insight is desired, a sensitivity analysis can be done. In a sensitivity analysis, it is investigated how the performance of the LR system changes when the choices, settings or data change. It is recommended to involve a statistician when performing a sensitivity analysis.

6. EXAMPLES OF ROUTE TO LR

6.1. Example 1: Cannabis, classification

6.1.1. Introduction

Cannabis is a generic term used for illegal products (like marijuana or hashish) derived from different hemp plants (*cannabis sativa*). As agricultural plant hemp fibre is used to manufacture ropes, textiles and special paper, but also numerous other products such as insulation materials and natural fibre composites. The seeds are used as food and animal feed; the oils extracted from them are also used as food, but also as cosmetics and as medicinal or technical oils.

The intoxicating effect of cannabis is caused by cannabinoids. The most important cannabinoid is delta-9-tetrahydrocannabinol (THC), which is most abundant in the plant's flowers. Two other main cannabinoids are cannabidiol (CBD) and cannabinol (CBN). However, seeds and roots are devoid of cannabinoids (UNODC 2022). Neither absolute THC content nor morphology allows discrimination of fibre cultivars and drug strains of *Cannabis sativa L*. unequivocally. THC values further depend on the developmental stage of the plant and environmental factors such as light and nutrients. Especially if the cannabis plant is very young and without inflorescence, the quantitation of only THC in leaves will not give an appropriate answer.

THC and the nonpsychoactive cannabidiol are both enzymatic products from the same precursor, cannabigerol (CBG). In contrast to the absolute THC content, the relative ratio of cannabinoids is fairly constant throughout the plant’s life and relatively unaffected by external factors (Staginnus 2014). Common cannabis expresses three discrete chemical phenotypes (chemotypes): a THC-predominant type, a CBD predominant type and an intermediate phenotype.

With the DNA testing using a PCR marker linked to a THCA synthase plants rich with THC could be identified. THC-predominant or THC-intermediate phenotype (i.e., a postulated BT/BT or BT/BD genotype) reveal an amplification product, whereas all CBD-predominant individuals (with a postulated BD/BD genotype) do not (Staginnus 2014).

Legal status of cannabis differs in countries: it´s legalised or totally forbidden (use, purchase, sale and possession) or something in between. Thus, in cannabis cases there can be arguments that the found plant is a fibre hemp plant not a drug hemp plant. If required, the determination of the hemp type shall be carried out on samples of cannabis. The question is of type classification.

6.1.2. Methods

Genotype determination

Plant DNA was isolated either from dried or fresh leaves or bracteoles, roots, or seeds, using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer′s instructions after grinding the material under liquid nitrogen.

Chemotype determination

The amounts of different cannabinoids found in the plant material are analysed (THC, CBD, CBN) by gas chromatography mass spectrometry (GC-MS) from methanol extract of the material. The identification is based on the retention time of the compounds and the mass spectrum typical of the substance by means of a spectral library.

The (THC+CBN)/CBD ratio (X) is calculated by the formula

$$X = ([THC] + [CBN]) / [CBD] \quad (6.1.1)$$

where [THC] is the THC response (area) obtained by the above method. Similarly, [CBN] and [CBD] are the responses of those compounds. Typically, the above-mentioned ratio is greater than one for drug hemp (THC+CBN dominant) and less than one for non-drug hemp (CBD dominant) (UNODC, 2022).

The (THC+CBN)/CBD ratio, which is interpreted as drug hemp by THC count, has a large number of both the above-mentioned cannabinoids (Sarma et al, 2020). According to Sarma, hemp types can be divided into three categories based on the main cannabinoid:

THC-dominant:	THC+CBN/CBD ratio ≥ 5; THC ≥ 10 mg/g, CBD ≤ 10 mg/g
THC/CBD-intermediate:	THC+CBN /CBD ratio 0.2-5, THC and CBD ≥ 10 mg/g (10 mg/g = 1 %).
CBD-dominant:	THC+CBN /CBD ratio ≤ 0.2, CBD ≥ 10 mg/g, THC ≤ 10 mg/g

6.1.3. Hypotheses

For this example, we propose the following hypotheses:

H_1*: The recovered hemp sample belongs to category X*
H_2*: The recovered hemp sample belongs to any other category (not X)*

where *X* is one of “THC-dominant”, “THC/CBD-intermediate” and “CBD-dominant”.

6.1.4. Background data

Cannabis samples were collected by NBI – Finland, and FSI - Zurich. The cannabinoids (THC, CBD, CBN) were analysed from these samples by GC-MS. The determined peak areas of THC, CBD, CBN were used to calculate the ratio to assign the samples to categories of THC-dominant, intermediate or CBD-dominant hemp. DNA analysis was performed from the same samples to determine if the plant in question contained a THC producing gene as well. This information was used as “ground truth” to build the LR model for classification of hemp samples.

It turned out that the collected data contained only a few intermediate samples. Thus, additional data of 134 intermediate samples, provided by LKA NRW – Germany, was included. The 134 added samples did not have DNA profiling results but were assigned to the intermediate class based on the chemical analysis. In total, the dataset comprised 332 samples.

6.1.5. Feature-based LR system

The feature used for classification was the ratio between the sum of THC and CBN to CBD, see equation 6.1.1.

The ratios of the 332 samples of hemp were calculated and could be plotted by densities per class.

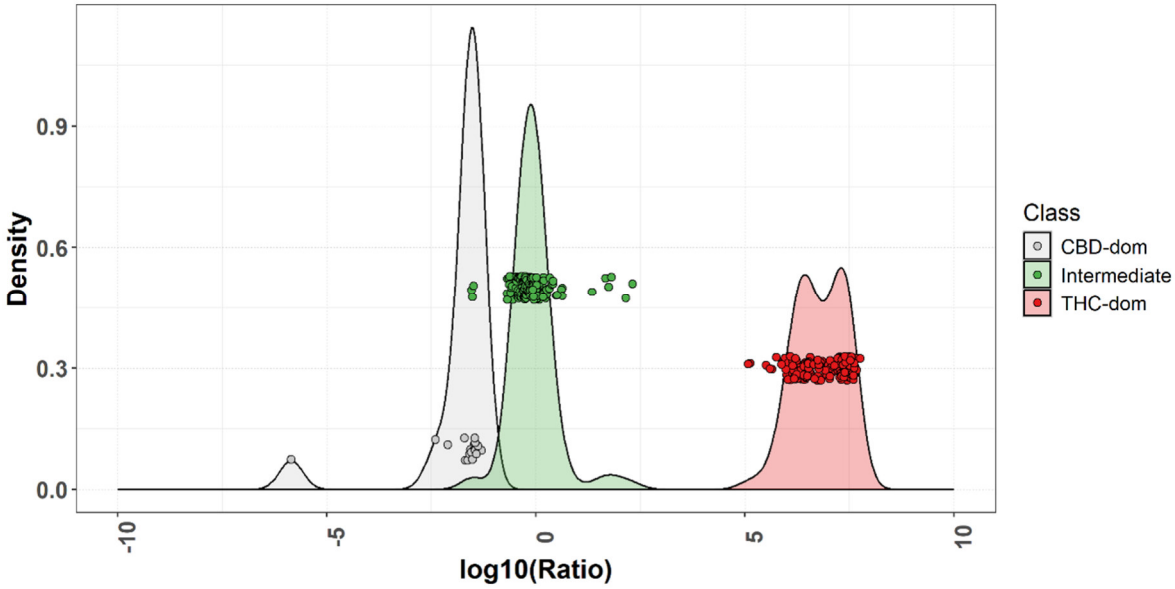

Figure 6-1. Distributions of the three classes are shown simultaneously as density plots and single point values.

As can be seen from Figure 6-1, the THC-dominant samples are clearly separated from the CBD-dominant and intermediate hemps, which are also reasonably (but not completely) separated from each other.

In the current scenario, the factfinder must select one class as the main hypothesis H_1, with the alternative hypothesis formed by the remaining two classes combined. Because the H_1 selection can be defined in three alternative ways, three distinct LR models are required—one for each class serving as H_1. Consequently, the LR system will not evaluate the classes as shown in Figure 6-1; instead, it will assess each class against the combination of the remaining two classes, as illustrated in Figure 6-2.

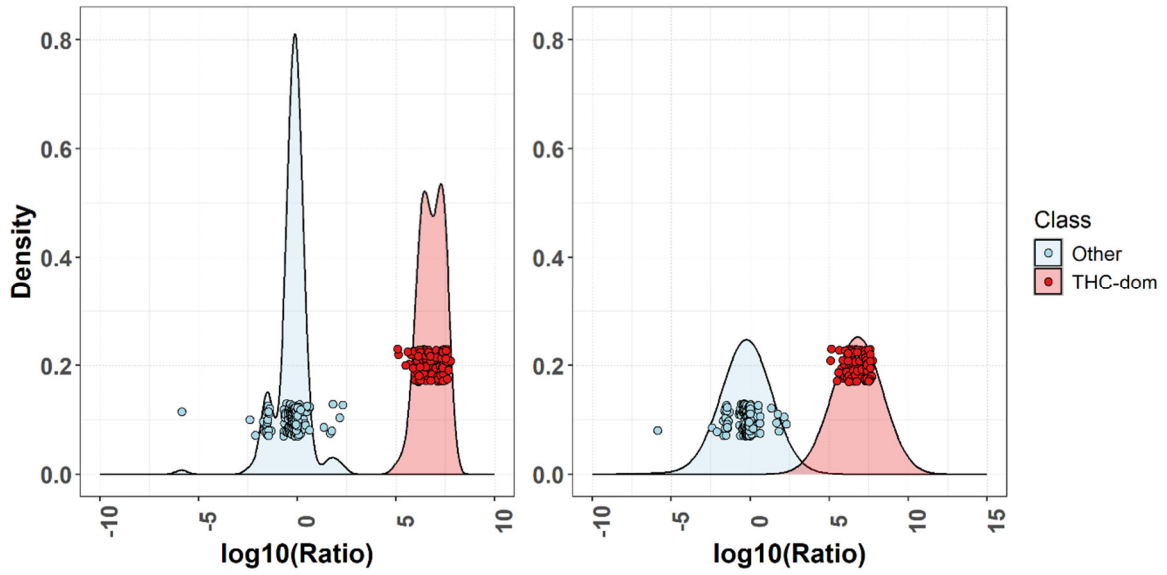

Figure 6-2. The class designated as H1, here THC-dominant, is shown in red, while the other two classes are combined and displayed in blue. The same data using two different bandwidths are shown; left: the default bandwidth and right: the optimized bandwidth.

The likelihood ratio can be calculated as the ratio between the red and blue density curves at any ratio value.

Bandwidth

When estimating density curves, the computation requires a parameter known as the bandwidth (cf. Section 5.2.1), which determines the width of the kernel used around each data point. It controls the smoothness of the density estimate: a small bandwidth produces a curve that closely follows the data, potentially showing random fluctuations, while a large bandwidth generates a smoother curve, see also Figures 5-3 and 5-4 in section 5.2.1. Choosing an appropriate bandwidth is crucial because it balances the trade-off between overfitting and underfitting in the density estimate. The computer automatically selects a default bandwidth, illustrated on the left in Figure 6-2. However, the bandwidth can be optimized, ideally by identifying the bandwidth that minimizes the Cllr. The Cllr-optimized bandwidth (we used the same bandwidth for both distributions) are shown on the right. As evident, the optimized bandwidth is larger, producing smoother density curves. Because all density plots are always normalized to have an area of 1, a wider base in the smoothed, optimized curve results in a lower peak height, as seen on the right in Figure 6-2.

The bandwidth was Cllr-optimized for all three hypotheses pairs (referred to below as models). The extrapolation boundaries, i.e. the range of LRs for which the LR method is considered robust, were determined in accordance with (Alberink et al, 2026).

6.1.6. Likelihood ratio model validation

The three LR systems were validated using *leave-one-out cross validation* (LOOCV). LOOCV for classification means that one sample is taken out while the evaluation model is fitted to the remaining data and then applied to the sample taken out. This is repeated for every sample in the dataset. The result is a set of likelihood ratios for samples where H_1 is true and the set of likelihood ratios for samples where H_2 is true. This constitutes the

grounds for evaluating the system performance (Cllr, EER etc).Figure 6-3 illustrates the distributions of the LRs when using the THC-dominant class as the main hypothesis. The distributions associated with the main hypothesis H_1 (represented in red) has a median value of approximately 4 000.

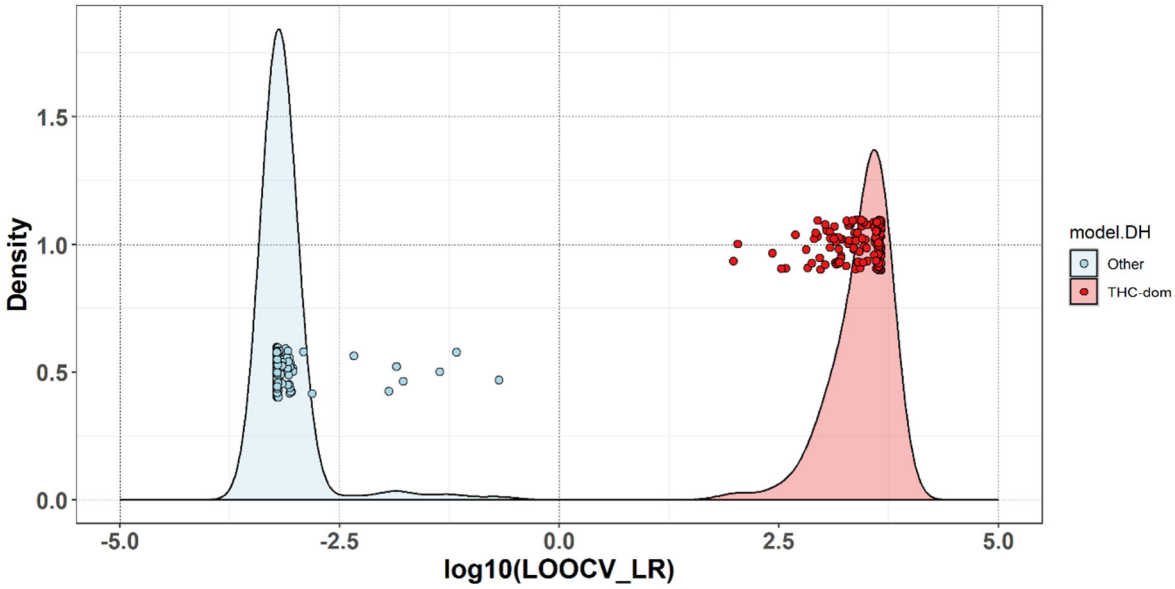

Figure 6-3. The distributions of LRs from the main hypothesis H1 (red), and the alternate hypothesis H2 (blue), for the THC-dominant model.

Table 6-1 shows that the specificity (true negative rate) and sensitivity (true positive rate) are satisfactory for all systems. The table also highlights that all performance metrics are better for the LR-system using the THC-dominant class as H_1, which is not surprising considering the separation shown in Figure 6-1.

Performance metric	THC-dominant	CBD-dominant	Intermediate
Sensitivity	100%	100%	98.0%
Specificity	100%	99.0%	99.5%
Cllr	0.006	0.066	0.054
EER	0	0.48	1.82

Table 6-1. Performance metrics of the three LR systems.

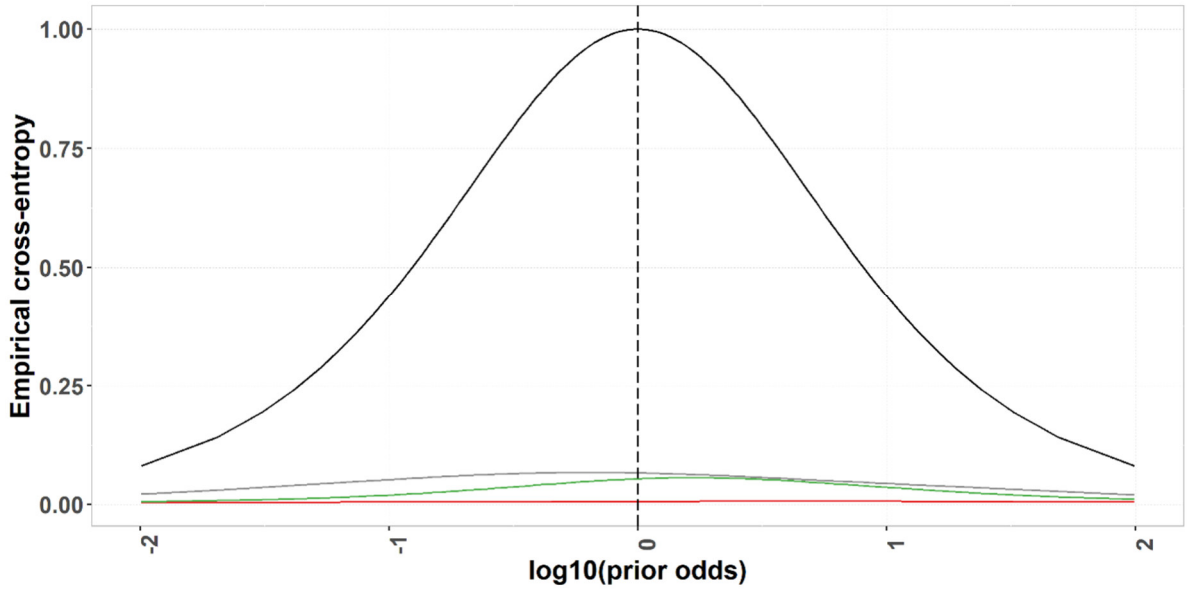

Figure 6-4. The ECE plot of the THC-dominant model (red), the CBD-dominant model (grey), the intermediate model (green), null model (black)

Figure 6-4 shows the ECE-plots of the three models. The plot displays that the model using the THC-dominant class as H_1 (red) delivers the strongest performance, while the other two models exhibit largely comparable results.

The equal error rate

The Equal Error Rate (EER) is zero for the model using the THC-dominant class as H_1 (red), as shown in Table 6-1. This means that the model completely discriminates between all datapoints in this class compared to the other two classes. The EER can be visualized in DET plots (see Figure 6-5), except for the THC-dominant class because its EER is zero.

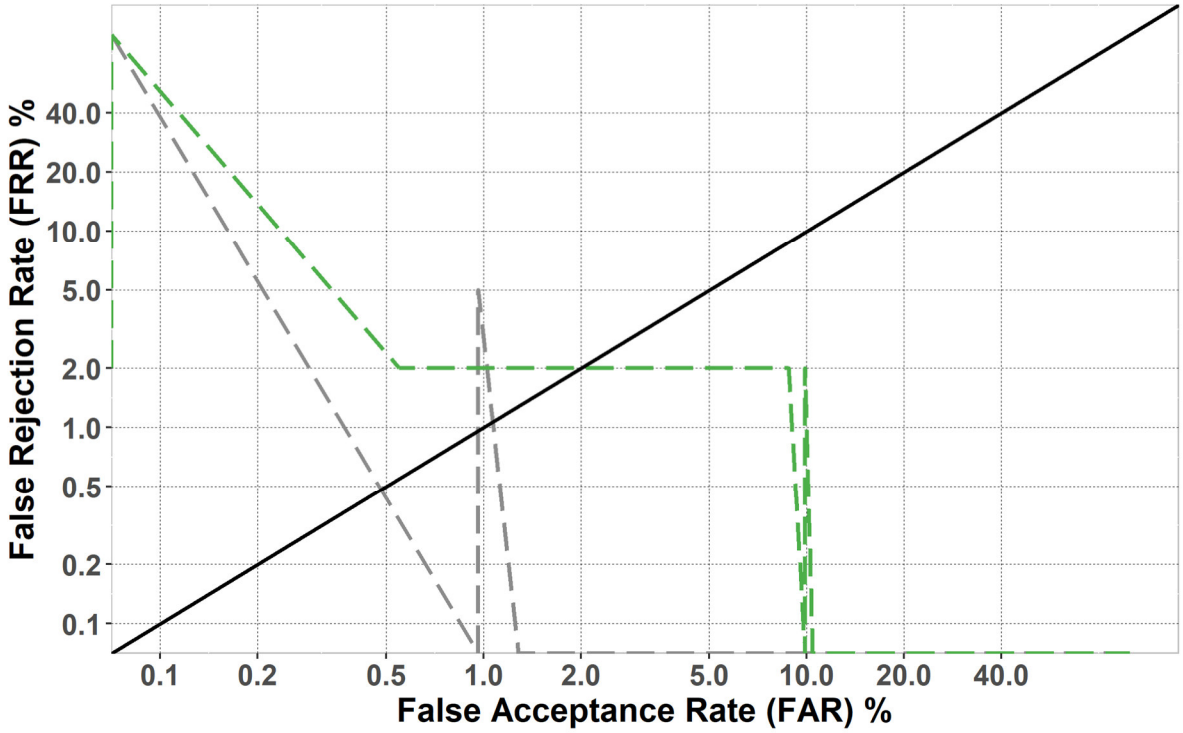

Figure 6-5. The DET plot of the CBD-dominant model (grey), and the intermediate model (green). The THC-dominant model is not visible in the plot since there is no overlap between the distributions. The black diagonal line shows where FAR and FRR are equal.

Conclusion and discussion

The three models perform well with strong performance metrics. Overall given these findings, it is reasonable to state all three models as fit-for purpose. By using these models, the forensic scientist can derive a likelihood ratio (LR) for the relevant hypotheses. As always, caution must be taken for extreme LR values.

6.2. Example 2: Comparison of different LR systems for the differentiation of diesel oil samples

In example 2, two distinct likelihood ratio (LR) systems are described to differentiate between diesel oil samples. The first is score-based, and the second is feature-based. Overall, there will be a scenario in which there are two oil samples, recovered and comparison, which are compared. Hypotheses used are of the following form:

Hypotheses:

H_1*: The recovered oil sample contains the same oil as the reference oil*
H_2*: The recovered oil is a different, unknown oil compared to the reference oil*

The example in the next section (6.3) describes an LR system on a related topic. It differentiates between mineral oil samples (which included other types of oil, next to diesel oil) and can deal with weathered oil samples as well. The LR system is developed using a third approach, based on elicitation of experts.

6.2.1. Background data

The dataset consists of 136 samples sourced from various suppliers. Samples from different sources were utilised to estimate the between-source variability of the oils. Each sample was analysed using gas chromatography–mass spectrometry (GC–MS) with either three replicates (n=89) or six replicates (n=47), to a total of 549 replicates. Replicates are needed to estimate the within-source variability of the oils.

Raw data
The evidential features employed in the analysis consist of ten peak ratios, designated as features R1–R10, see Table 6-2.

Ratio	Numerator compound	Denominator compound
R1	n-C_{17}	Pristane
R2	n-C_{18}	Phytane
R3	Pristane	Phytane
R4	1,3,4-trimethyladamantane	2-ethyladamantane
R5	Bicyclic sesquiterpene 1	Bicyclic sesquiterpene 2
R6	Bicyclic sesquiterpene 5	Bicyclic sesquiterpene 6
R7	Bicyclic sesquiterpene 8	Bicyclic sesquiterpene 9
R8	meta-Octyltoluene	ortho-Octyltoluene
R9	Hexylbenzene	Heptylbenzene
R10	Heptylbenzene	Decylbenzene

Table 6-2. Compounds used for ratios.

The specific ratios were derived from compounds chosen for their ubiquity in diesel oils and their capacity to differentiate between samples (Malmborg et al, 2020). For instance, the feature R1 represents the peak height ratio between the n-alkane C_{17} to the isoprenoid pristane.

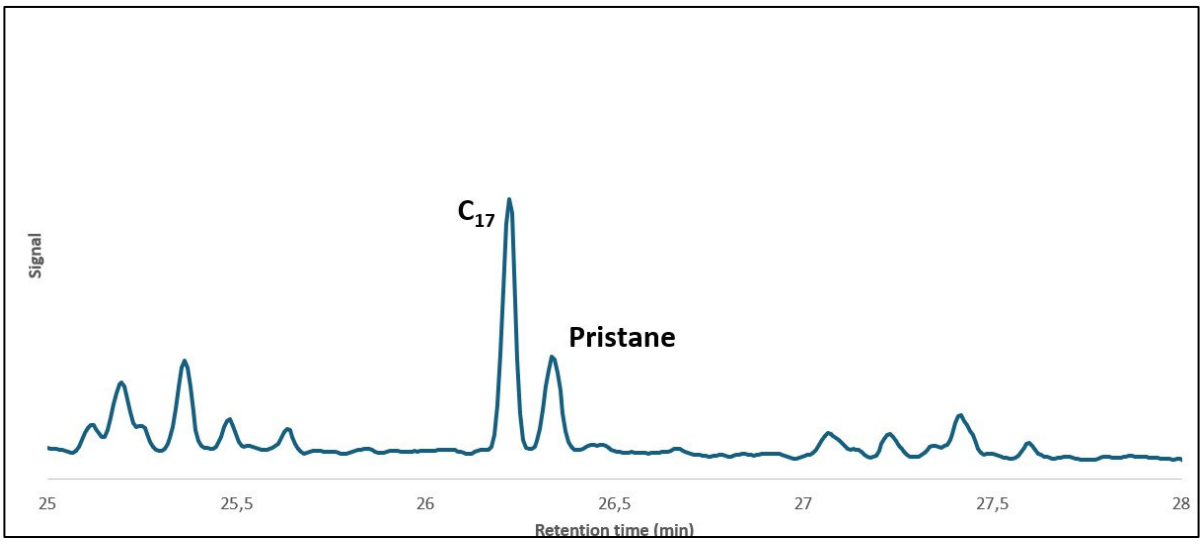

Figure 6-6. The two chromatographic peaks for the feature R1.

The ratio between C_{17} and pristane, as shown in Figure 6-6, is approximately 2.6. However, in a different oil, this ratio (feature) may vary, typically falling within the range of two to four. This variability forms the basis for differentiating between oil samples by applying this single feature.

The ten ratios are not independent, meaning that a relatively high value in one ratio may e.g. increase the likelihood of observing high values in others. As a result, likelihood ratios derived from individual features cannot be directly multiplied to produce an overall likelihood ratio under the stated propositions. Such an approach would overestimate the evidential strength by failing to account for these dependencies. To accurately compute a composite likelihood ratio, these interdependencies must be properly incorporated into the model.

Graphical representations can help elucidate dependencies. For instance, in a bivariate model involving two features, the values across all samples can be visualised in a scatter plot, as shown in Figure 6-7. The dependency shows that increases in R1 are generally associated with corresponding slight increases in R2.

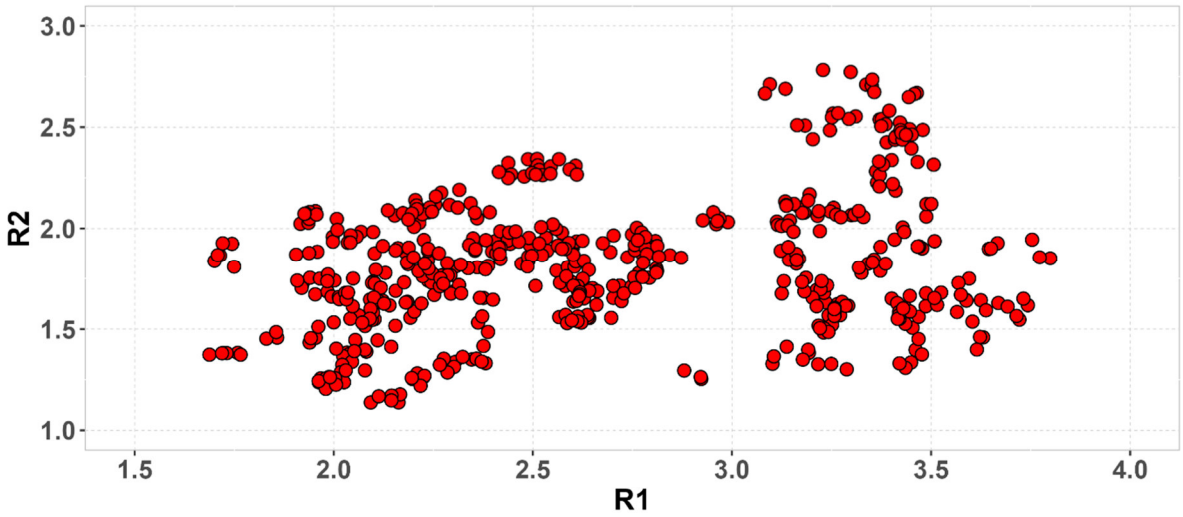

Figure 6-7. The values of the two features, i.e. ratios R1 and R2, shown as a scatter plot.

Data preprocessing

No data preprocessing was necessary.

6.2.2. Score-based LR system

An effective approach to manage dependencies is the use of a score-based likelihood ratio (LR) system. Score-based methods consolidate multiple features into a single metric, or score, which quantifies the similarity or dissimilarity between the items being compared. LRs are then estimated based on these univariate scores. By reducing dimensionality to a single value, score-based methods demonstrate high robustness, particularly when working with limited datasets. However, this simplification comes at the expense of information loss, as the typicality of individual features is not explicitly considered.

Score-based LR systems estimate probability distributions for a vector representing the differences or similarities between two objects. Because this vector is derived from pairwise comparisons, the system explicitly models score distributions for object pairs. For instance, a vector might be constructed by measuring the differences between several chromatographic peak ratios of two objects. The score is subsequently computed from this vector using a selected similarity or distance metric, such as Euclidean distance, Manhattan distance, or cosine similarity.

The score-based LR system for comparing samples of diesel oil was developed as follows. The propositions used are typically that the samples are from a common source (H_1) or from a different source (H_2). In terms of (Bovens et al, 2019), what is answered is a comparison question.

Comparisons of feature vectors given both propositions were performed as follows, mainly following the step-plan of (Leegwater et al, 2024). The 136 samples were split up into 5 equally sized parts, so-called folds. Using these folds, cross-validation was applied on the LR system. This was done as follows.

Given any fold in the cross validation, the feature vectors were either used in the training (4 of the 5 folds) or in the validation set (1 of the 5 folds). Preprocessing was applied by converting the feature vectors by a *z*-score transformation. For each peak height ratio, this means that the average observed value for that specific peak is subtracted from the ratio, after which it is divided by the standard deviation. This transformation is determined based on the training set only and applied to both the training and validation data. For the training and validation data respectively, every combination of two feature vectors is gathered and Euclidean distances between the feature factors are used to quantify similarity. The Euclidean distance is described by

$$dist_{Euclidean}(p, q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \qquad (6.2.1)$$

where *p* and *q* are the samples to be compared, and *i* represents each feature. Kernel Density Estimation is then used to approximate probability distribution functions of comparison scores given both hypotheses, as described in section 5.2.

Likelihood ratio calculation

Following the KDE functions corresponding to the training data, Likelihood ratios were gathered for the validation data. This was done for all folds, and the results were combined. Lower and upper bound values for LR values, see section 5.3.1. on extrapolation boundaries, were 1/598 and 668, respectively.

The resulting LR values given both hypotheses are depicted by histograms in the Figure 6-8, on a logarithmic scale.

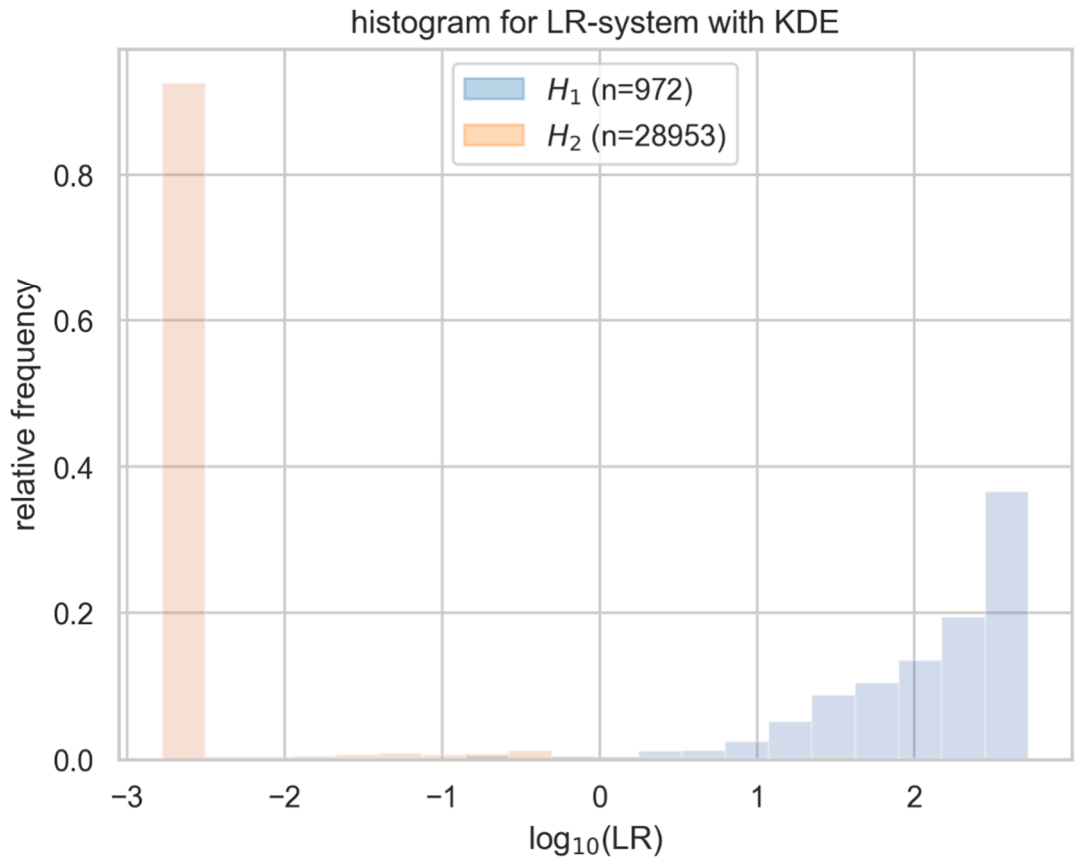

Figure 6-8. Histograms of LR values given both hypotheses, using the score-based LR system.

The performance of the outcomes of the cross-validation was tested by an ECE plot, see Figure 6-9. In the plot the red curve denotes the ECE plot of the original LRs that were determined, and the green the LRs after they have been transformed by the Pool Adjacent Violators algorithm.

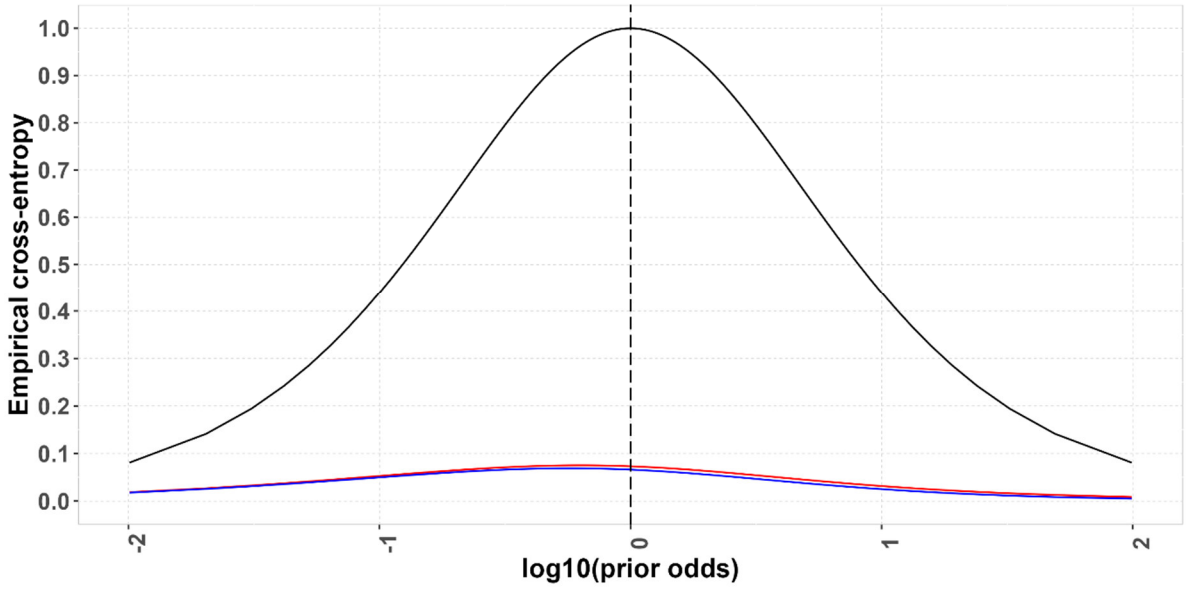

Figure 6-9. ECE plot of the original LRs of the cross-validation, in red, and LRs after they have been transformed by the PAV algorithm, in blue.

The system does have a rather good performance with not an excessive amount of misleading evidence. The Cllr value (ECE for prior logarithmic odds of 0) is 0.074.

If in a case for example the observed similarity score is 1.5, this corresponds to a likelihood ratio of $f(s=1.5 \mid H_1)/f(s=1.5 \mid H_2)=0.273/0.0267\approx10$.

6.2.3. Feature-based LR system

Dependencies between ratios can also be addressed using a multivariate feature-based likelihood ratio (LR) model. Feature-based LR systems consider the probability distributions of direct measurements of objects, where a feature might represent the refractive index of glass, a chromatographic peak area ratio, or the weight of a tablet.

Unlike score-based systems, feature-based LR systems require the modelling of feature distributions under both competing hypotheses, rather than just the similarity or distance between two items. While higher-dimensional relationships are challenging to visualise graphically, they can be effectively managed using multivariate modelling techniques.

The primary strengths of a multivariate feature-based model lie in its ability to preserve information and account for the typicality of individual features. However, as more ratios are incorporated, the dimensionality of the multivariate continuous feature space increases significantly. This expansion necessitates a larger dataset to accurately estimate the covariance matrix, which defines the shape of the density function. As a result, the model becomes more complex.

Dimensionality reduction

When dealing with ten features, dimensionality reduction is often necessary. Techniques such as Principal Component Analysis (PCA) can be employed to identify a smaller set of uncorrelated components that capture the majority of the variance in the data. Alternatively, the most discriminative features can be selected based on their performance in distinguishing between samples. These methods simplify the model while retaining critical information. Without proper management, feature-based systems risk producing erroneous conclusions. Feature-based systems generally retain superior discriminating power because they avoid the loss of information inherent in reducing multidimensional data to a single scalar score (Bolck et al, 2017; Ishihara and Carne, 2022). However, they require larger datasets to accurately estimate covariance matrices and are more sensitive to specification errors in the distribution model.

The ten peak ratios R1-R10 were further analysed by Malmborg and Nordgaard (2016) to optimise a feature-based LR system. Their study identified that three ratios R1, R5 and R10, together formed the most effective feature-based method. Using only two ratios reduced the system's discriminative power, while adding a fourth ratio provided only marginal improvement in discrimination. However, this added complexity and introduced some extreme likelihood ratio values, indicating that the dataset was too limited to effectively support a fourth ratio. Consequently, the feature-based model using the three identified ratios was constructed here.

Calculation and results: The within-source distribution

The 549 replicates were utilised to calculate the mean values for each of the 136 samples, i.e. one mean for each ratio and sample. The mean of the relevant sample was later subtracted from each replicate's ratio value, producing 549 residuals per ratio. The collection of residuals represents the distribution of analytical within-sample variation. The distribution for ratio R1 is depicted in Figure 6-10 and 6-11.

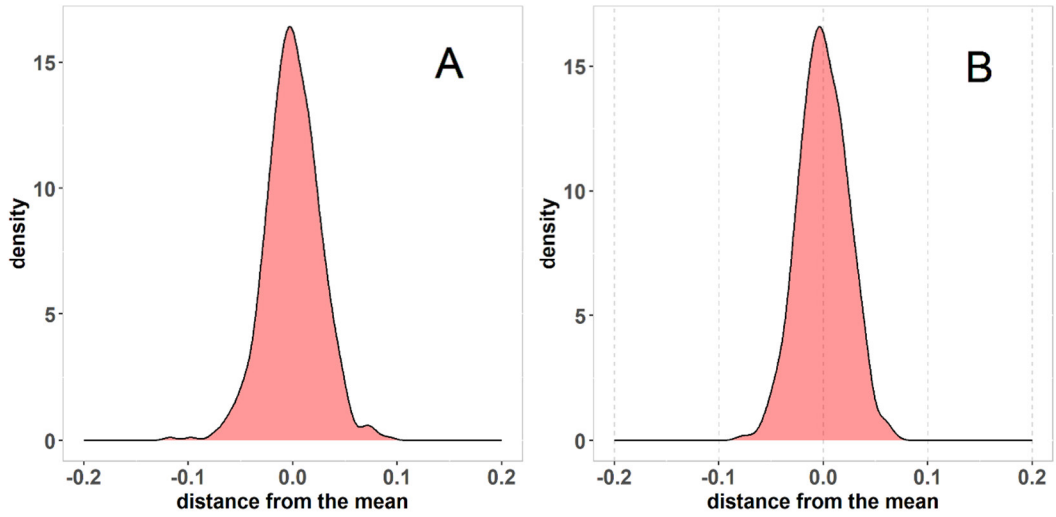

Figure 6-10. The distribution of all ratio R1 residuals A) before normalisation B) after normalisation.

The statistical framework employed for the feature-based model assumes a Gaussian distribution to capture variation within a single source. In other words, the replicates of each sample should be approximately normally distributed around the mean value for each ratio. It is essential to verify the normality assumption early in the statistical analysis.

It is reasonable to assume that peak heights themselves are normally distributed across sample replicates. However, the ratio between two normally distributed variables, such as the peak heights of n-C_{17} and pristane, typically exhibits more pronounced tails (leptokurtic distribution) than a standard normal distribution. Indeed, normality tests revealed that for all three peak ratios R1, R5 and R10, the within-source variation did not adhere to the assumption of normality.

The ratios R1, R5 and R10 were therefore subjected to normalisation, in accordance with Malmborg and Nordgaard (2021). The transformation of heavy-tailed data to approximate normality is not a straightforward process, and we will not delve into the details here. After applying this transformation, the three ratios conformed to an approximately normal distribution. The normalized residuals for ratio R1 are shown in Figure 6-10, right.

The impact of deviating from the normality assumption can be assessed by evaluating the performance of the LR system. When comparing the results using the normalized data to those using the non-normalized data, we observed an improvement, although it was not dramatic. In other words, the data in Figure 6-10 A are not far from normal, and omitting the normalisation step would not have led to significantly erroneous results in this case. However, it is crucial to carefully manage and verify any assumptions made within the LR system.

Calculation and results: The between-source distribution

The between-source distribution is typically not normal. For instance, the ratio of n-C_{17} to pristane in oils may vary depending on the crude oil's origin. Oils from the North Sea may cluster around one ratio value, while Russian oils may cluster around a different value, resulting in a bimodal distribution with two distinct peaks. Other factors, such as differences in refinery techniques, can further contribute to the variation between oils. Consequently, the overall distribution becomes uneven and multi-modal, not conforming to any particular distribution pattern (see Figure 6-11, right).

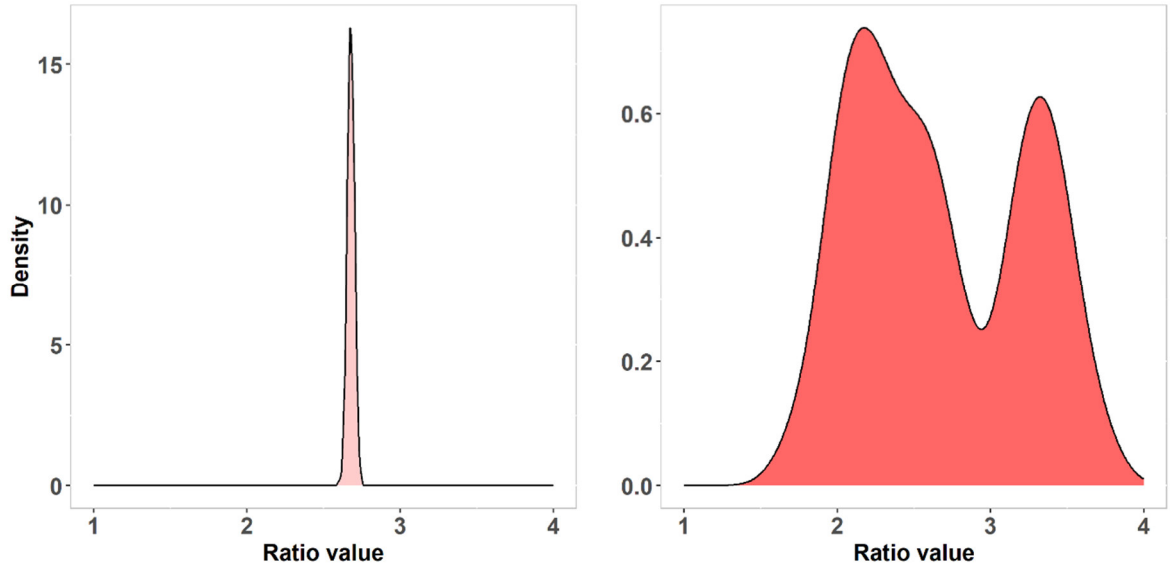

Figure 6-11. The distributions of R1. The x-axis is kept consistent for both figures but note the difference in the y-axis. Left) The within-source variation is normal; the mean is from a random sample, and the standard deviation is the same as in 6-10. Right) The KDE distribution for the between-source variation.

Gaussian kernel density estimation can be used to estimate the probability density function of the variation between sources. In this approach, each sample value in the distribution is replaced by a normal distribution, where the original sample value serves as the mean, and the bandwidth (see section 5.2.1) defines the standard deviation.

In the current study, the bandwidth was determined according to the method outlined by Silverman (1998).

The bandwidth can also be optimised by adjusting it based on the performance of the likelihood ratio (LR) system, i.e., iteratively altering the bandwidth to achieve the best system performance.

Likelihood ratio calculation

The likelihood ratio (LR) is calculated as the ratio between the probability densities of the within- and between-source distributions at the point of interest. This point lies in the three-dimensional space spanned by R1, R5 and R10. For visibility, Figure 6-12 illustrates the calculation using only one dimension, R1.

The light red curve, which is normally distributed, represents the within-source variation, with the mean positioned at the comparison sample mean. The probability density under

the same-source proposition at the point of interest is indicated by a black dot on the light red curve. This dot sits at the apex of the curve only if the comparison and recovered samples have exactly the same ratio value.

The dark red curve represents the between-source variation of ratio R1. As with the within-source variation, the point of interest is marked by a black dot. This illustrates the importance of considering the typicality of the ratio value.

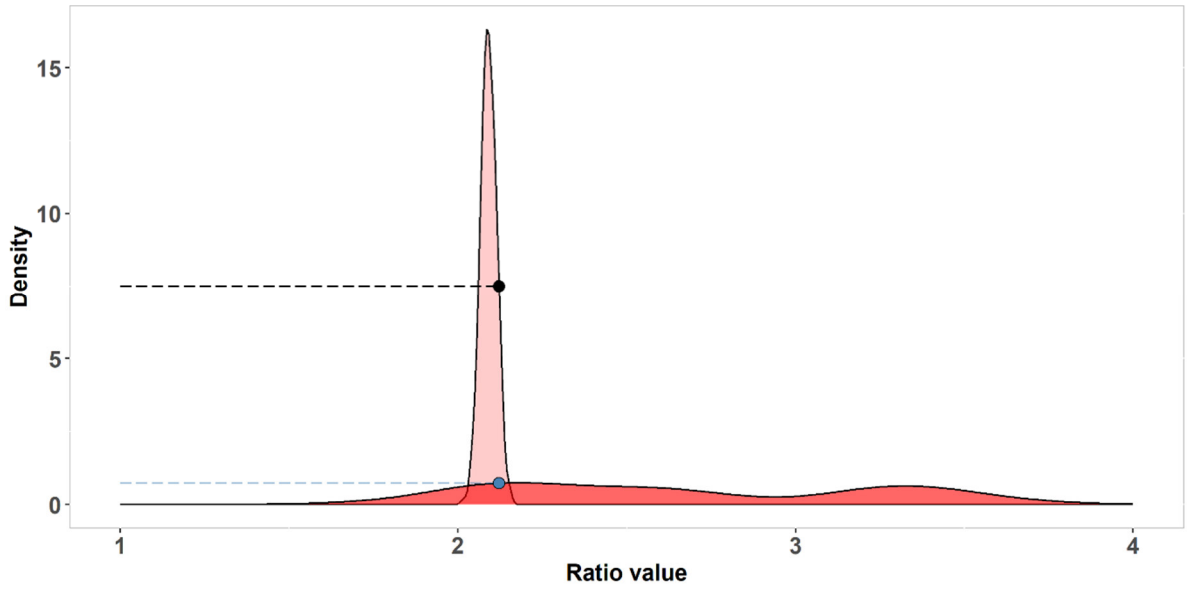

Figure 6-12. The probability densities for ratio R1 under the same-source proposition (light red), and different-source proposition (dark red) at the point of interest.

Finally, the likelihood ratio is calculated as the ratio between the two probability densities, in the Figure 6-12 example LR = 16.45 / 0.72 = 23. It is important to note that this example is based on only the R1 feature. When all three features are used the LR will increase substantially, assuming that R5 and R10 also match well.

The extrapolation boundaries, i.e. the range of LRs for which the LR method is considered robust, were determined in accordance with (Alberink et al, 2026) to 1/1939 (lower) and 9494 (higher).

6.2.4. Likelihood ratio model validation: Performance of the score-based and feature-based systems

Figure 6-13 illustrates the distributions of the likelihood ratios (LRs) for the two systems that were described. The distributions associated with the main hypothesis H_1 (represented in red) reveal median values of approximately 3,200 for the feature-based model and 180 for the score-based model.

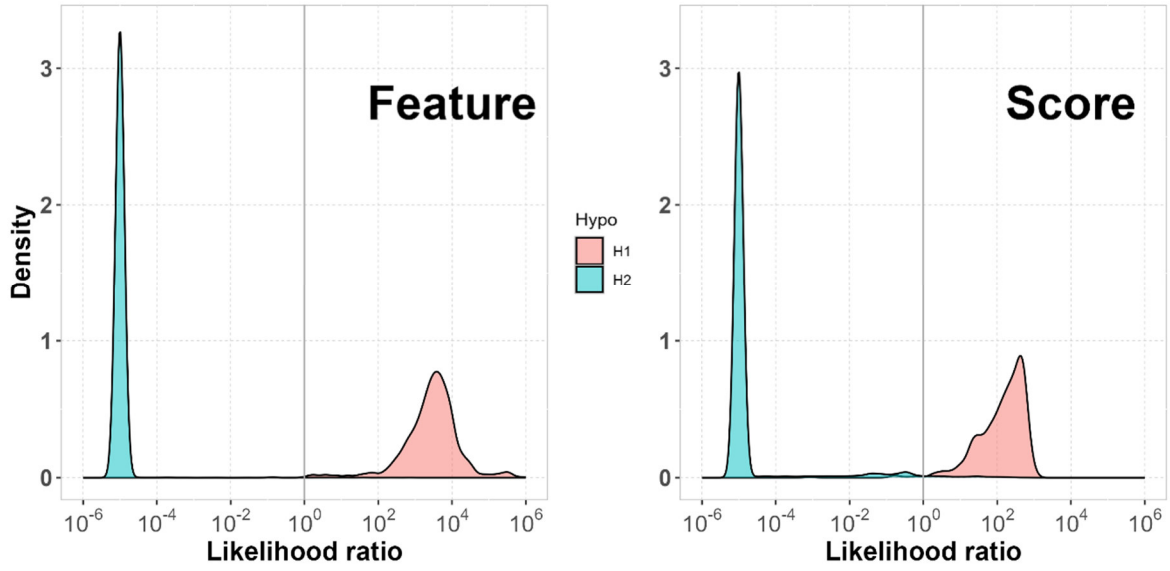

Figure 6-13. The distribution of LRs obtained when H1 is true (red) and when H2 is true (blue) for the feature-, and score-based models. For cases where the alternate hypothesis H_2 (blue) is true, lower LRs have been truncated at 10^{-5}.

The red distributions also reveal considerable variation in the maximum LR values across the models. The score-based model has a maximum LR of about 1 000, while the feature-based model (C) exhibits a significantly higher maximum LR of approximately 2.2×10^6. The extreme LR value observed in the feature-based model highlights the challenge of limited data availability in the tails of the multidimensional distribution. Such outliers should be interpreted with caution, and it is recommended to apply restrictions to the calculated LR values to avoid overestimating the likelihood ratios (Vergeer et al, 2016, Alberink et al, 2026). The separation of the distributions is generally satisfactory for both the feature-based model and the score-based model, although some overlapping LR values are observed.

Performance metric	**Score-based model**	**Feature-based model**
Sensitivity	98.9%	99.9%
Specificity	98.1%	99.9%
Cllr	0.074	0.010
EER	1.54	0.11

Table 6-3. Performance metrics of the compared LR system.

Table 6-3 shows that the specificity (true negative rate) and sensitivity (true positive rate) are satisfactory for both systems. The table also highlights that the performance metrics Cllr and EER are lower for the feature-based system compared to the score-based system. This result is expected, as the dimensionality reduction in the score-based systems leads to a greater loss of information.

Figure 6-14 presents the ECE plots for the models.

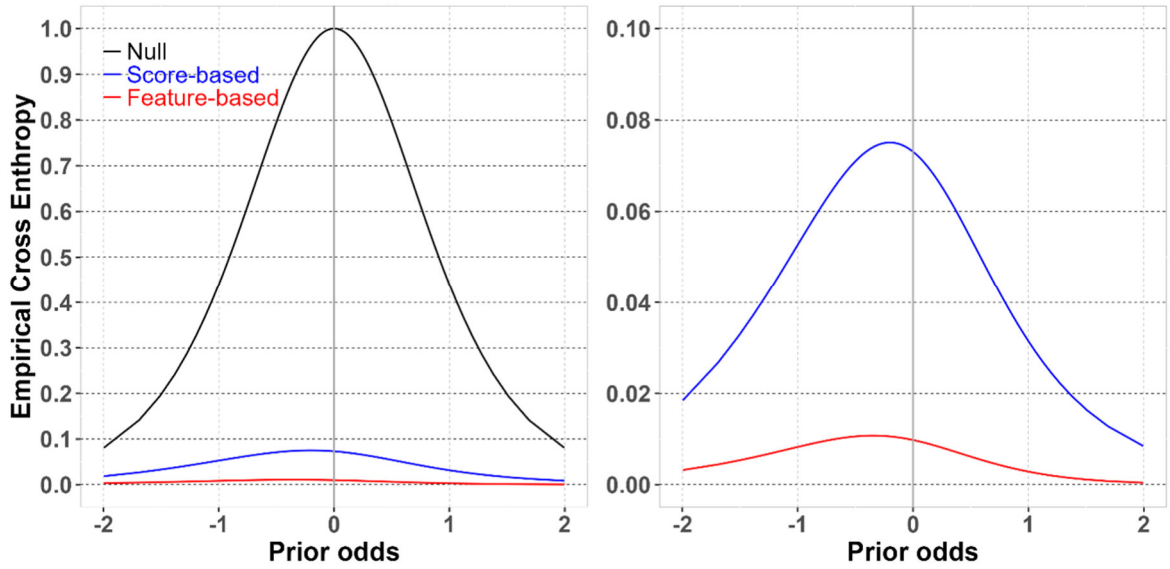

Figure 6-14. The ECE plot of the score-based model (blue), and the feature-based model (red). The curves are zoomed in on the right.

The Equal Error Rate

The Equal Error Rate (EER) follows a similar pattern to the Cllr, with the feature-based model outperforming the score-based model, as shown in Table 6-3. This difference in performance is further highlighted in the DET plots, Figure 6-15, where the best-performing model minimises the area under the curve in the lower-left corner.

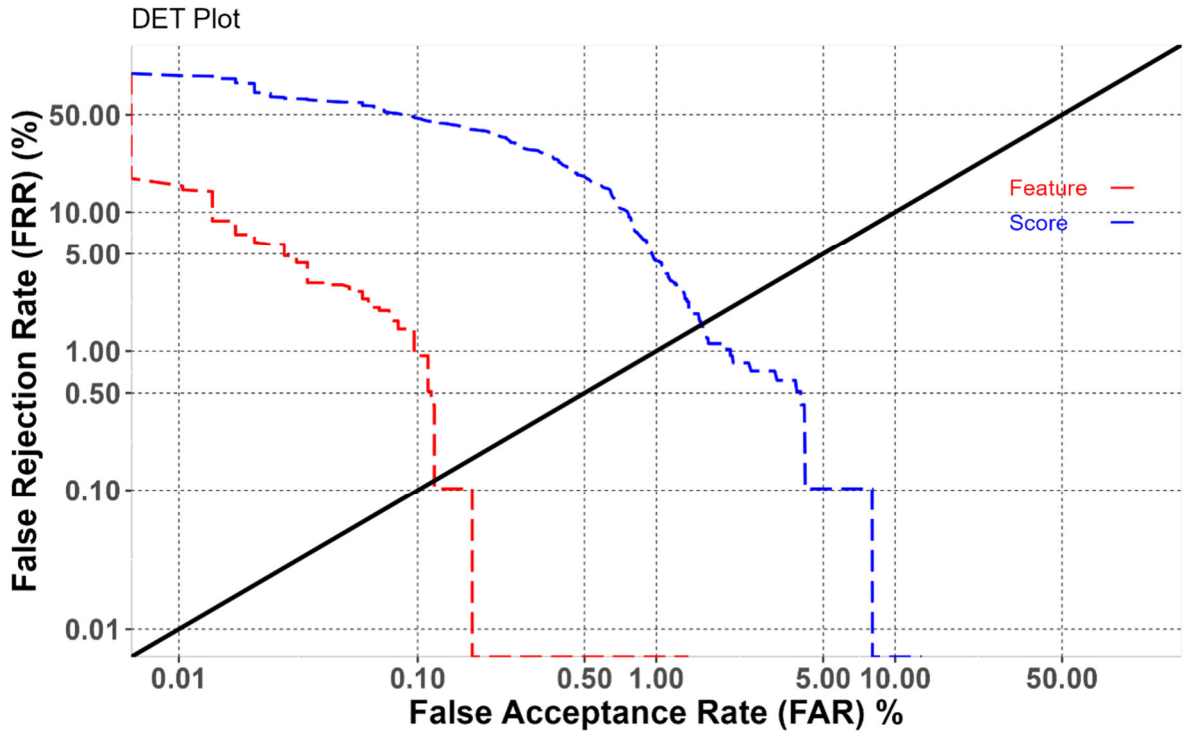

Figure 6-15. The DET plot of the score-based model (blue), and the feature-based model (red).

Conclusion and discussion

Overall given these findings, the feature-based model would likely be the preferred option if a single method were to be chosen from the two evaluated. However, it is crucial to interpret extreme LRs with caution. In such cases, implementing restrictive measures is recommended to minimise potential inaccuracies, thereby ensuring that the results are both reliable and interpretable.

6.2.5. Case Example

An oil spill at sea was detected by aerial surveillance. By analysing the routes of nearby vessels, two ships (A and B) emerged as suspected sources. The coast guard boarded the ships and sampled oil from the respective bunker tanks. Both spill and reference samples were sent to the forensic laboratory. The scientist postulated two competing hypotheses, i.e. the two oils are the same vs the two oils are different.

The samples were analysed. The oil types were all diesel oil, and the weathering of the spilled oil was limited. The comparison was thus in scope of the current LR-systems. The values of the ten peak ratios of these two samples are given in Table 6-4.

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
Spill	2.13	3.50	2.61	0.96	1.64	1.51	1.18	2.37	1.26	4.32
Ref. A	2.07	3.24	2.54	0.97	1.65	1.49	1.20	2.34	1.28	4.43
Ref. B	2.85	2.61	2.06	0.64	1.86	1.03	1.73	2.20	1.39	5.61

Table 6-4. Compound ratio values, i.e. the features, of the case example.

For both approaches, the same procedure was followed to develop the LR-systems as before but now using all available samples as training set. It is assumed that the performances of these systems are at least as good as was determined during cross-validation.

Score-based model

The training data was first pre-processed (z-transformation), then scores for all possible pairs of measurements were calculated, resulting in 972 H_1-scores and 149454 H_2-scores. Based on these two sets of scores, the LR-system was determined and applied to the case data. The score for the measurements on the spill and on reference A is 0.71, leading to a likelihood ratio of 366. This value falls between the extrapolation boundaries of this LR-system (1/1762 and 689). The score for the measurements on the spill and on reference B is 4.7, leading to a likelihood ratio that is equal to the lower extrapolation bound of 1/1762.

Feature-based model

After analysis, the data was pre-processed (normalised) for the feature-based approach, and we then apply the three-dimensional between-source model. It is a space in three dimensions, and we plot only one dimension, ratio R1. In the next step, the three ratio values from the reference sample are used to locate the same-source density curve (light red curve) in the three-dimensional space; but again, plotted using R1.

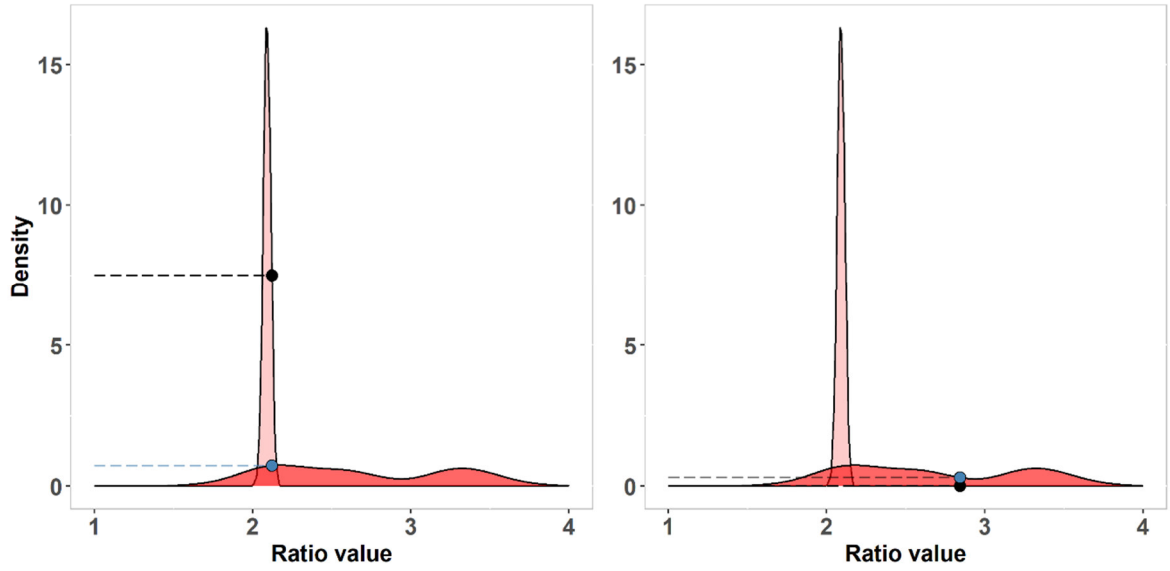

Figure 6-16. Ratio R1 depicted for the spill sample against the oil from ship A (left) and ship B (right). The black dot and line represent the same source probability density, and the blue dot and line represent the different source probability density.

Finally, the likelihood ratio is obtained by dividing the two probability densities in the three-dimensional space, yielding a value of 3289 for the oil sample from ship A and 2.6 x 10^{-308} for the oil sample from ship B. Because the value for ship A lies within the validated model's extrapolation limits (1/1939 to 9494), the likelihood ratio is reported as 3289. For ship B, the calculated value falls far below the lower boundary, so the reported likelihood ratio is set to 1/1939.

6.3. Example 3 Obtaining LRs through expert elicitation

As described in section 5.2.2, LRs may be obtained through expert elicitation. This approach is especially useful when obtaining empirical data is challenging due to cost or complexity of the problem, but experts still have substantial special knowledge e.g. through practical and theoretical understanding of chemistry or physical processes. Here we demonstrate this approach in case of oil spill comparisons. This approach is widely applicable to different forensic disciplines, but successful application requires interaction between statisticians and forensic examiners to produce a useful and accurate model.

The basic question considered in this example is the typical case where the courts or investigators are interested in whether two samples contain the same oil. One might wish to directly ask about the source of the oil, but defining a source in case of oil spills can be problematic and introduces a multitude of variables that are difficult to control, including among others the actual rarity of the oil and exact location of where the spill originated. Thus, the problem here is considered at a lower level with the propositions

H_1*: The recovered oil sample contains the same oil as the reference oil*
H_2*: The recovered oil is a different, unknown oil compared to the reference oil*

The analytical data obtained includes FID-chromatograms mainly for visual inspection and GC-MS spectra with 50 different compounds and their integrated peak areas. As before, the task is then to obtain the probabilities of the observations given each proposition. While the analytical method provides numerical data, it is notoriously hard to obtain sizeable and

representative databases of measurements of known same and different oil sample pairs. This hinders the development of an empirical LR system. Thus, a method based on expert elicitation was chosen to aid experts to consistently and reproducibly assign likelihood ratios to comparisons.

The elicitation in this case proceeds in the following steps:

- Determination of variables necessary to adequately describe all relevant observations.
- Establishing statistical dependencies between the variables.
- Assigning conditional probability tables for each observational variable.

In each step, the experts provide information to the statistician responsible for producing the final model through several workshops. In addition, the experts are required to assign conclusions on a likelihood ratio scale to different example cases that are used as the goal for the final statistical model.

Determination of the necessary variables

In this step, the experts are tasked to consider the data used in examination as a whole and try to divide it into useful variables. A natural starting point is found to be separation of the data by the instrument that produced it, namely FID and GC-MS. Furthermore, it is clear that the experts are mainly considering the chemical similarity of the samples and as such it is decided that the variables must represent degrees of similarity. Additionally, it is observed that certain compounds are shared over these methods, and their effect is separated into their own variable. With the interest of keeping the model as simple as possible, the relevant data is grouped under as few variables as possible while leaving the possibility of expanding the model at a later stage open. In the end the model has three observational variables: MS, for mass spectral observations, FID for observations in the chromatogram and PrPh for pristanes and phytanes which are compounds partly measured by both methods providing significant overlap.

In addition to these observational variables, two background variables – related to the quality of the information in the sample material in either method – are introduced. These variables are meant to capture influence of effects such as weathering or other degradation in the samples resulting in lack of information. In practice, this is relevant when due to degradation some compounds have disappeared from the samples. This can result in a situation where whatever is still observed provides high degree of similarity between the samples, but as information is missing, it cannot be valued as strongly as if the samples had been in perfect condition. That is, one must not draw as strong conclusions from partial data.

Finally, a variable representing the proposition or hypothesis is introduced to enable the model to account for different ground truth states.

For the observational variables five possible states representing degrees of similarity are determined ranging from –2 to +2, with –2 indicating extreme dissimilarity and +2 extreme similarity. For these states, the experts determine several criteria leading to a scoring system that allows the experts to assign a variable state based on the analytical findings. For the background quality variables three possible states are given ranging from 0 to –2,

with 0 meaning good quality (i.e. highly retained information) and –2 indicating high level of degradation in the information from the sample.

Determination of dependencies between variables

A statistical dependency between two variables means that knowledge of the state of one variable provides information on the state of the other variable. That is, knowing the state of one variable allows one to make predictions on the other variable. This is a critical concept as ignoring the dependence structure between variables in a statistical model can lead to over- or underestimating the influence of observations.

In the case of oil comparison, naturally all observational variables depend on the variable representing the proposition: if H_1 is assumed to be true, it will definitely be expected for the observations to indicate higher similarities than if H_2 is true. Furthermore, the observational variables FID and MS naturally depend on their corresponding quality variables as the states of the quality variables directly affect the probabilities of observations: if the sample is degraded, more precise observations allowing for the strongest (dis-)similarities are less likely. Besides these simple dependencies, another dependency is determined. While the mass spectral analysis provides much more information than the FID analysis, knowledge of the results of the FID analysis does inform the experts' expectations for the results of MS analysis: high similarity indicated in FID results makes it more likely that MS results also support high similarity and vice versa. Therefore, there is a dependency between variables FID and MS.

The results of the above are interpreted using a Bayesian network. We will not go into the details of this network since it is out of scope of the guideline, but the network is illustrated in Figure 6-17. In the figure, each variable is represented by an ellipse, and the arrows indicate the dependence structure. For instance, an arrow from H to PrPh indicates that the variable PrPh is statistically dependent on the variable H. To fully reproduce the network, one would also require the so-called conditional probability tables for each variable (see e.g. (Taroni et al, 2014) for more on Bayesian networks). The value of the Bayesian network is that it allows a concise visual summary of the model.

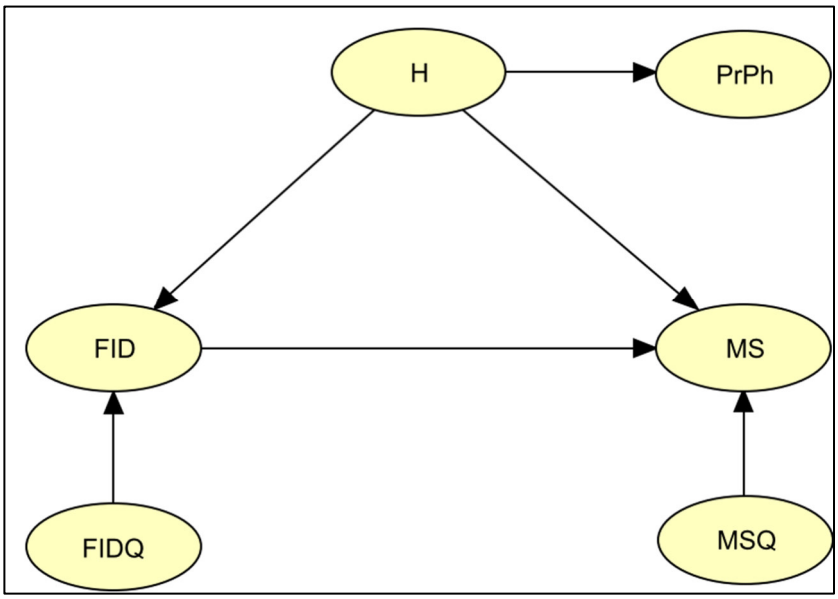

Figure 6-17. Illustration of a Bayesian network for oil comparisons. The arrows describe dependencies between variables which in turn are represented by the ellipses.

Assigning conditional probability tables to the observational variables and using the model to arrive at an LR

This essentially involves defining a probability distribution over all possible states of a variable under each possible configuration of other variables it is dependent on. This is done by means of elicitation from experts. As the aim of the model is to represent expert opinions as closely as possible, the training data for this purpose is given by expert assignments of likelihood ratios on a scale to example cases corresponding to real cases and artificial configurations.

Once the conditional distributions for each variable are obtained by fitting the model to training data, the Bayesian network can be used to obtain the likelihood ratios in a straightforward way. Effectively, after performing the analyses, the experts need to map their observations to variable states as defined in the Bayesian network. This allows computation of the probability of the observations under both competing propositions from which the LR is obtained as their ratio.

As a simple practical example, suppose there has been an oil spill, and two ships have been identified as possible sources for it. Samples from both the spill and the ships have been obtained and delivered to the laboratory for analysis. Figure 6-18 shows the FID chromatograms obtained for all three samples.

Comparing the spill to the first suspected source, the expert determines that there are no substantial issues related to the sample material quality for either the MS or FID analyses and the related background variables are set to state 0. The expert observes clear differences in results from the FID and MS analyses between the sample and as such the corresponding variable states are set to –2. For the overlapping compounds represented by one of the variables, PrPh, there are differences, but they are less obvious, and the state is assigned as –1. Inserting these values into the Bayesian network yields a likelihood ratio in the order of 1 / 100 000, providing strong support for the proposition that the oil spill is from a different source than the suspected first source.

Comparing the spill sample to the second suspected sample, the expert determines again the quality of the sample material to be without issue. However, the results from MS and FID analyses yield some similarities between the samples and the corresponding variable states are set as 1. For the variable PrPh, the similarities are clear, and its state is set as 2. Given these variable states, the Bayesian network yields a likelihood ratio in the order of 1 000 providing moderately strong support that the spill oil and the suspected source oil are originally the same oil.

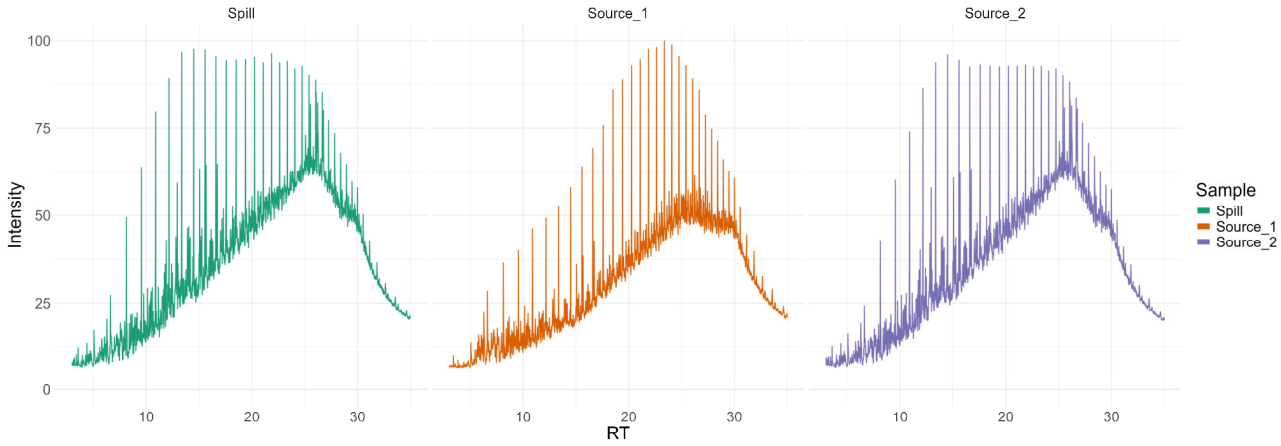

Figure 6-18. FID chromatograms of the spill sample and the two possible sources.

6.4. Example 4: Clustering of several and comparison of two heroin samples

6.4.1. Introduction:

Example 4 describes the clustering of heroin by hierarchical cluster analysis to find chemical links between heroin samples and second an LR approach to compare two samples of heroin.

Comparisons of drug samples from different seizures are frequently asked by law enforcement and court to forensic laboratories. The idea behind this is that if there are relations of characteristic properties out of the material that lead to the conclusion that the material is from a common source, then a more serious impact on the accused person or crime group will be assumed. These properties of the material can be elucidated by the characterisation of chemical markers followed by a comparative analysis. This way of analysis is named profiling. Nowadays it is an important tool to complement routine law enforcement and forensic analysis. Drug characterisation studies have shown that it is possible to identify links between drug samples, to classify material from different seizures into groups of related samples and to attribute their common source or origin. Such information can be used for evidential (judicial, court) purposes or it can be used as a source of intelligence to group the samples that may have a common origin or history (UNODC, 2023).

The empirical basis for Heroin Profiling presented here is a collection of seized heroin samples since 1979 at the Federal Criminal Police Office (BKA), Germany, as well as authentic heroin samples from different production regions like Southwest Asia, Southeast Asia, Mexico and South America. The collection also includes consignments from European and non-European countries as part of international mutual legal assistance.

The heroin samples are systematically collected and examined in a standardised examination procedure, taking into account the external characteristics, the overall composition, the form of preparation and the chemical traces produced during manufacturing process.

The data were stored in a database and evaluated with regard to the following questions:

- Linking of heroin preparations from different seizures
  (different preparations are traced back to a common bulk amount)
- Differentiation of heroin preparations with the same external characteristics
  (preparations with the same appearance may still originate from different bulk amounts)
- Linking heroin preparations on the basis of cutting agents
- Determination of the region of origin
  (Southeast Asia, Southwest Asia, Mexico, South America, et cetera)
- Drawing conclusions about manufacturing processes (like synthetic routes)
- Observation of the illegal market with regard to discover or prove trafficking and distribution routes
- Detecting changes in manufacturing process of heroin (like synthetic routes)
- Detecting contamination with harmful substances and pathogens

6.4.2. Example 4a: Heroin Clustering

A common situation is that law enforcement officers seize illicit drugs during an ongoing case file at different sites. At a certain point of the investigation the suspicion arises that the seized drug material may come from one bulk amount and is divided during the distribution in several smaller amounts. Sometimes independent investigations lead to the assumption that several individual drug dealers in different cities or regions may be connected to the same bulk material. Chemical characterisation followed by clustering of gained data can help to answer such type of questions.

If material from individual portions originate from one bulk amount it will show the same chemical characteristics and represent a cluster. It is also common that several individual portions of material can form multiple clusters. Each of these clusters represent in this case a different bulk amount of material. If this clustering is done by a trained scientist, it is called *expert assessment*. The application of an explorative, multivariate method like hierarchical cluster analysis can speed up the process of clustering by using selected parameters of chemical investigations to find clusters of material showing the same chemical characteristics.

A trained and validated model is applied that is based on datasets
of clusters found by expert assessment as ground truth.

A typical case example for this is as follows:

At a house search in city A, approx. 6 kg of heroin, approx. 2 kg of additives (paracetamol, caffeine) and cutting agent (mannitol) was seized. Two suspects found at the site were arrested, accused to possess this material. The heroin was found in 20 different portions/packages. At the same time in two other cities (B and C) two separate amounts of heroin in portions of 0.5 kg each were seized. Further police investigation led to the

assumption that the two arrested suspects were the owners of the 20 packages of heroin and also accused to deal with heroin seized in city B and C.

Question: Do all these heroin preparations belong to a certain (same) class?
Are the heroin preparations from the two different cities (B and C) subsets of the heroin seized at the house search in city A?

At least 10 g of each of the 20 packages of heroin seized in city A were sent to the laboratory. The material was individually homogenised, and aliquots were analysed to quantify the main composition of alkaloids and specified selected impurities formed in the manufacturing process of heroin.

For multivariate analysis numerical data of the quantitative constitution of main alkaloids of heroin preparations found in analysis by Gas Chromatography-Flame Ionisation Detection (GC-FID) were selected (UNODC, 2005). To minimise the influence of hydrolysis, ratios of the following alkaloids were calculated.

- Total morphine (sum of diacetylmorphine, 6-acetylmorphine and morphine, TM) and total codeine (sum of acetylcodeine and codeine, TC)
- Total morphine (TM) and narcotine (N)
- Total morphine (TM) and papaverine (P)
- Papaverine (P) and narcotine (N) [5]

The ratios are expressed as: TM/GC, TM/P, TM/N and P/N.

For impurities the peak areas of 25 selected components were chosen gained by GC-FID analysis of an acetic toluene extract according to Strömberg et al (2000).

[5] The P/N ratio can be mathematically constructed from the two presuming ratios. All three are included, because they each individually give different information on the production process.

ID	TM/AC	TM/P	TM/N	P/N	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	Sum P21-P23
6900	13.093	28.960	2.366	0.082	95.439	0.000	10.846	8.217	61.536	0.688	115.516	1.602	2.390	3.595	13.220	15.832	1.709	13.392	9.458	14.516	1.971	18.887	15.764	8.604	132.408
6899	13.067	30.053	2.459	0.082	89.024	0.000	9.974	7.386	58.660	0.000	97.654	1.153	1.982	3.113	11.874	13.680	0.000	10.372	7.911	13.186	1.827	17.084	14.229	7.454	111.312
6898	11.582	12.869	0.976	0.076	176.009	0.000	9.051	8.559	102.229	0.000	206.537	0.000	11.975	6.524	30.056	30.909	2.808	20.022	15.280	18.411	2.240	22.608	19.491	13.569	155.621
6897	11.587	12.921	0.975	0.075	175.216	0.000	9.145	8.916	102.118	1.988	208.061	0.000	11.195	6.734	30.451	31.414	2.660	19.837	15.198	18.954	2.102	23.091	19.562	13.825	155.011
6896	11.574	12.987	0.981	0.076	189.244	0.000	9.709	9.419	109.891	0.982	224.555	0.000	11.795	7.049	32.352	32.498	2.589	20.265	15.234	19.620	2.113	24.695	20.810	14.063	157.462
6895	11.579	12.793	0.971	0.076	184.177	0.000	9.475	9.697	110.603	1.792	214.451	0.000	9.978	7.056	31.507	31.766	2.608	19.581	15.147	19.249	1.847	24.647	20.389	13.508	148.552
6894	11.442	29.212	1.697	0.058	33.279	0.000	1.172	6.679	45.502	2.979	244.227	2.183	2.878	4.158	17.341	19.320	0.000	14.975	7.719	10.820	4.267	15.189	29.538	5.186	21.736
6893	11.402	28.590	1.628	0.057	34.253	0.000	1.160	6.768	43.554	2.751	258.252	2.281	2.907	4.367	17.747	20.144	1.444	16.615	8.057	10.694	4.565	15.678	31.794	5.644	23.652
6892	14.312	24.200	0.934	0.039	62.259	2.943	3.215	17.056	30.004	0.905	69.385	48.762	1.560	2.704	1.758	12.221	2.347	3.258	4.663	7.465	1.489	15.396	8.925	8.327	52.211
6891	11.555	12.356	0.961	0.078	189.010	0.000	9.846	8.617	89.328	0.879	249.460	0.000	3.984	6.934	33.023	36.767	4.187	28.205	19.535	17.644	3.915	22.435	20.979	16.415	189.867
6887	15.294	27.786	1.670	0.060	88.346	0.000	8.581	9.675	43.673	0.000	159.118	12.889	3.196	4.346	7.780	16.913	2.062	16.120	6.656	16.558	1.374	16.479	12.072	10.570	150.698
6886	15.177	22.903	1.730	0.076	112.250	0.000	10.636	10.234	44.452	0.698	154.006	11.478	3.351	4.946	8.034	18.091	2.868	18.731	7.167	22.493	1.422	19.163	12.767	10.609	169.493
6885	15.293	27.689	1.658	0.060	89.618	0.000	8.595	9.825	44.912	0.000	157.034	11.927	3.186	4.398	7.704	16.599	2.114	15.849	6.636	16.839	0.959	16.153	11.771	10.257	150.913
6883	13.281	25.116	1.937	0.077	101.298	0.000	9.163	8.514	45.980	0.000	122.102	3.325	2.943	3.637	13.183	16.862	1.922	14.120	9.610	13.488	2.002	17.213	13.149	9.070	127.213
6882	13.277	25.281	1.937	0.077	104.558	0.000	9.510	8.729	50.135	0.000	130.245	3.489	3.170	3.519	13.140	17.391	2.134	15.117	9.601	13.878	2.167	18.299	14.326	9.376	136.365
6881	11.607	12.686	0.983	0.077	177.538	0.000	9.040	8.902	74.377	1.084	236.834	0.000	4.273	6.423	30.306	34.644	3.921	26.758	18.438	16.370	3.365	21.505	18.686	14.782	173.916
6880	12.983	23.959	1.828	0.076	92.713	0.000	8.204	7.416	44.660	0.731	116.854	2.629	2.861	3.407	13.279	16.342	1.839	13.765	9.561	12.727	1.788	14.531	11.836	8.118	116.811
6879	12.966	23.906	1.821	0.076	90.638	0.000	7.623	6.914	38.699	0.000	110.183	2.549	2.908	3.171	12.049	15.192	1.769	13.010	8.552	11.056	1.608	13.190	10.933	7.511	107.631
6878	12.989	22.960	1.786	0.078	105.553	0.000	9.043	8.331	43.858	0.000	125.715	2.897	3.646	3.734	14.267	17.803	2.148	15.571	10.386	12.826	2.918	16.110	12.720	8.742	127.146
6877	12.967	22.600	1.744	0.077	104.764	0.000	8.121	7.558	42.842	0.740	128.241	2.668	4.580	3.784	14.481	18.054	2.051	15.434	10.328	12.053	2.710	16.939	12.116	8.672	120.311
6875	14.387	26.684	1.629	0.061	82.470	0.000	7.305	9.195	37.423	0.000	118.071	9.689	2.300	3.563	8.969	13.950	1.555	11.147	6.936	13.875	0.983	14.887	10.528	8.046	111.147
6874	13.317	25.708	2.001	0.078	84.253	0.000	7.639	7.427	38.908	0.000	93.321	2.748	1.835	2.958	11.276	13.360	1.394	10.501	8.152	12.373	1.506	15.735	10.150	6.583	94.895

Table 6-5. Analytical data of the seized heroin samples.

The background data used to develop the clustering method contained 606 items of the Heroin Database of the Federal Criminal Police Office (BKA), Germany, collected in the years 2016 - 2022. For these, the 4 quotients of the main component analysis and the 16 most important peak areas of the impurity profile were selected for data analysis resulting in total of 20 variables (TM/TC, TM/P, TM/N and P/N), impurity peaks (P1, P3, P4, P5, P7, P10, P11, P12, P14–P20, sum P21+P22+P23). As all these variables were on different scales, a simple standardization of the data was performed (subtraction of the mean, division by the standard deviation).

The aim of the method was to cluster the heroin samples based on analytical data into groups that correspond as closely as possible to an expert's assessment of different classes of heroin with a class here implying the samples originate from a common source. According to expert assessment, there were total of 66 groups consisting of at least two items and 250 items not assigned to any group. A distance metric was used to assess the similarity between the samples and prior expert assessment of existing groups to calibrate the method. Hierarchical cluster analysis (HCA) was used to obtain the actual clusters as it is a rather simple and widely applicable clustering method.

A large number of samples in the database were a priori known to not belong to any class. As such large number of singleton classes could confuse clustering algorithms, another preprocessing step was implemented to first filter out as many of these singletons as possible. This was accomplished by considering for each item its distance to all the other samples and then observing, how many samples were within a given distance threshold from that item. If no other sample was within the threshold distance of the sample under scrutiny, the sample would be considered an outlier and not included in the clustering. The threshold was set so that minimum number of items classified by the expert would be discarded. This constitutes a step that is taken before any clustering is performed with its own threshold value independent from the cluster analysis.

Next the remaining samples were clustered based on their pairwise distances using HCA. To provide a clustering solution, HCA requires a cutoff value to determine at which level the hierarchical cluster tree is cut. This cutoff was obtained by maximizing the similarity between the obtained clustering and the expert assignment of classes while ignoring the known singleton clusters. The adjusted Rand index (ARI) (see e.g. Warrens, 2022 for a detailed description) was used to measure this similarity. ARI in essence provides a numerical value between 0 and 1 for any two groupings of a dataset with 0 implying complete mismatch and 1 implying complete agreement.

In practice, the Manhattan distance was chosen as the metric for between sample distances for its simplicity and ability to adequately separate the items in the dataset. This distance is given by the equation

$$d(x,y) = \sum_{i=1}^{20} |x_i - y_i| \qquad (6.4.1)$$

where the summation is over the measured and pre-processed values for two samples. Further, the threshold for filtering the excess outliers (i.e. items that belong to no class) was set to 2.75, which allowed the vast majority of singleton classes in the data to be excluded from the analysis with minimal effect on other classes. However, some expert assigned classes were affected as the numerical data by itself was not enough to identify connections between the items belonging to these classes. This is explained by the fact

that the expert assessment of the classes included additional information e.g. related to the seizures and visual inspection of chromatograms.

After filtering out the outliers, HCA was applied to the dataset using different linkage rules. For each linkage rule, the hierarchical cluster tree given by HCA was cut at different cutoff values to provide candidate clusters. The ARI value between the candidates and the expert assigned classes was calculated for each cutoff threshold by first removing any known remaining singleton classes from the calculation. For each linkage, the best ARI score obtained thus was retained. The so-called “average” linkage rule was found to perform the best on the data providing an ARI score of 91.9% using a cutoff threshold of approximately 3.71. This threshold value in the context of average linkage HCA indicates that all obtained clusters have average pairwise distances below 3.71.

The known singletons were excluded from ARI calculations as HCA would typically assign them to clusters among themselves which would needlessly reduce ARI. This is undesirable as it is not expected that the expert has truly inspected every possible pairwise comparison in the dataset and some of the supposed singleton samples in reality form groups based on chemical similarity found in the included variables. These extraneous groups might in fact be of interest in themselves and provide new insights for the expert. By excluding the singleton classes from ARI calculations, the performance criterion focuses on the algorithm’s ability to correctly group the classes properly determined by the expert.

This method was then applied to the case data. It was found immediately that none of the items in the case data were close to any items in the background dataset. However, the case samples formed clear clusters. Only one case sample (6892) fulfilled the previously determined outlier criterion, but it was still included in the HCA computation as a single outlier was not expected to confuse the calculations significantly. The results of HCA are illustrated in the so-called dendrogram graph given in Figure 6-19. A dendrogram shows the hierarchy of clusters given by HCA where each node indicates combining two groups corresponding to the branches leaving the node into a new, bigger cluster. The height of the node shows at which point the joining of two clusters occurred and its interpretation depends on the linkage rule. For “average” linkage this means the average distance within clusters.

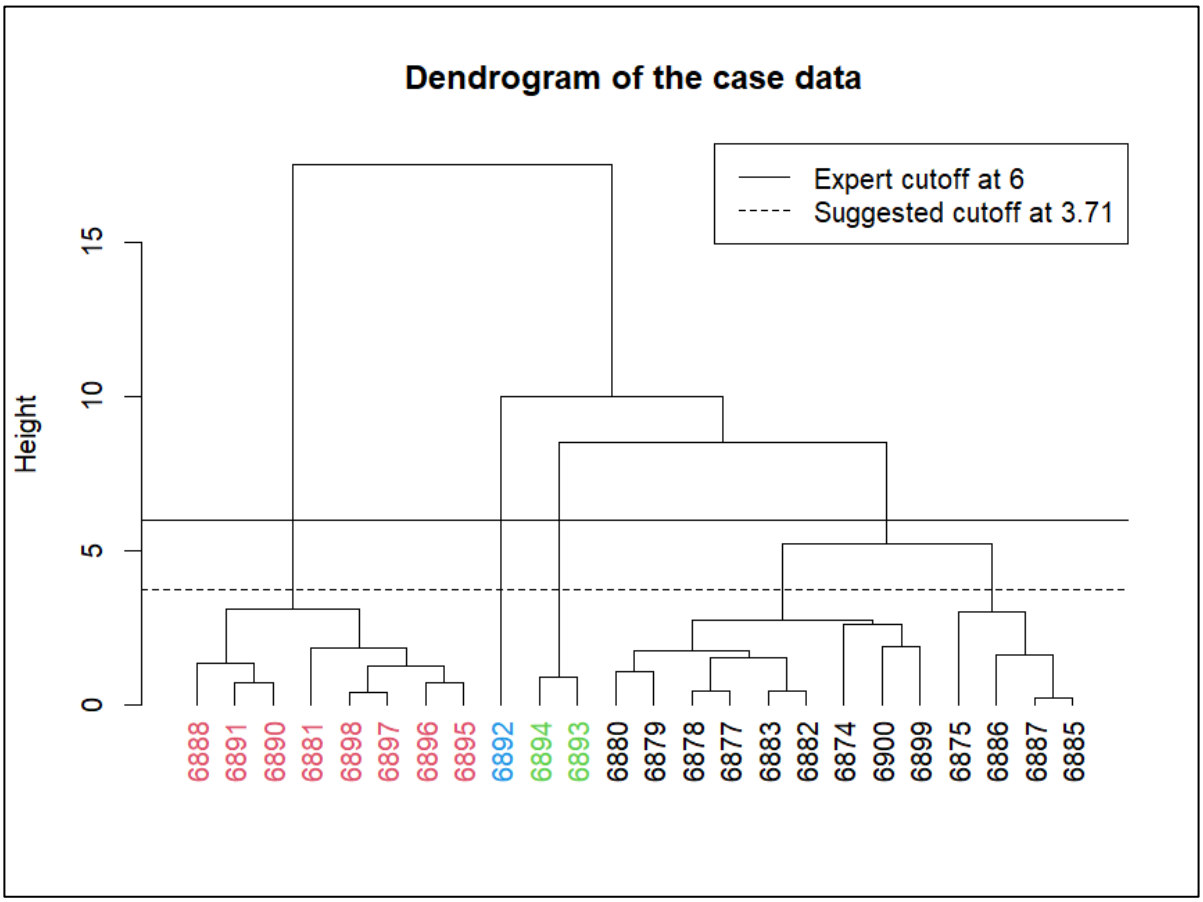

Figure 6-19. Dendrogram of Hierarchical Clustering using "average"-linkage on heroin case samples. Colours indicate classes assigned by expert. Horizontal lines indicate dendrogram cutoffs based on expert assessment and training data respectively.

Using the 3.71 cutoff provides some clear clusters but additional information, based on the chemical information from the expert indicates that the items on the right-hand side of the dendrogram belong together and, in expert opinion, a cutoff value of 6 would be more reasonable. This goes to show that the given data does not fully represent all the information the expert has access to with respect to assigning items to classes, because in the HCA only the numerical values of the peak areas and quantitative constitution can be assessed. Other information like external characteristics (e.g. salt form and similar) are not transferrable to the HCA model. However, the model provides a good starting point. In effect, an expert should always review the clustering results to ensure reasonable results. It should be noted that HCA is a fairly simple clustering method and a more advanced method with a more sophisticated measure of similarity could produce superior results. Regardless, the numerical data itself is likely too rough to accurately determine the exactly correct clusters in this case and as such expert input is still recommended for further analysis. It could also be beneficial to introduce more data e.g. in the form of the full chromatograms.

For the case example the conclusion is:

According to the threshold value of 6 at least four clusters can be assigned.

In the expert opinion the heroin samples No. 6899 and 6900 were clustered to a group of samples of 13 items marked in black at the right side of the dendrogram. Sample No. 6899 was from city B and sample No. 6900 was seized in city C. All other samples were seized in

city A. This type of result is congruent with results of items that are from a common origin, indicating a common bulk amount and show a link of the heroin seized in the three cities.

The remaining heroin samples belong to three additional classes. Eight heroin samples No. 6888 to 6895 at the left branch (colour red) also form a cluster that can assign to a second bulk amount of heroin. The samples No. 6893 and 6894 form a cluster (colour green) of a third bulk amount of heroin while heroin sample No. 6892 (colour blue) forms a fourth cluster of a single element. Heroin samples that show such similarities originate from three different common sources or bulk amounts.

The obtained clustering solution fully answers the question given to the expert in this case and as such calculating a likelihood ratio is not necessary. If a further, more precise question is posed e.g. regarding the source of two specific items, an LR approach could be applied.

6.4.3. Example 4b: LRs for comparison of two heroin seizures

The ultimate forensic question about two seizures of heroin would be: “Do the two seizures have a common origin?”. However, following the arguments of section 6.4.1, such a question cannot be answered for the current heroin production. There is not enough information to conclude on the very origin (site of production) of a seizure of heroin. We therefore interpret common origin here as common bulk of heroin samples. With bulk we mean a class as defined in section 6.4.1.

The propositions of interest are formulated as

H_1: *The two seizures of heroin belong to the same bulk origin.*
H_2: *The two seizures of heroin belong to different bulk origins.*

The background data were the same as in example 4a (section 6.4.2). They consist of the total of 606 seized heroin materials, clustered into 85 groups. In this grouping there is one heterogeneous group with materials that could not be allocated to an explainable group and will hence be left out from this example. From the 25 variables used to cluster the raw data presented in section 6.4.2, the four alkaloid ratios (TM/TC, TM/P, TM/N and P/N), nine of the impurity peaks (P1, P4, P7, P10, P11, P12, P14, P19 and P20), and a sum of three impurity peaks (P21+P22+P23) were retained for the analysis. Because of small peak areas, overlapping peaks as well as asymmetrical peak shapes, the others were not taken into account. The impurities of peak No. 21, 22, and 23 transform into each other. Therefore, the sum of these three peak areas was calculated and used as a single peak.

Taking out the heterogeneous group and removing materials with missing data (peaks), 407 materials distributed over 83 groups with at least two materials per group remain. These constitute the ground truth for fitting and assessing a statistical model for obtaining likelihood ratios for comparison. The 83 groups represent the 83 bulks for which the comparison model should be valid.

Since there are 14 variables to be included in the model, and in addition 83 classes to separate, several with very few materials, it is not feasible to fit a multivariate probability distribution to data, i.e. using feature-based evaluation. Therefore, score-based evaluation will be applied.

With score-based evaluation a measure of distance (or similarity) must be chosen (cf. section 6.2.2). There are many such measures and at present no generic grounds for choosing one over the other.

The similarity score chosen here – for purpose of illustration - was a modification of the so-called *quotient method statistic*. This was proposed by Jonsson and Strömberg (1993) as a way of concluding on potential matches between amphetamine seizures. However, this method is merely chemometric with very limited possibilities to go from the calculated statistic to a likelihood ratio.

Our modified statistic is described as following:

For two seized materials there are n variables monitored (for the heroin example n = 14). The observed variables are denoted $x_1, \ldots, x_n$ for material 1 and $y_1, \ldots, y_n$ for material 2. The quotients, $q_i = x_i/y_i$, $i = 1, \ldots, n$, are then calculated. Such quotients represent a comparison between the two materials, and should all quotients be equal to one this would mean that the variables are identical between the two materials. Hence, a reasonable heuristic conclusion would be that the materials have a common origin (for the heroin seizures a common bulk). A similar conclusion would be drawn if all quotients are equal but not equal to one. Such a situation indicates that the origin of the materials is the same, but that one of the materials has an overall shift in level of the variables. However, we cannot expect such a result even if the materials have a common origin. Nevertheless, the ratios should not deviate too much from each other even if we must expect that a quotient with x and y both high in value is larger than a quotient with x and y both low in value. This is so since the variation in chemical measurements increases with the magnitude of the measurand.

The modified quotient statistic is

$$Q = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{|q_i - q_j|}{q_i + q_j} \qquad (6.4.2)$$

where $|q_i - q_j|$ is the absolute difference between q_i and q_j, and the summation is over all pairs of quotients. The $q_i + q_j$ sees to that the absolute difference may vary with the magnitude of the variables.

The structure looks similar to the Canberra distance[6] but the difference is that the comparison is here done in the calculation of quotients, and it is the accumulated distances between quotients that is the score from the comparison. The measure Q varies continuously and hence allows more variation for fine-tuning separation between bulks.

The measure Q was calculated for all comparisons of two materials within a class and for the comparisons of two materials from different classes randomly sampling two materials from each class. This procedure was intended to make the difference in number smaller between within-class comparisons and between-classes comparisons to obtain better balance. The total number of within-class comparisons is 1 537 and the total number of between-classes comparisons is 13 612. In Figure 6-20 are shown histograms of the

[6] $\sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$

calculated *Q* statistics (scores) for within-class comparisons (blue) and for between-classes comparisons (red).

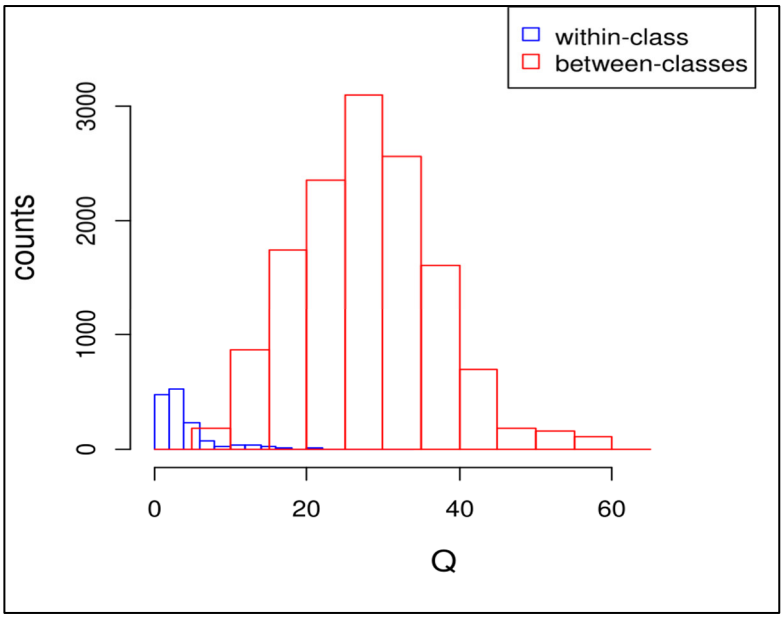

Figure 6-20. Histograms of the quotient statistic Q for within-class comparisons (blue) and between-classes comparisons (red).

The histograms in Figure 6-20 shows that the quotient statistic (score) was generally low for within-class comparisons and more widely distributed for between-classes comparisons, however with low counts for small values. A histogram gives a non-smoothed image of a distribution but suffers from irregularities due to that the number of observations graphed may not be sufficient to mirror the general shape of the distribution.

To get a better picture of how the score varies in distribution between within-class and between-classes comparisons, kernel density estimation was applied to each of the two sets of scores (see section 5.2.1). The resulting densities are shown in Figure 6-21.

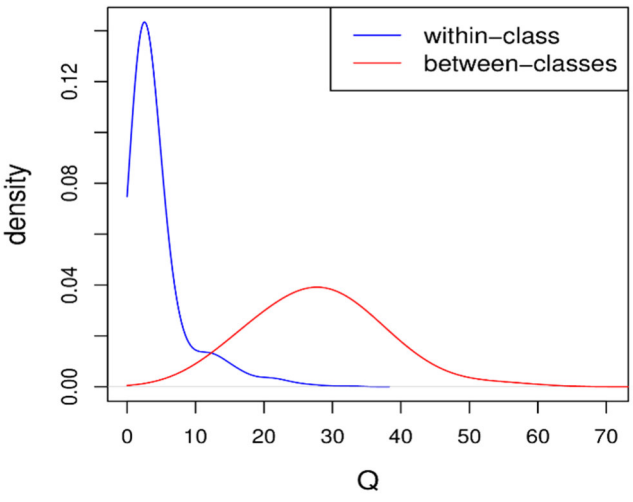

Figure 6-21. Kernel densities for the quotient statistic for within-class comparisons (blue) and between-classes comparisons (red).

For any comparison of two seizures of heroin the resulting score (i.e. value of *Q*) can be compared against the two densities to obtain the likelihood ratio. As an example, suppose the score *Q* is about 5. Reading off the densities on the blue and red curves in Figure 6-21 we obtain the numerator and the denominator of the likelihood ratio as approximately 0.090 and 0.004 respectively (see Figure 6-22). The resulting likelihood ratio is thus 0.090/0.004 = 22.4.

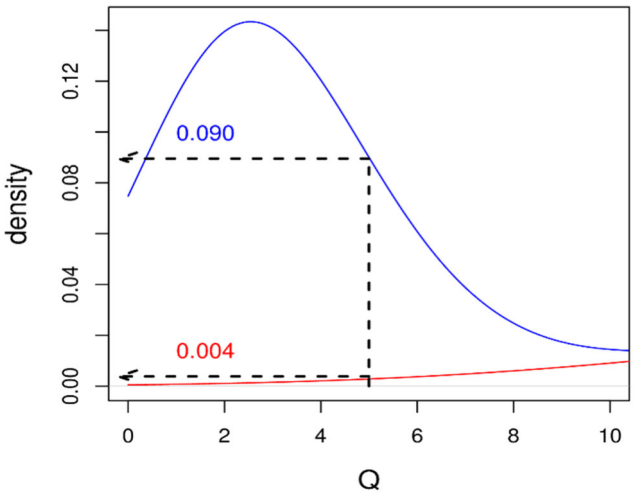

Figure 6-22. Reading off numerator and denominator of the likelihood ratio from a score of 5, based on the kernel densities shown in Figure 6-21. The numerator is 0.090 (crossing of vertical line with blue curve) and the denominator is 0.004 (crossing of vertical line with red curve).

Validation

The score-based method presented above was validated using so-called *leave-one-out cross validation* (LOOCV) on a subset of the ground truth. LOOCV for comparison means that two materials at a time is taken out from the ground truth, the evaluation model is fitted to the remaining data and then applied to the two materials left out, i.e. here make a score-based evaluation of the comparison of these two. The reason for using a subset was partly that several classes comprised only two materials, which complicated the assessment of within-class comparisons, partly because applying LOOCV to the entire dataset is very time-consuming.

Therefore, for the validation only the materials used to fit the complete model (resulting in the kernel densities of Figure 6-21 and only classes with at least three materials were used.

In Figure 6-23 histograms are shown of *logarithms* of likelihood ratios for comparisons of materials known to belong to the same class (blue bars) and for comparisons of materials known to belong to different classes (red bars).

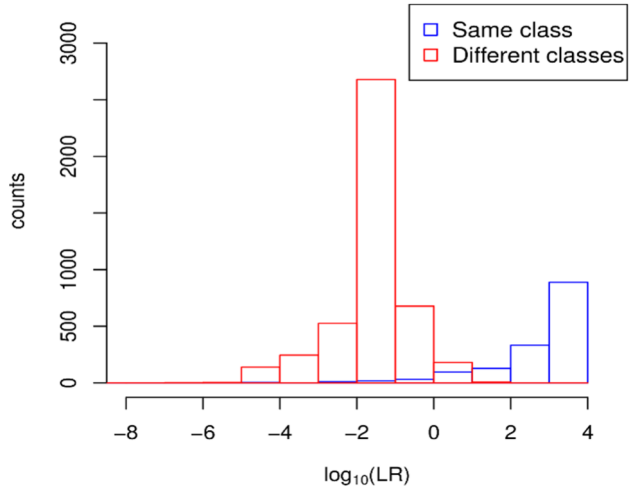

Figure 6-23. Histograms of calculated logarithm of likelihood ratios for the hypothesis H_1: "The two seizures of heroin belong to the same class" vs. the hypothesis H_2: "The two seizures of heroin belong to different classes" for materials known to be within the same class (blue) and materials known to be within different classes (red).

In Figure 6-23, the blue bars for logarithms of likelihood ratios lower than 0 contain the number of comparisons where a false negative conclusion was made, in total 68 comparisons. The total number of same source comparisons were 1511. The red bars higher than 0 contain the number of comparisons where a false positive conclusion was made, in total 189. The total number of comparisons were 6384, which gives a false positive rate of 189/6384 ≈ 3.0% and a false negative rate of 68/1511 ≈ 4.5%.

In Figure 6-24 is shown the ECE plot from the validation study. The curve for the computed likelihood ratios (red, observed) is well below the neutral evidence curve (black, null) and close to the calibrated likelihood ratios curve (blue). The Cllr is ≈ 0.188 quite close too cllrmin ≈ 0.160. In Figure 6-25 the DET plot for the computed likelihood ratios is shown.

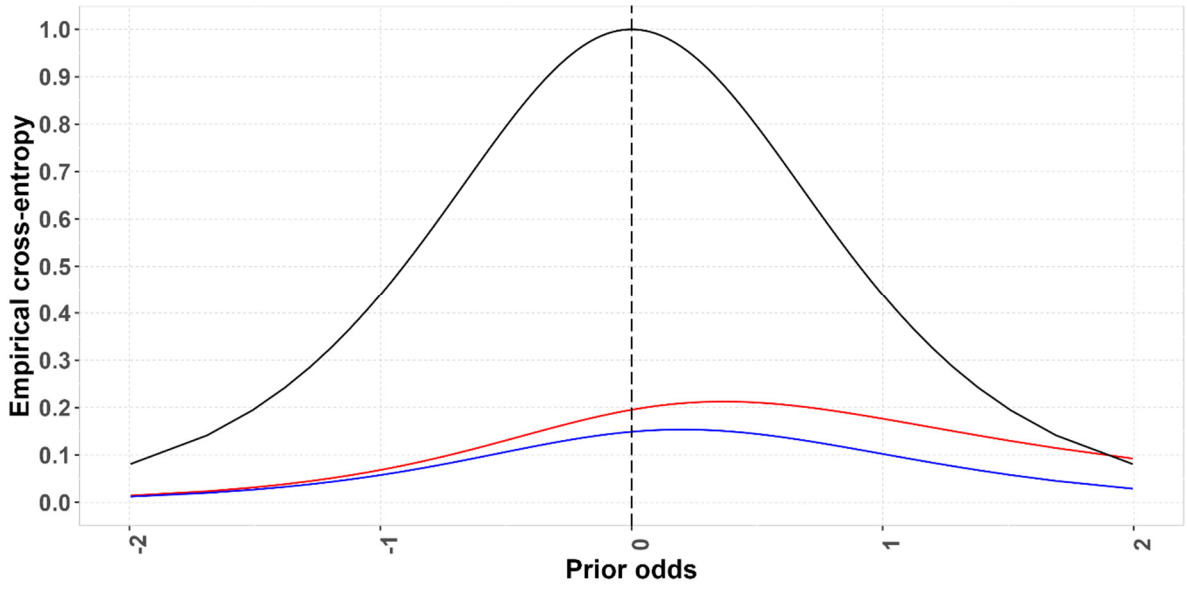

Figure 6-24. Empirical cross-entropy plot of the likelihood ratios obtained in the validation study.

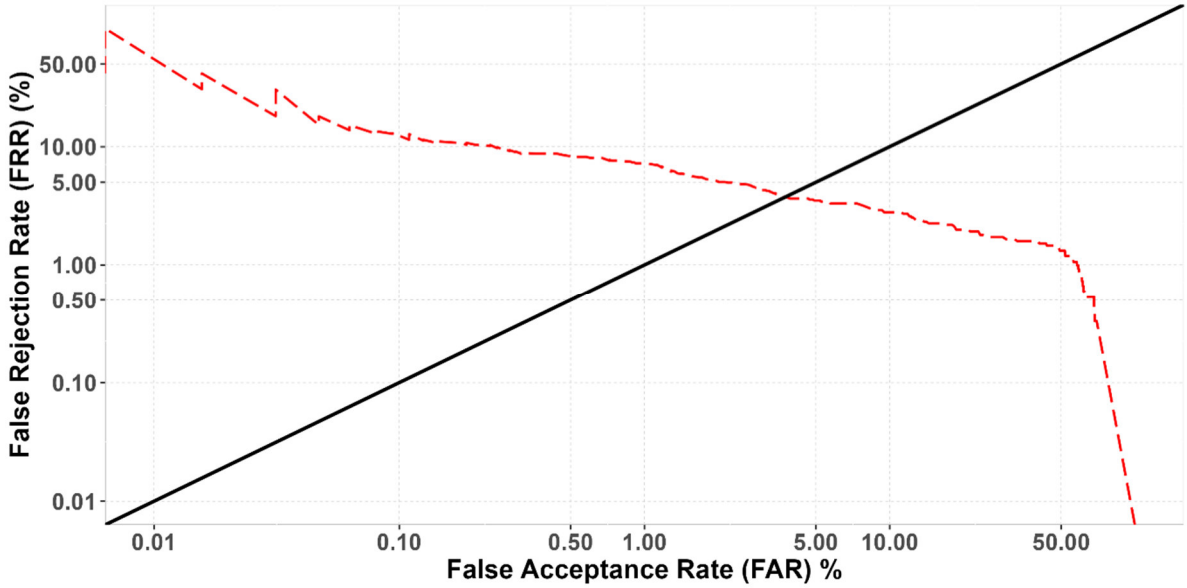

Figure 6-25. DET plot of the likelihood ratios obtained in the validation study. EER is about 3.71.

Case example

Assume we have two seizures, A and D of heroin. Their chromatograms are presented as mirrored in Figure 6-26.

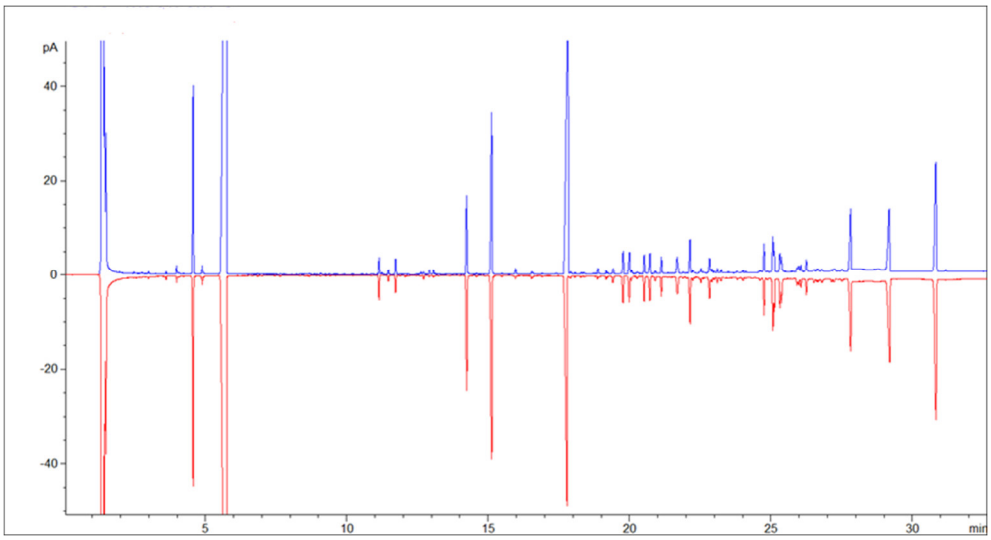

Figure 6-26. Chromatograms from the analyses of seizure A (upper, blue) and seizure D (lower, red).

An experienced chemist would draw an initial conclusion that the two seizures come from the same class, since the chromatograms are very similar. To assess this observed similarity, we apply the quotient statistic to the variables selected for the LR model developed above (i.e. TM/TC, TM/P, TM/N and P/N, P1, P4, P7, P10, P11, P12, P14, P19 and P20, and the sum P21+P22+P23). Table 6-6 gives the values of these variables for the two seizures.

Variable	*Seizure*		*quotient (q_i)*
	A	D	
TM/TC	13.317	13.093	1.017
TM/P	25.708	28.960	0.888
TM/N	2.001	2.366	0.846
P/N	0.078	0.082	0.863
P1	84.253	95.439	0.883
P4	7.427	8.217	0.904
P7	93.321	115.516	0.808
P10	2.958	3.595	0.823
P11	11.276	13.220	0.853
P12	13.360	15.832	0.844
P14	10.501	13.392	0.784
P19	10.150	15.764	0.644
P20	6.583	8.604	0.765
P21+P22+P23	94.895	132.408	0.717

Table 6-6. Values of selected variables for comparison for the two seizures A and D and the pairwise quotients.

The value of the quotient statistic *Q* is 5.53. Using the kernel densities of Figure 6-21 the likelihood ratio becomes 0.077/0.004 = 19.25. Hence, the results of the comparison are 19.25 times more probable if seizure A and D come from the same class (bulk of heroin) than if they come from different classes. Also, the value of the likelihood ratio seems low, in forensic practice in comparing heroin a value of approx. 19 is corresponding to a close similarity that is shown by samples coming from a common bulk amount.

Another seizure, B, was also compared with seizure D. Their mirrored chromatograms are shown in Figure 6-27.

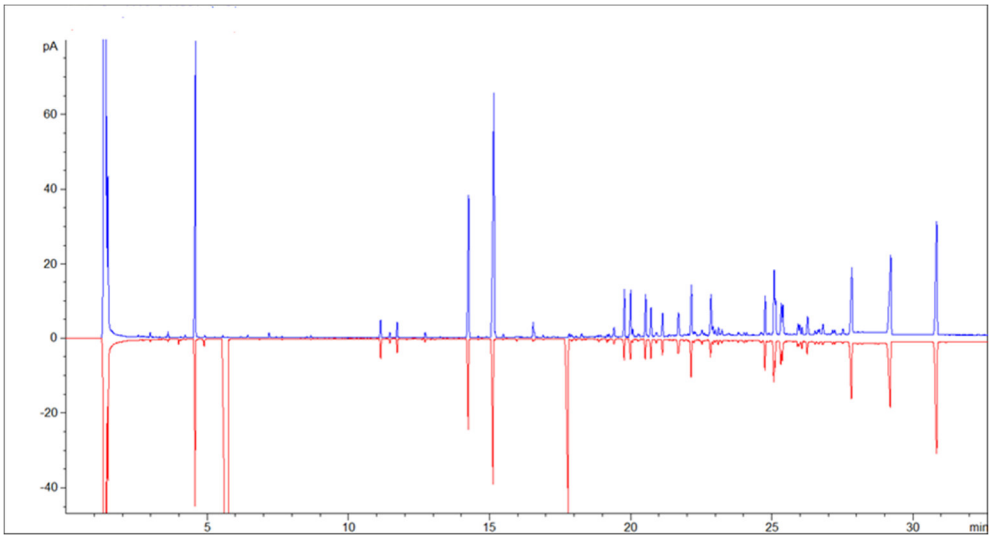

Figure 6-27. Chromatograms from the analyses of seizure B (upper, blue) and seizure D (lower, red).

The experienced chemist would conclude that these two seizures belong to different classes. This is so since the chromatograms obviously differ much. However, what is the evidential strength of such a hypothesis? Table 6-7 gives the values of the variables selected for the LR model used to evaluate the comparison between seizure A and D.

Variable	*Seizure*		*quotient (q_i)*
	B	D	
TM/TC	11.574	13.093	1.017
TM/P	12.987	28.960	0.888
TM/N	0.981	2.366	0.846
P/N	0.076	0.082	0.863
P1	189.224	95.439	0.883
P4	9.419	8.217	0.904
P7	224.555	115.516	0.808
P10	7.049	3.595	0.823
P11	32.352	13.220	0.853
P12	32.498	15.832	0.844
P14	20.265	13.392	0.784
P19	20.810	15.764	0.644
P20	14.063	8.604	0.765
P21+P22+P23	157.462	132.408	0.717

Table 6-7. Values of selected variables for comparison for the two seizures B and D and the pairwise quotients.

The value of the quotient statistic is 25.72. Using the kernel densities of Figure 6-21 the likelihood ratio becomes 0.001/0.036 = 0.028. Hence, the results of the comparison are 1/0.028 = 35.71 times more probable if seizure B and D come from different classes (bulks of heroin) than if they come from the same class.

6.5. Example 5: Classification of car paint samples

Classification is the assignment of a person or an object to a particular category (ENFSI, 2015). In this section we will examine the rationale behind employing likelihood ratios in forensic classification.

Hypotheses
H_1: *The sample belongs to class X*
H_2: *The sample belongs to any other class*

6.5.1. Background data

The EUCAP (European Collection of Automotive Paints) database is maintained by the Paint, Glass & Taggants working group of ENFSI, and contains comprehensive analytical data in forms of FTIR (Fourier Transform Infrared Spectroscopy) transmission measurements on free-standing thin sections of car coatings (ENFSI, 2022). While other

techniques (e.g. Raman spectroscopy, pyrolysis-gas chromatography mass spectrometry etc.) have been documented in the literature for the analysis of automotive paint samples, FTIR remains the primary method employed for their characterization due to the ease of use and database comparison possibilities, cf. (Duarte et al, 2022), (Fasasi et al, 2015), (Perera et al, 2018), (Kwofie et al, 2018).

The part of the EUCAP repository that are discussed here includes clear coatings only. Apart from spectral data, each sample is also attributed to a manufacturer of cars in the database. An obvious application would be to train a model that classifies paint samples according to its manufacturer. It is appealing to do so as a quick car brand prediction could serve as useful forensic intelligence. Also, the ground truth (which car brand a sample belongs to) is known from the database. Our initial efforts therefore focused on predicting car brands from FTIR samples. However, we were not successful in doing so.

The project group consulted with the paint experts who originally supplied the data. They explained that our difficulties could stem from the fact that FTIR spectra of clear coatings are heavily influenced by the type of binder used, and a car brand may use several different binder systems in its production lines. Conversely, the same binder system can be found across different brands. This overlap makes it difficult to reliably determine a car's brand based solely on its FTIR spectrum.

What is possible though is classifying the spectra based on binder type. This approach may have less direct forensic utility than predicting a car brand, especially given the absence of definitive ground truth in the database, which would be which binder type a sample belongs to. We chose to proceed in order to demonstrate the process of deriving classification likelihood ratios from spectral data.

We opted for a two-step approach. First, we classified the samples into groups using clustering of their FTIR spectra. These clusters are expected to reflect the type of binder used in the coatings. Secondly, the clusters were treated as ground truth, and we trained an LR model for assigning a new paint sample (object) to a specific cluster.

Raw data
For this exercise, we extracted FTIR data from 7 829 original paint samples and 48 automobile manufacturers from the EUCAP database.

The spectral data for a sample can be seen as a vector of intensities, with each intensity corresponding to a specific wavelength. These wavelengths serve as features for the classification of different samples. Table 6-8 shows a sample of how the data might appear in rows and columns in e.g. Microsoft Excel.

Wavenumber	**602**	**606**	**610**	**614**	**618**
Sample 1	0.105	0.108	0.108	0.111	0.114
Sample 2	0.092	0.093	0.090	0.098	0.100

Table 6-8. Example of data structure with samples in rows and features (wavenumbers) in columns. Only a few samples and wavenumbers are shown.

Data preprocessing
Most of the preprocessing can be executed with a software provided with the FTIR instrument (Lavine et al, 2014), (Kwofie, 2020). In this study, however, preprocessing was performed using the software R (R core team, 2020).

Wavelength alignment: In order to perform the analysis all samples must share the same features, in this case the wavelengths. The majority of samples in the database was represented by 788 wavelengths, starting from 602 and then every fourth integer wavelength as in table 6-8. However, some samples were represented by different wavelengths. For samples with wavelength measurements not aligned with the majority, the following correction was applied: for each missing wavelength (e.g., 606 nm), the intensity was estimated by linearly interpolating between the intensities at the two adjacent wavelengths (e.g., 605 nm and 607 nm). Linear interpolation assumes a straight-line relationship between the neighbouring points.

Slope correction: Baseline drift in FTIR spectra represents noise that needs to be corrected. The correction for baseline drift in spectra was performed by using asymmetric least squares (ALS). ALS fits a smooth baseline iteratively by minimizing a weighted least squares error, where weights adjust asymmetrically to penalize negative deviations more than positive ones. ALS can effectively remove baseline drift while preserving peak shapes. Figure 6-28 illustrates the impact of this correction on a spectrum.

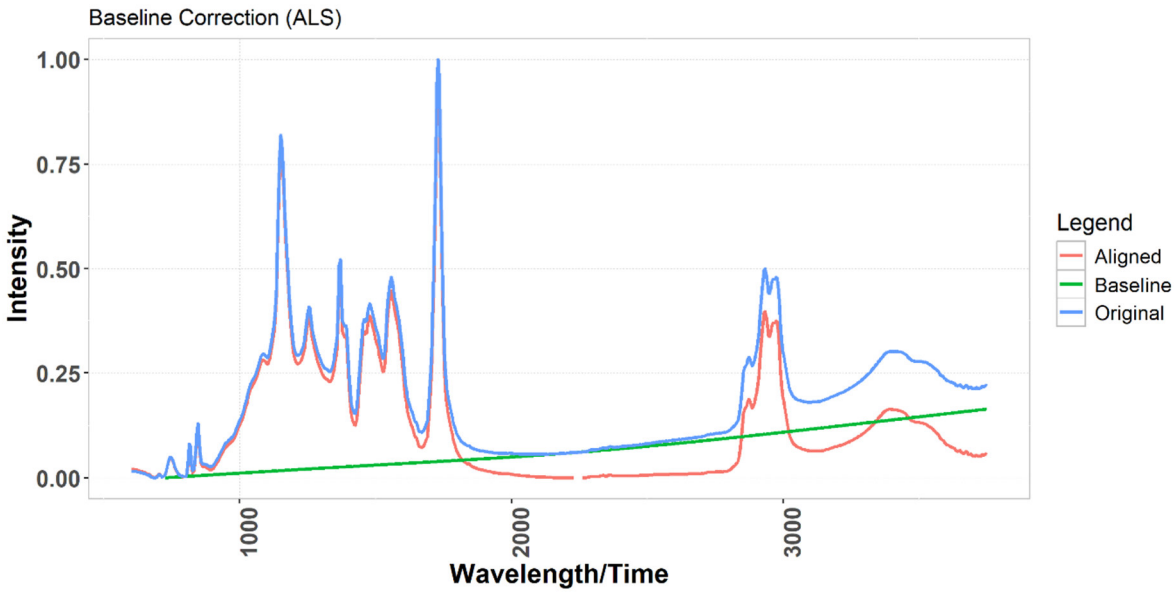

Figure 6-28. Asymmetric least square baseline correction method.

Wavelength filtering: Only the fingerprint (600–1900 cm^{-1}) and C–H stretching (2750–3150 cm^{-1}) regions of the spectra were selected to exclude interferences such as the asymmetric stretching band of CO_2. After the filtering, 427 wavelength features remained.

Wavelength standardisation: The final preprocessing required is standardisation of the variables. This preprocessing unifies the scales of the variables, to avoid domination by wavelengths of high intensity.

$$St\ Intensity = \frac{Intensity - \mu}{\sigma} \quad (6.5.1)$$

where μ is the mean, and σ is the standard deviation of intensity within each sample. Standardisation over each individual spectrum removes scale differences between samples but preserves relative patterns *inside* each sample.

6.5.2. Clustering for "ground truth"

Details of the clustering procedure are beyond the scope of this booklet. The result of the clustering was that each sample was assigned to a cluster, where each cluster is considered to correspond to a particular binder type. In the analysis, five clusters emerged, i.e., each sample was assigned to one of five clusters, named Cl_1 to Cl_5.

6.5.3. Dimensionality reduction

The resulting preprocessed matrix is of dimensions 7829 x 427, and its interpretation vastly exceeds the capabilities of human perception. To compute an LR based on the propositions mentioned earlier, we want to reduce the data's dimensionality. Specifically, we aim to decrease the number of features (427) without reducing the number of samples (7829).

Principal Component Analysis (PCA) simplifies the dataset by transforming the high-dimensional data into a lower-dimensional space and allows for visualization of the data in 2D or 3D plots, see Figure 6-29.

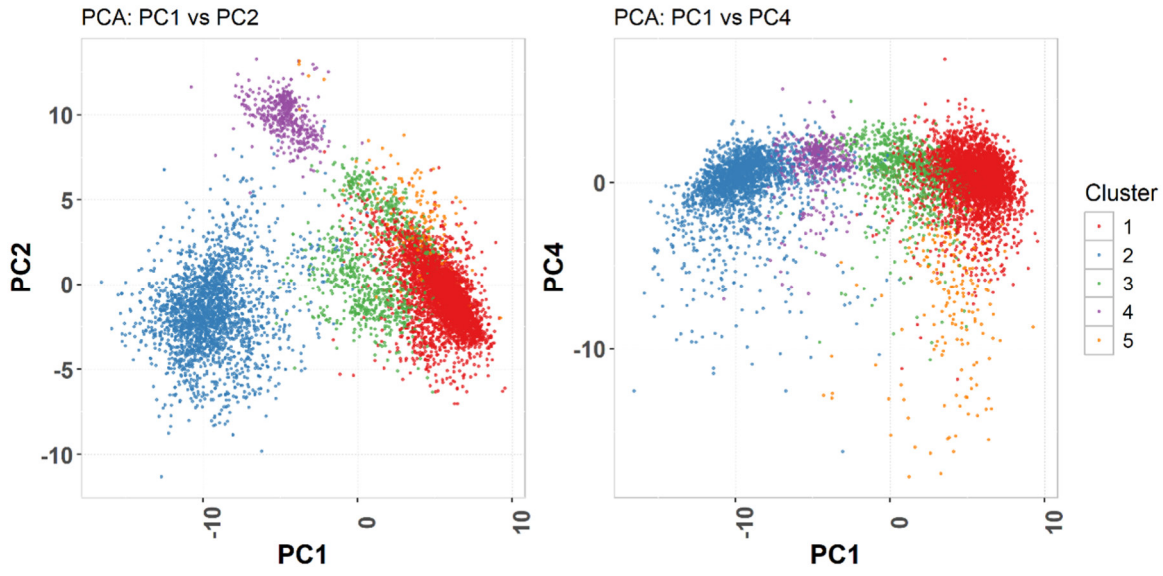

Figure 6-29. PCA scatter plots of PC1 vs PC2 (left) and PC1 vs PC4 (right).

The figure shows that clusters Cl_1–Cl_4 can be reasonably separated using PC1 and PC2, while PC4 contributes by separating cluster 5 (yellow) to some extent. This suggests that using these four principal components (instead of the original wavelengths) may effectively discriminate between the samples.

PCA doesn't directly provide any probabilistic interpretation for classification. Additional steps are needed to reach a classification based on LRs. One feasible additional step is to use Linear Discriminant Analysis (LDA). Unlike PCA, LDA directly computes discriminant functions that can be used to apply likelihood ratios for classification. We could have tried to apply LDA directly on the 7 829 x 427 matrix, but its underlying assumptions—such as multivariate normality—are difficult to satisfy in 427 dimensions. Reducing the dimensionality is better, and it also helps to stabilize the covariance matrix during LDA.

Moreover, as we plan to follow PCA with LDA it makes sense not to use more principal components (PCs) than the number of possible dimensions in LDA, which is always one less than the number of classes (here: clusters). In this case, with five clusters, only four dimensions are available for LDA. The eigenvalues of the PCs also support this choice, as the first four components account for a substantial portion of the variance—81%. In PCA, eigenvalues represent how much of the total variance in the data is captured by each principal component. Therefore, we let the first four PCs be the features of our car paint samples during LDA.

6.5.4. Linear discrimination analysis

LDA requires the data to be reasonably jointly normally distributed. Approximate normality for the features, i.e. the four PCs, can be checked using e.g. histograms. For the samples in this example, the PC distributions were slightly skewed. By applying a logarithmic data transformation to the PCs, the skewness was less pronounced. As numbers must be positive to apply a logarithmic transformation, we subtracted the minimum value in each principal component to all values in that PC. This assured all values were positive, except for the minimum value itself, now being zero. By also adding 1 to all values, we ensured all values were of value 1 or more before logarithmic transformation as in equation 6.5.2:

$$transformed\, x_i = \log\left(x_i - \min\left(x_{pc_i}\right) + 1\right) \quad (6.5.2)$$

Where x_i is the sample value in the i^{th} principal component.

The Linear Discriminant Analysis (LDA) requires prior probabilities for each sample's likelihood of belonging to a given cluster. To compute the likelihood ratio for testing whether a sample belongs to cluster X versus any other cluster, the priors are defined as:

- Cluster X: 0.5
- Each of the remaining four clusters: 0.5 / 4 = 0.125

This prior configuration was applied regardless of which cluster Cl_1 to Cl_5 was being tested. A separate model was trained for each cluster. When the question is "does the sample belong to cluster Cl_1?" we use the model for that cluster. When the question instead concerns cluster Cl_2, the corresponding model for Cl_2 is applied.

After checking the normality assumption, LDA was performed on the four features of the 7829 samples. The LDA plot for the Cl_1 model, i.e. using prior 0.5 for Cl_1 and 0.125 for other classes is shown in Figure 6-30. The cluster Cl_1 is shown as red points.

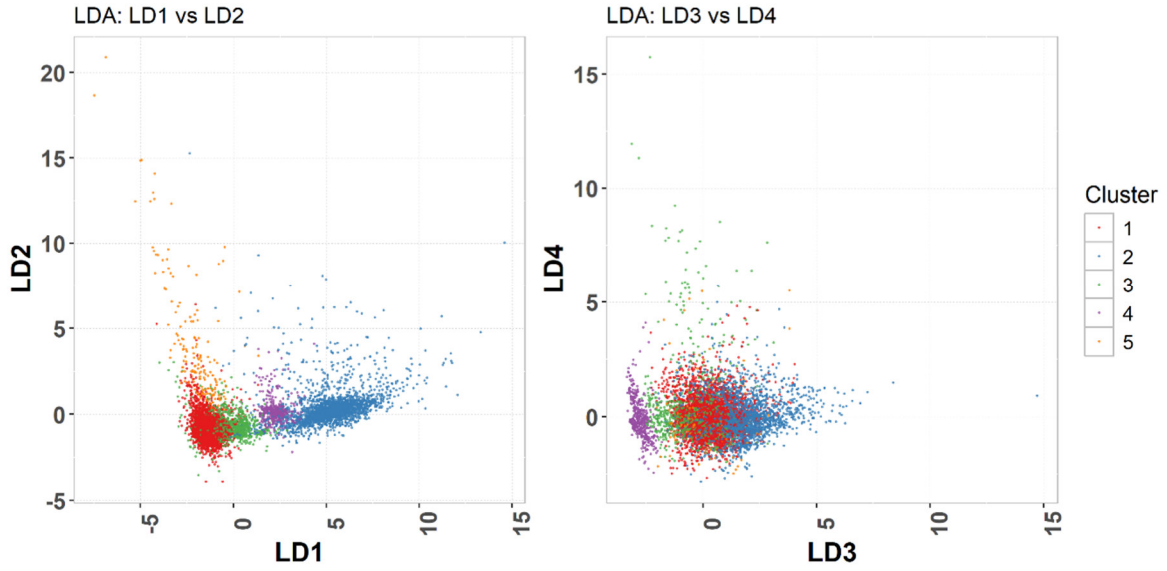

Figure 6-30. Scatter plots of LD1 vs LD2 (left) and LD3 vs LD4 (right) for cluster 1 model.

Likelihood ratio calculation

LDA estimates the posterior probability (p_{po}) associated with each class, and these probabilities are returned as the output of the analysis. The posterior odds ($odds_{po}$) are found by using the equation

$$odds_{po} = \frac{p_{po}}{1-p_{po}} \tag{6.5.3}$$

After this operation, by dividing the posterior odds by the prior odds ($odds_{pr}$):

$$odds_{pr} = \frac{p_{pr}}{1-p_{pr}} \tag{6.5.4}$$

a likelihood ratio can be obtained:

$$LR = \frac{odds_{po}}{odds_{pr}} \tag{6.5.5}$$

Since the prior odds is 1, the LR here equates the posterior odds ($odds_{po}$).
The extrapolation boundaries, i.e. the range of LRs for which the Cl_1 method is considered robust, were determined in accordance with (Alberink et al, 2026) to 1/515 (lower) and 90 (higher).

6.5.5. Likelihood ratio model validation

The five LDA models were validated using 5-fold cross-validation. Given any fold in the cross validation, the feature vectors were either used in the training (4 of the 5 folds) or in the test set (1 of the 5 folds). The posterior probabilities for cluster X were collected across all folds. Since the true cluster membership of each sample was known, the corresponding likelihood ratios could be classified as either H_1-true (sample belongs to cluster X) or H_2-true (sample belongs to another cluster). After this operation, density plots for H_1-true LRs and H_2-true LRs was constructed, see Figure 6-31.

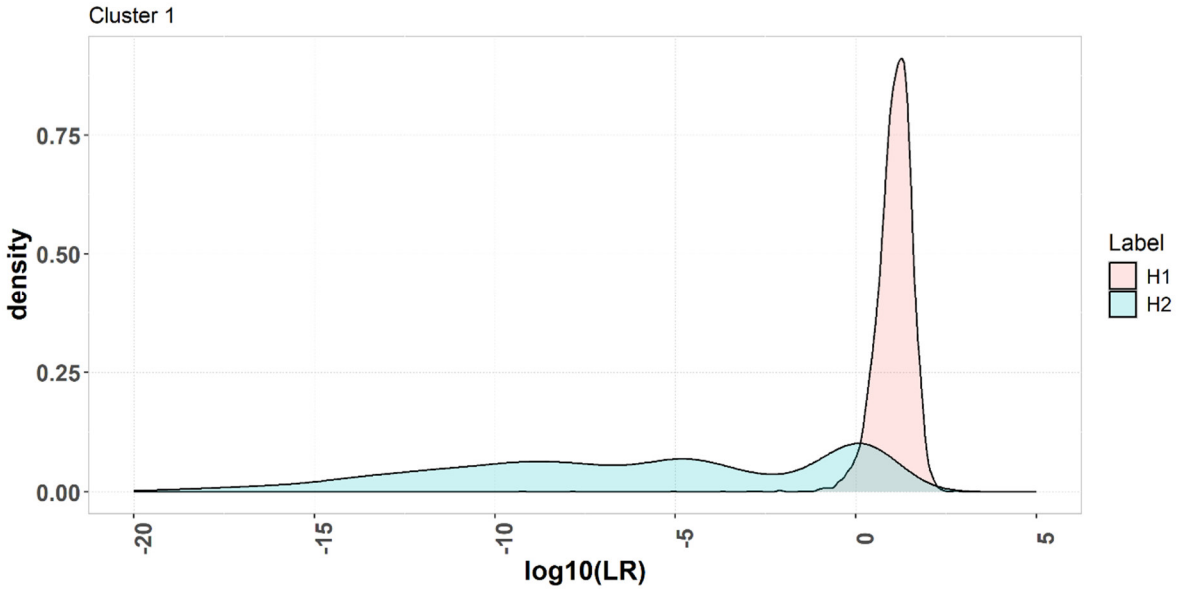

Figure 6-31. Density plots of likelihood ratios for the cluster 1 (red points in Figure 6-30). Blue curves represent the LR distributions for samples within the true class (H_1), while the red curve shows the LR distribution for samples from all other classes (H_2).

As can be seen from the distributions, the separation of curves is not perfect. This can also not be expected considering the overlap of colours in Figure 6-30. The maximum LR is ca. 4 000. Because this value exceeds the method's validated upper boundary (LR = 90), the result is reported as the upper limit for high likelihood ratios.

Performance metric	Cluster 1 (red)	Cluster 2 (blue)	Cluster 3 (green)	Cluster 4 (purple)	Cluster 5 (yellow)
Sensitivity	84.5%	99.9%	66.0%	100%	75.8%
Specificity	97.2%	95.6%	97.8%	97.5%	98.8%
Cllr	0.302	0.266	0.571	0.041	0.61
EER	8.6	1.6	13.2	0.4	4.3

Table 6-9. Performance metrics of the five LR systems.

Table 6-9 shows that performance varies among the different clusters. It is not surprising that the difficulty to separate varies among the clusters. For instance, the strong

performance of cluster 4 is consistent with what can be observed in the LDA plot for that cluster, see Figure 6-32. Since the cluster (purple dots) is not overlapping with other clusters, it is expected to perform well.

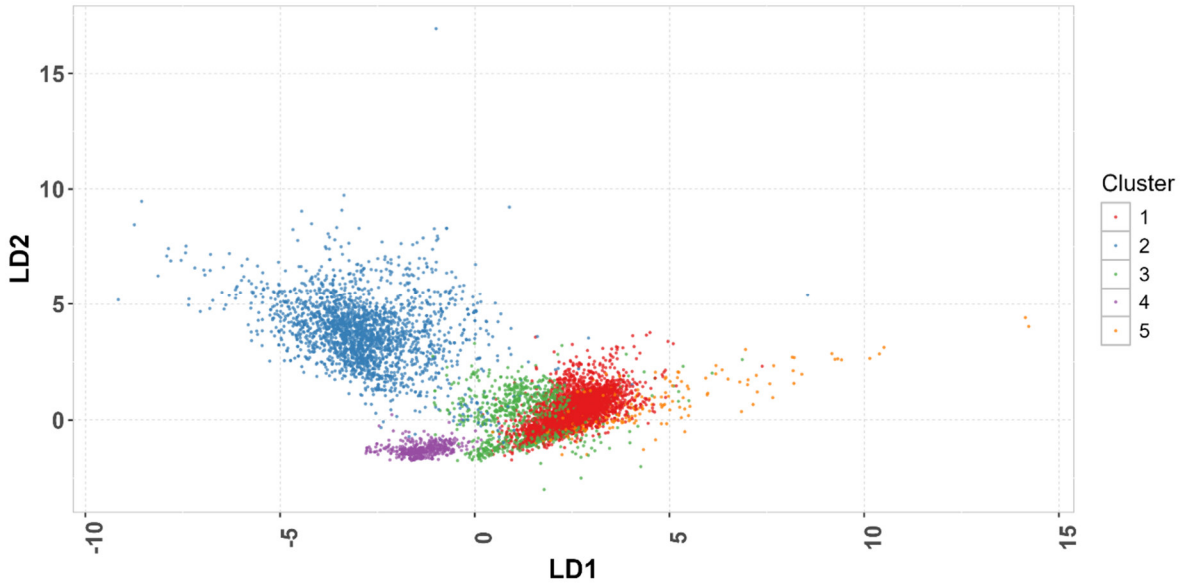

Figure 6-32. Scatter plots of LD1 vs LD3 for cluster 4 model.

It can also be seen from Table 6-9 that Cl_2 performs better than Cl_1 in terms of specificity and sensitivity, but their Cllr values are not that different. This indicates that the Cl_2 model produces fewer but more extreme misclassifications. In fact, several false negatives with LRs below 10^{-5} appear in the Cl_2 dataset.

Figure 6-33 presents the ECE plots for the Cl_1 model.

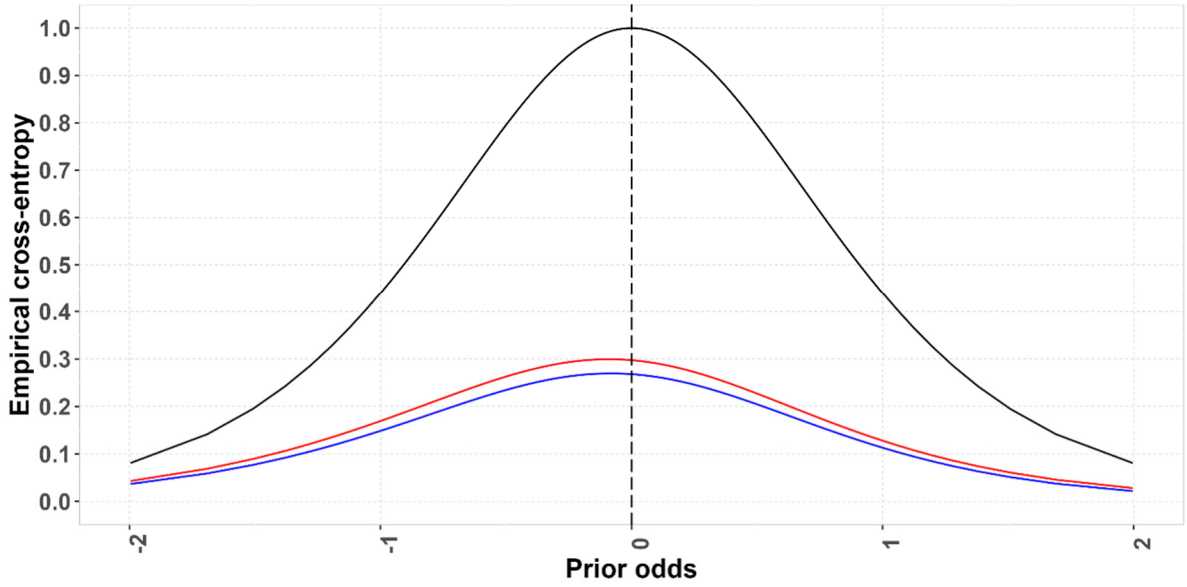

Figure 6-33. The ECE plot of the cluster 1 model.

The equal error rate

The Equal Error Rate (EER) varies extensively, as shown in Table 6-9. This difference in performance is further highlighted in the DET plots, Figure 6-34, where the best-performing model minimises the area under the curve in the lower-left corner.

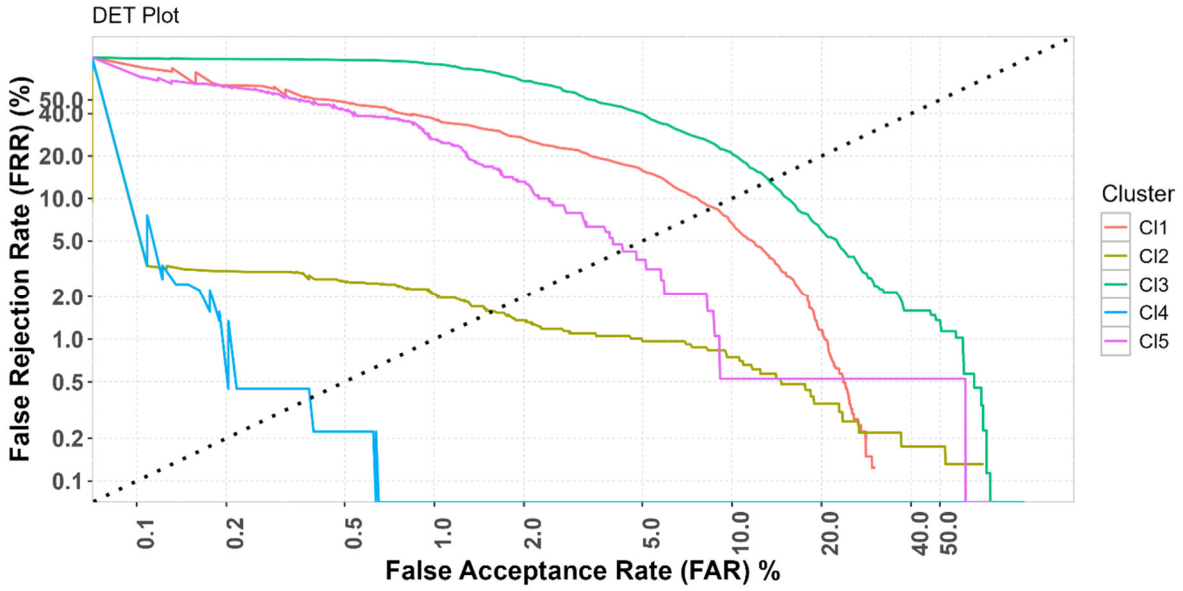

Figure 6-34. The DET plot of the five models, one per cluster.

Conclusion and discussion

This approach results in five likelihood ratio models—one for each cluster—where each model estimates the likelihood ratio corresponding to the question, “Does the sample belong to cluster X?” The performance of the models varies depending on how well each cluster can be separated from the others in the four-dimensional LDA space. Cluster 4 shows good separability, yielding higher likelihood ratios, while Cluster 3 exhibits substantial overlap with other clusters.

Case example

A car struck a pedestrian, and the driver fled the scene. The pedestrian thinks the car was of brand A. The investigator contacts brand A, who specifies which binder they generally use. This binder belongs to cluster 1. Paint from the clothes of the victim was sent to the laboratory, with the question “Does the sample from the clothes belong to cluster 1?”.

The forensic scientist hereby postulates two competing hypotheses, i.e. the sample belongs to cluster 1 vs the sample belongs to any other cluster. After FTIR analysis of the sample the generated data was pre-processed. The resulting data were then entered into the PCA model described above, producing scores on the four principal components required as input for the LDA analysis.

The sample was processed using the LDA-model validated for cluster 1, and the results are shown in Figure 6-35.

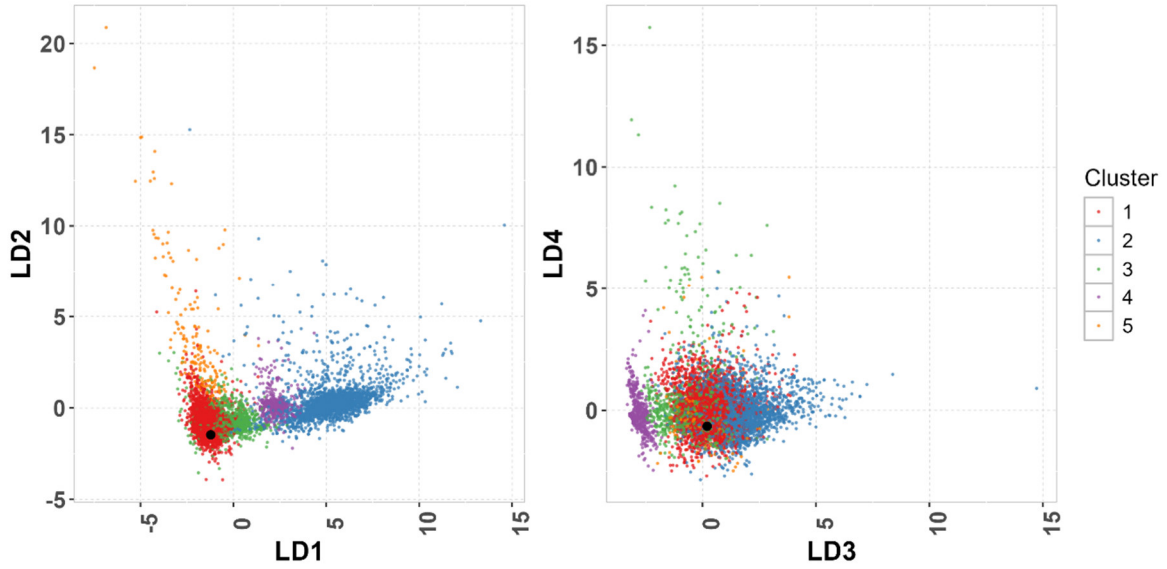

Figure 6-35. Scatter plots of LD1 vs LD2 (left) and LD3 vs LD4 (right) for cluster 1 model. The recovered sample is shown as a black dot.

LDA-analysis returned the posterior probability for cluster 1, in this example it was estimated to 93.68%. To obtain a likelihood ratio, we then apply the operation described above:

$$\text{odds}_{\text{po}} = \frac{0.9368}{1-0.9368} = 14.8 \qquad (6.5.6)$$

As the prior odds was set to 1:1 the likelihood ratio will equal the posterior odds. As the likelihood ratio is within the boundaries of the LR system, the conclusion from the forensic analysis will be that it is 14.8 times more likely to get this result given that the sample belongs to cluster 1 compared to that it belongs to any other class.

7. ASSESSMENT OF THE APPLICABILITY OF AN LR SYSTEM IN CASEWORK

Chapter 5 focused on the empirical validation of LR systems, while this chapter will address the scope of the validation. Scope is the range of conditions for which the method has been tested, and refers to the consideration: is the LR system applicable to the specific case at hand? In other words, even a valid and well-functioning LR system will produce unreliable results if provided with poor-quality or inappropriate input data. The scope should be defined prior to the validation of the LR system (Meuwly et al, 2017).

The expertise of the forensic chemist is essential for this task. The assessment of applicability demands knowledge of both chemical and statistical aspects that may influence the LR system. Therefore, the forensic practitioner is responsible for understanding the scope and limitations of the applied LR system. Blindly trusting the system's output without considering its scope is never advisable.

We will discuss the constraints associated with using the LR systems described in the diesel oil example, see Section 6.2.

7.1. Diesel oil example

The data used to construct the LR models for oil comparison (score- and feature-based) consist of the ratios between the heights of selected chromatographic peaks found in unweathered diesel oils. The forensic chemist involved in an oil comparison case should therefore consider:

Type of oil samples in the case: The database contains data only from diesel oils. If the case refers to e.g. lubricating oils, the LR model is not well suited. Diesel and lubricating oils are chemically very different. The peak ratios selected for the LR model will have a very different distribution for lubricating oils. Certain peaks, e.g. the paraffins C_{17} alkane and pristane, may even be absent as paraffinic substances are undesirable in such oils. The ratios from a lubricating oil sample will have a high risk of being outliers in a database of diesel oils, and the evidential strength is likely to be exaggerated due to this overestimated rarity.

Contamination of case samples: A case might involve contamination of (usually) the recovered sample by another fossil distillate product; alternatively, an unknown foreign substance may be co-eluting with a peak used for the LR system. A chemist can visually detect and adjust for the contamination. In the data driven approach, however, the contamination will create a distance between the samples that may underestimate the evidential strength.

Weathering of samples: Weathering is the effect the environment has on the oil once discharged, e.g. evaporation, biodegradation, photooxidation or the dissolution of compounds into the water phase. These processes affect oil substances differently, substantially changing the composition of the oil. Here, the database contains only unweathered oils; they have since their production been protected in enclosed volumes. There is a multitude of literature on oil weathering, and a trained forensic chemist can spot and account for weathering effects during the manual evaluation. However, the data driven LR approach will perform poorly on weathered oils as the weathering will result in a

separation between the original and the weathered samples leading to an underestimation of the evidential strength. In this case, retraining the model with weathered samples may not be the solution, as there are indefinite varying degrees of weathering. It is likely that the chemist will need to resort to machine learning methods or evaluate the case without relying on a model.

Other factors may also need to be considered, such as whether the model is trained on an instrument comparable to the one used in the case. To summarize, if the data in the database used to train the model is deemed relevant to the case, the case falls within the scope of the validation, and the LR model will provide valuable assistance. If not, the LR system should not be employed. The competence of the forensic chemist is very important here. If necessary, they can adjust the models according to their expertise, but then the models need to be revalidated.

7.2. Updating and maintaining the LR system with new data

After an LR system has been validated and employed new relevant data may arrive to assist the system. It is generally beneficial not to add new data to the method continuously, as new data require re-validation, but rather do an LR-system update e.g. on a yearly basis.

7.3. What to do with available information not included in the LR system

It is probably more of a rule than an exception that an LR system uses only a portion of the information available to calculate a likelihood ratio. In the oil example above, the LR system encompasses selected ratios, but one might also want to include the overall shape and the fine structure of the chromatogram.

Another example may be a glass comparison where the chemist may have an LR system for the refractive index, but also wants to incorporate the elemental composition, for which no system may be readily available. The system proposes an LR for the refractive index, and then the forensic chemist arrives at a separate LR for the elemental composition based on professional beliefs. How are these two LRs combined? One might be inclined to multiply the LRs to reach a final conclusion. However, the principle behind multiplying likelihood ratios is that the two investigative results are independent, meaning that each result is not influenced by the other. Multiplying LRs is generally not valid if there are dependencies between the factors being considered.

When there are dependencies between the factors, the joint LR needs to account for the dependence between the pieces of evidence. In such cases, a more complex statistical approach is required to properly handle the dependencies, such as using conditional LRs or Bayesian networks. A more complex, but joint, statistical approach requires a substantial amount of data on both refractory index and elemental composition from the same samples. When such composite data is not available, the LRs cannot be mathematically joined.

8. HOW TO COMMUNICATE LR RESULTS?

An assigned likelihood ratio is a numerical value. A numerical value itself says nothing without a contextual interpretation. Therefore, it must always be communicated together with the two hypotheses involved with its assignment.

As an example, suppose that one hypothesis, H_1, states that a recovered paint flake on a hit car comes from a specific car suspected to have made the hit and that the alternative hypothesis, H_2, states that the paint flake has another origin. The assigned likelihood ratio for H_1 against H_2 is 450. This can be communicated as

“The forensic findings are 450 times more probable if the recovered paint flake comes from the specific car than if it has another origin”

This way of reporting demonstrates that a likelihood ratio of 450 has no absolute interpretation. It can be compared to other likelihood ratios with respect to the strength of the forensic findings, but standing alone it cannot be substituted for, say, a probability that the recovered paint flake comes from the specific car.

However, despite the fact that communication of the reported likelihood ratio as described is technically correct, it may be difficult for a factfinder (e.g. a judge) to understand. While a chemist may understand quantitative analysis, a lawyer has no training or professional experience in quantitative reasoning. One cannot expect a judge to possess a numerically expressed prior probability of the hypothesis supported by the likelihood ratio (i.e. like the hypothesis H_1 above). Hence, the judge cannot numerically multiply the reported likelihood ratio with prior odds to obtain posterior odds (and subsequently a posterior probability) for the hypothesis supported.

Instead of reporting the likelihood ratio in terms of how much more (or less) probable the forensic findings are if one hypothesis is true compared to if another hypothesis is true, it can be reported as to which degree the forensic findings reinforce (or weaken) the current evidence for the hypothesis. As an example, a likelihood ratio of 450 can be interpreted as "the current amount of evidence for the hypothesis in question is reinforced by a factor of 450". A likelihood ratio of 0.20 can be interpreted as "the current amount of evidence is weakened by a factor of 0.20" (weakened to a fifth of its current strength). However, reporting degrees of reinforcement as pure numbers is still problematic if there is no corresponding number to multiply with. Therefore, it is wise to report the magnitude of the likelihood ratio instead of its specific numerical value. A likelihood ratio is always assigned based on the information available at that time. Should more information be added, the likelihood ratio may change. Hence, a reported value of 450 may the next day be 460 or 440 if new information has been acquired. The differences, though, usually have no significant effects on the evidentiary strength, but that may be hard to understand without daily familiarity with probability calculations.

To report the magnitude of the likelihood ratio, an ordinal scale may be used. There are several possibilities to construct such a scale, but examples may be found in the literature, e.g. in (Nordgaard et al, 2012). The ENFSI guideline on evaluative reporting (ENFSI, 2015) discusses the use of a scale and its implications. It is common practice to implement a scale that is common for all likelihood ratio assignments within a laboratory. This is so since a factfinder otherwise cannot compare the strength of evidence with different forensic findings.

The levels of an ordinary scale correspond to intervals of likelihood ratios. In Table 8-1 is shown an example of such a scale.

Level	Interval of likelihood ratios
A	$1\,000\,000 \leq LR$
B	$10\,000 \leq LR < 1\,000\,000$
C	$100 \leq LR < 10\,000$
D	$10 \leq LR < 100$
E	$2 \leq LR < 10$
F	$0.5 < LR < 2$
G	$0.1 < LR \leq 0.5$
H	$0.01 < LR \leq 0.1$
I	$LR \leq 0.01$

Table 8-1. An example of an ordinal scale with scale levels corresponding to intervals of likelihood ratios.

The scale in Table 8-1 divides the interval from 2 to infinity into 6 subintervals. This division signifies differences in magnitude of a likelihood ratio, why it can be communicated like “LR is at least 2”, “LR is at least 10”, etc. Naturally, when passing the border between two levels the magnitude is the same just below this border as just above it. Hence, if the assigned likelihood ratio falls just above a border, it shall be considered whether the level used should rather be the level below that border. This depends on the amount of information underpinning the assignment of the likelihood ratio. If that information is robust in the sense that it is not expected that additional information will change the likelihood ratio for it to fall below the current border, then the level above the border should be chosen. Otherwise, it is recommended to choose the level below the border.

The latter choice may be criticised in that it deflates the reported magnitude. Without any specification of the actual numerical value of the likelihood ratio, a recipient of the report may underestimate the strength of the forensic findings. In particular, if the likelihood ratio supports an incriminating hypothesis, it can be considered as weak by a defence lawyer. On the other hand, for many forensic findings it is difficult to assign the likelihood ratio with high resolution. Hence, it is merely about being convinced that the likelihood ratio is in a specific interval, or rather that it is above a border (lower interval limit).

In the scale in Table 8-1 there is one level, F, corresponding to “neutral evidence”. Strictly, neutral evidence would give a likelihood ratio equal to one, but it should be considered an exception to having assigned a likelihood ratio to be exactly one. Using an interval around one reflects that evidence that is practically neutral may still imply assigned likelihood ratios both slightly above one and slightly below one, but they should in such cases not be interpreted as supporting any of the two hypotheses involved.

The levels G and I in Table 8-1 comprise likelihood ratios that do not support the hypothesis of interest (e.g. H_1 above), but its alternative. A scale with levels on both sides of the likelihood ratio being equal to one is implemented when the conclusion from the forensic investigation always should address the question put to the forensic chemist. For instance, the question “Do these two seizures of heroin have a common origin?” has only two possible true answers, “yes” or “no”. A likelihood ratio (clearly) above one would then be on the “yes”-side and a likelihood ratio (clearly) below one would be on the “no”-side. It can be assumed that the commissioner has reason to suspect that the two seizures have a common origin, and a reported scale level among A-E is then interpreted as supporting his/her suspicion, while a reported scale level among F-I does not. In Table 8-1 the number of levels on the “no”-side is lower than the number of levels on the “yes”-side. This is so since it is usually much harder to assign likelihood ratios below one, in particular when the likelihoods are expressed as probabilities. A probability cannot exceed one, so the denominator has this limit. Hence, to assign a very low likelihood ratio it must be possible to assign a very low probability of obtaining the forensic findings if the hypothesis of interest is true, i.e. the numerator probability. Most often in such cases, the findings are rare no matter which of the two hypothesis is true, but they must be rarer under the hypothesis of interest than under its alternative. Assigning the ratio them between may therefore be a very difficult task, and to divide the interval from zero to 0.01 into subintervals is therefore of no use. Having said that, there are scales in use (for instance at NFC, Sweden) that are symmetric around one.

If the forensic expert is free to choose the order of the hypotheses, it could be more practical to use a one-sided scale, i.e. a scale based on intervals larger than or equal to one. In the report it must then be clearly stated which hypothesis is supported by the assigned likelihood ratio. Such a scale is in use at NFI, The Netherlands.

9. SUMMARY

During the past decades, results and conclusions of forensic investigations have become increasingly objective, quantitative and standardized. A general example is the adoption of likelihood ratios (LRs) for evaluative reporting in various forensic disciplines. Another example is the growing use of chemometrics within the fields of forensic chemistry.

Within forensic chemistry, the following types of question are commonly encountered:

- Identification: What is this white powder?
- Multiclass classification: To which chemical group does this substance belong?
- Binary classification: Does this mixture contain any cocaine?
- Quantification: How much cocaine does this mixture contain?
- Comparison (individualisation): Do these two mixtures have an identical source?

Some of these question types can be answered using chemometrics, some using LRs, and some using a combination of these.

Open questions (identification and quantification) are in general not suitable for answering using LRs. When a suitable analytical method is available and the research material is of sufficient quality, identification questions can in practice often be answered with certainty. The answers to quantification questions still involve a certain amount of uncertainty; they usually consist of a best estimate in combination with an uncertainty interval around it. A part of the classification questions can possibly also be answered with certainty (‘categorical classification’), again given a suitable method and sufficient material quality. A previous guideline discusses how chemometrics can help to investigate and answer these types of questions, using several practical examples.

The current guideline describes how the combination of chemometrics and LRs can be used to answer the remaining types of questions. Answers to another part of the classification questions still involve expressing uncertainty (‘probabilistic classification’). These answers should involve quantification of the uncertainty, preferably using an LR. Comparison questions are in general to be answered using LRs; they typically involve additional uncertainty associated with the alternative scenario that is considered. This guideline provides general theoretical background on LRs, and it explains how to construct and validate LR-systems. Several practical examples are presented.

The first example is about the classification of cannabis plants. Based on GC-MS peak ratios, a feature-based LR-system is implemented in order to decide whether the type of hemp is THC-dominant, THC/CBD-intermediate or CBD-dominant.

Example two illustrates the use of LRs for the comparison of diesel oil samples. Based on ten peak ratios from GC-MS measurements, both a feature-based and a score-based approach are used to develop an LR-system. The performance of these approaches is compared, and both are applied to a case example.

The third example focusses on comparing (weathered) mineral oil samples using LRs. Instead of using measurements as background data, the LR-system is based on expert knowledge that is captured in numbers using an elicitation approach.

Example four is about heroin seizures. First, clustering by total composition in combination with selected impurities is used to identify different groups of seizures. Based on the identified clusters, individual samples are tested to form a joint group with one of the clusters Next, LRs are used to compare two heroin seizures, quantifying whether their similarities are more probable when they belong to the same group or when they belong to two different groups.

The fifth example considers measurements on car paints classification. The analysis is based on spectral data (FT-IR). First clustering is used to determine a proxy ground truth as to several groups of car paints. Then an LR model is trained for assigning a new paint sample (object) to a specific cluster.

After this, applicability of LR-systems in casework is discussed, together with how to communicate the results to a factfinder.

REFERENCES

Ahrens B, Alberink I, Bovens M, Huhtala S, Malmborg J, Nordgaard A, Salonen T, 2021, Guideline for the use of Chemometrics in Forensic Chemistry, https://enfsi.eu/wp-content/uploads/2021/09/Guideline-for-the-use-of-Chemometrics-in-Forensic-Chemistry.pdf.

Aitken C.G.G, Taroni F, Bozza S, 2020, Statistics and the evaluation of evidence for forensic scientists, 3rd ed. Wiley.

Alberink I, Leegwater A.J, Malmborg J, Nordgaard A, Sjerps M, van der Ham L, 2026, A transparent method to determine limit values for likelihood ratio systems, Law, Probability and Risk.

Berger C, 2010, Criminalistics is reasoning backwards, Nederl. Juristenbl. 85, pp. 784-789.

Bolck A, Ni H, Lopatka M, 2015, Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison, Law, Probability and Risk, vol. 14, no. 3, pp. 243-266.

Bovens M, Ahrens B, Alberink I, Nordgaard A, Salonen T, Huhtala S, 2019, Chemometrics in forensic chemistry — Part I: Implications to the forensic workflow, For. Sci. Int. 301, pp. 82–90.

Brümmer N, du Preez J, 2006, Application-independent evaluation of speaker detection, Comput. Speech Lang. 20 (2), pp. 230–275, https://doi.org/10.1016/j.csl.2005.08.001.

Dahlman C, 2018, Beviskraft: Metod för bevisvärdering i brottmål, 1 ed., Norstedts Juridik AB.

Duarte J.M, Sales N.G.S, Braga J.W.B, Bridge C, Maric M, Sousa M.H, Gomes J, 2022, Discrimination of white automotive paint samples using ATR-FTIR and PLS-DA for forensic purposes, Talanta 240, 123154. doi: https://doi.org/10.1016/j.talanta.2021.123154.

European Network of Forensic Science Institutes (ENFSI), 2020, Best Practice Manual for controlled drug analysis, BPM DWG-CDA-001-001, https://enfsi.eu/wp-content/uploads/2017/06/BPM-Control-drug-Analysis-final-version.-21-02-2020.pdf

European Network of Forensic Science Institutes (ENFSI), 2016, Drug working group (DWG), Guideline on quantitative Analysis of seized drugs, https://enfsi.eu/wp-content/uploads/2016/09/guidelines_quant_sampling_dwg_printing_vf4.pdf.

European Network of Forensic Science Institutes (ENFSI), 2024, Guideline for calculating measurement uncertainty in quantitative forensic investigations, https://enfsi.eu/wp-content/uploads/2024/04/Guideline-for-Calculating-Measurement-Uncertainty-in-Quantitative-Forensic-Investigations-1.pdf.

European Network of Forensic Science Institutes (ENFSI), 2015, Guideline for evaluative reporting in forensic science. Strengthening the evaluation of forensic results across Europe (STEOFRAE).

European Network of Forensic Science Institutes (ENFSI), 2022, GUIDELINE for the forensic examination of paint by Fourier-transform infrared spectroscopy, https://enfsi.eu/wp-content/uploads/2022/06/EPG-GUIDELINE-002-Guideline-for-the-forensic-examination-of-paint-by-FT-infrared-spectroscopy.pdf

Evett I.W, 1998, Towards a uniform framework for reporting opinions in forensic science casework, Science & Justice 3, pp. 198-202.

Fasasi A, Mirjankar N, Stoian R.I, White C, Allen M, Sandercock M.P, Lavine B.K, 2015, Pattern recognition-assisted infrared library searching of automotive clear coats, Applied Spectroscopy 69, pp. 84-94.

Huhtala S, Nordgaard A, Ahrens B, Alberink I, Salonen T, Bovens M, 2023, Chemometrics in Forensic Chemistry – Part III: Assessment and interpretation of results, Forensic Sci Int 348:111612, doi: 10.1016/j.forsciint.2023.111612.

Ishihara S, Carne M, 2022, Likelihood ratio estimation for authorship text evidence: An empirical comparison of score- and feature-based methods, Forensic Sci Int. 334, doi.org/10.1016/j.forsciint.2022.11126.

ISO 17025 Testing and calibration laboratories. https://www.iso.org/obp/ui/en/#iso:std:iso-iec:17025:ed-3:v1:en

ISO 21043 Forensic Science Part 3 Analysis and Part 4 Interpretation. https://www.iso.org/obp/ui/en/#iso:std:iso:21043:-3:ed-1:v1:en and https://www.iso.org/obp/ui/en/#iso:std:iso:21043:-4:ed-1:v1:en

Jonsson C.S.L, Strömberg L, 1993, Computer Aided Retrieval of Common-Batch Members in Leuckart Amphetamine Profiling, J Forensic Sci 38, 1472-1477, doi: 10.1520/JFS13553J.

Kwofie F, Perera U.D.N, Allen M.D, Lavine B.K, 2018, Transmission infrared imaging microscopy and multivariate curve resolution applied to the forensic examination of automotive paints, Talanta 186, pp. 662-669.

Kwofie F, Perera U.D.N, Allen M.D, Lavine B.K, 2020, Application of infrared microscopy and alternating least squares to the forensic analysis of automotive paint chips, Journal of Chemometrics 35, e3277.

Lavine B. K, Fasasi A, Mirjankar N, Sandercock M, 2014, Development of search prefilters for infrared library searching of clear coat paint smears, Talanta 119, 331-340. doi: https://doi.org/10.1016/j.talanta.2013.10.066

Leegwater A.J, Vergeer P, Alberink I, van der Ham L.V, van de Wetering J., El Harchaoui R, Bosma W, Ypma R.J.F, Sjerps M.J, 2024, From data to a validated score-based LR system: A practitioner's guide, For. Sci. Int. 357, 111994.

van Lierop S., Ramos D, Sjerps M, Ypma R, 2024, An overview of log likelihood ratio cost in forensic science–Where is it used and what values can we expect?, Forensic Science International: Synergy 8, 100466. doi: https://doi.org/10.1016/j.fsisyn.2024.100466.

Malmborg J, Kooistra K, Kraus U.R, P. Kienhuis P, 2020, Evaluation of light petroleum biomarkers for the 3rd edition of the European Committee for Standardization methodology for oil spill identification (EN15522-2). Environmental Forensics, pp. 1-15.

Malmborg. J, Nordgaard A, 2016, Forensic characterization of mid-range petroleum distillates using light biomarkers, Environmental Forensics 17, 244-252. doi: https://doi.org/10.1080/15275922.2016.1177758https://doi.org/10.1080/15275922.2016.1177758

Malmborg J, Nordgaard A, 2021, Validation of a feature-based likelihood ratio method for the SAILR software. Part I: Gas chromatography–mass spectrometry data for comparison of diesel oil samples. Forensic Chemistry 26:100375.

Meuwly D, Ramos D, Haraksim R, 2017, A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. Forensic science international 276, pp.142-153.

Nordgaard A, Ansell R, Drotz W, Jaeger L, 2012, Scale of conclusions for the value of evidence, Law, Prob. & Risk 11. doi: https://doi.org/10.1093/lpr/mgr020

Perera U.D.N, Nishikida K, Lavine B.K, 2018, Development of infrared library search prefilters for automotive clear coats from simulated attenuated total reflection (ATR) spectra, Applied Spectroscopy 72, pp. 886-895.

Ramos D, Meuwly D, Haraksim R, 2020, C.E.H. Berger, Validation of Forensic Automatic Likelihood Ratio Methods, in: Handbook of Forensic Statistics 1st Edition, CRC Press, Boca Raton, USA.

Ramos D, Gonzalez-Rodriguez J, 2013, Reliable support: Measuring calibration of likelihood ratios. Forensic science international 230, pp. 156-169.

R Core Team, 2020, R: A language and Environment for Statistical Computing, Vienna, Austria: R Foundation for Statistical Computing. Available at https://www.R-project.org/.

Rice J.A, 2007, Mathematical Statistics and Data Analysis, third edition, Thomson.

Salonen T, Ahrens B, Bovens M, Eliaerts J, Huhtala S, Nordgaard A, Alberink I, 2020, Chemometrics in forensic chemistry — Part II: Standardized applications – Three examples involving illicit drugs, Forensic Science International 307, 110138.

Sarma N. D. A. Waye, M. A. ElSohly, P. N. Brown, S. Elzinga, H. E. Johnson, R. J. Marles, J. E. Melanson, E. Russo, L. Deyton, C. Hudalla, G. A. Vrdoljak, J. H. Wurzer, I. A. Khan, N. Kim, and G. I. Giancaspro I, 2020, Cannabis Inflorescence for Medical Purposes: USP Considerations for Quality Attributes, Journal of Natural Products 83 (4), pp. 1334–1351.

Silverman B.W, 1998, Density Estimation for Statistics and Data Analysis, Chapman & Hall/CRC, Boca Raton.

Staginnus C, Zörntlein S, de Meijer E, A PCR marker Linked to a THCA synthase Polymorphism is a Reliable Tool to Discriminate Potentially THC-Rich Plants of Cannabis sativa L., Forensic Sci, July 2014, Vol. 59, No. 4doi: 10.1111/1556-4029.12448

Strömberg L, Lundberg L, Neumann H, Bobon B, Huizer H, van der Stelt N. W, 2000, Heroin impurity profiling. A harmonization study for retrospective comparisons, Forensic Sci Int. 114(2): pp. 67-8, doi: 10.1016/s0379-0738(00)00295-4.

SWGDRUG Recommendations for the Analysis of Seized Drugs, 2024, https://www.swgdrug.org/Documents/SWGDRUG%20Recommendations%20Version%208.2%20Final%20for%20posting.pdf

Taroni F, Biedermann A, Bozza S, Garbolino P, Aitken C, 2014, Bayesian networks for probabilistic inference and decision analysis in forensic science, 2nd ed. Wiley.

Taroni F, Biedermann A, Aitken C, 2023, Statistical Interpretation of Evidence: Bayesian Analysis, Encyclopedia of Forensic Sciences, 3rd ed., Max M. Houck (Ed.), Oxford: Elsevier, pp. 656-663.

United Nations Office on Drugs and Crime (UNODC), 2005, Methods for Impurity Profiling of Heroin and Cocaine, Method A4, pp. 20-21, ST/NAR/35.

United Nations Office on Drugs and Crime (UNODC), 2023, Methods for Impurity Profiling of Heroin and Cocaine, ST/NAR/35 Rev.1

United Nations Office on Drugs and Crime (UNODC), 2022, Recommended Methods for the Identification and Analysis of Cannabis and Cannabis Products, pp. 26, ST/NAR/40.1 rev.

Vergeer P, van Es A, de Jongh A, Alberink I, Stoel R, 2016, Numerical likelihood ratios outputted by LR systems are often based on extrapolation: When to stop extrapolating? Sci. & Just. 56 (6), pp. 482–491.

Warrens M.J., van der Hoef H, 2022, Understanding the Adjusted Rand Index and Other Partition Comparison Indices Based on Counting Object Pairs. J Classif 39, 487–509. https://doi.org/10.1007/s00357-022-09413-z

AUTHORS IN ALPHABETICAL ORDER

Dr. Björn Ahrens
Federal Criminal Police Office, Wiesbaden, KT-45, Äppelallee 45, D-65203 Wiesbaden, Germany
bjoern.ahrens@bka.bund.de

Dr. Ivo Alberink
Netherlands Forensic Institute, Laan van Ypenburg 6, 2497 GB, Den Haag, The Netherlands
i.alberink@nfi.nl

Dr. Michael Bovens (Chairman ENFSI Drugs Working Group Subcommittee Chemometrics)
Zurich Forensic Science Institute, P.O. Box, 8021 Zurich, Switzerland
michael.bovens@for-zh.ch

Dr. Leen van der Ham
Netherlands Forensic Institute, Laan van Ypenburg 6, 2497 GB, Den Haag, The Netherlands
l.van.der.ham@nfi.nl

Sami Huhtala (Team leader ROTOR)
National Bureau of Investigation, Tikkurilantie 30, P.O. Box 285, FIN-01301 Vantaa, Finland
sami.huhtala@poliisi.fi

Tuomas Korpinsalo
National Bureau of Investigation, Tikkurilantie 30, P.O. Box 285, FIN-01301 Vantaa, Finland
tuomas.korpinsalo@poliisi.fi

Jonas Malmborg
National Forensic Centre, Swedish Police Authority, 58194 Linköping, Sweden
jonas.malmborg@polisen.de

Dr. Anders Nordgaard
National Forensic Centre, Swedish Police Authority, 58194 Linköping, Sweden
anders.nordgaard@polisen.se

ENFSI wishes to promote the improvement of mutual trust by encouraging forensic harmonization through the development and use of Best Practice Manuals. Furthermore, ENFSI encourages sharing Best Practice Manuals with the whole Forensic Science Community which also includes non ENFSI Members.

Visit www.enfsi.eu/documents/bylaws for more information. It includes the ENFSI framework document Framework for the Process for the Creation of Technical Documents within ENFSI (code: QCC-FWK-001).

doi:10.5281/zenodo.21370909