Analytical Methods

RSC Publishing

PAPER

View Article Online
View Journal | View Issue

Cite this: *Anal. Methods*, 2013, **5**, 6788

The evidential value of microspectrophotometry measurements made for pen inks

Agnieszka Martyna,[a] David Lucy,[b] Grzegorz Zadora,*[c] Beata M. Trzcinska,[c] Daniel Ramos[d] and Andrzej Parczewski[ac]

Received 18th September 2013
Accepted 20th September 2013

DOI: 10.1039/c3ay41622d

www.rsc.org/methods

Three colour systems, defined by the International Commission on Illumination (CIE), have been used to parametrise spectra from microspectrophotometry in the visible range for ten replicates of each of forty inks. The parametrised spectra were used to calculate a likelihood ratio (LR) for pairwise comparisons under the propositions implying that any ink from some suspect document came from the same pen, as that from a control document, *versus*, the converse proposition which implies the ink from the suspect document came from some other pen. Both univariate and bivariate likelihood ratios for each colour system were calculated. Empirical cross-entropy was selected as an appropriate measure of performance for each system. The bivariate combinations of the CIE-*xyz* colour system achieve the best results as well as a bivariate combination of *a* and *b* variables within the CIE-*Lab* colour system.

1. Introduction

1.1. Analysis of inks for forensic purposes

Inks are routinely examined in cases where suicide notes, forged wills, blackmail notes, and documentation from other more general offences of fraud may form part of the evidence. In many cases inks may be used as part of a comparative analysis, whereby an ink from some suspect document may be compared to some ink from a control document for evidence that the inks on the two documents came from the same pen, printer, or other source of ink.[1]

This is considered to be a "comparison problem", and is commonly used for analytical chemistry in forensic science. Various instrumental techniques have been employed to make these comparisons. Examples are high performance thin layer chromatography (HPTLC),[2] high performance liquid chromatography (HPLC),[3] time-of-flight secondary ion mass spectrometry (TOF-SIMS),[4] and laser desorption ionisation-mass spectrometry (LDI-MS),[5] desorption electrospray ionisation (DESI)[6] and direct analysis in real time (DART).[7,8]

These techniques provide more information than non-destructive techniques, however, they disrupt the integrity of the sample. As a consequence, the evidence is not available for further testing or review at a later stage in the investigation. Therefore, in the forensic sciences, non-destructive methods, such as optical techniques are preferred. Attenuated total reflectance (micro-ATR-FTIR) is a non-destructive measurement method for Fourier transform infrared (FT-IR) spectroscopy. It has been used to estimate the sequence of intersecting lines,[9] and to analyse blue ballpoint pen inks,[10,11] as well as red seal inks.[12] Raman spectroscopy can be useful for characterising and discriminating between inks based on their composition,[13,14] and microspectrophotometry in the visible range (MSP-Vis; 380–800 nm), which enables the chemist to estimate the colour of objects in an objective and reproducible way.

1.2. Colour analysis

A measurable feature which varies between all objects is their colour.[15] In humans colour recognition is only possible due to the eye's ability to receive, and the human brain to interpret light stimuli.

To eliminate such subjectivity, attempts to introduce standardisation of colour notion using a 2° (two degrees) standard observer were made in 1931. The notation 2°, along with the later introduced 10° observer (introduced in 1964), refers to the viewing angle under which the standard observer sees the colour. The International Commission on Illumination (CIE) defines standards for colour transmittance, or reflection, for any illumination/measurement system where a measured spectrum is weighted by some standard colour, under standard conditions. This is undertaken for the standard primary colours red, green and blue, which are known as the tristimulous values of the CIE-*XYZ* system. The tristimulous values *X*, *Y* and *Z* can be used to provide co-ordinates in a three dimensional colour space. However, standardised values, termed the chromaticity coordinates, can also be used to form the CIE-*xyz* system. As an alternative, a set of weightings of the tristimulous values has been used to produce the CIE-*Lab* system, where the first feature

[a]*Jagiellonian University, Faculty of Chemistry, Ingardena 3, 30-060 Krakow, Poland*
[b]*Department of Mathematics & Statistics, Lancaster University, Lancaster, LA1 4YF, UK*
[c]*Institute of Forensic Research, Westerplatte 9, PL-31-033 Krakow, Poland. E-mail: gzadora@krakow.ies.pl; Tel: +48 124228755*
[d]*Escuela Politecnica Superior, Universidad Autonoma de Madrid, Calle Francisco Tomas y Valiente 11, 28049 Madrid, Spain*

is lightness (L) ranging from black ($L = 0\%$) to white ($L = 100\%$) and the other two parameters a and b are the chromatic values changing from green to red and blue to yellow, respectively.

Microspectrophotometry in the visible range is often used for the analysis of writing materials. It is a non-destructive technique. It is generally used in transmission mode, with writing materials being extracted from the base, and forming relatively thin and representative samples. It can also be used in reflectance mode, which has been recommended since it is non-destructive.[1,16–18]

1.3. Evaluation of the evidential value of MSP-Vis spectra

Analytical methods return some output which is related to the chemistry of the specimen being examined. However, the chemical composition of a specimen of ink is by itself, in a forensic context, meaningless, and needs a further stage of interpretation. The Scientific Working Group for Materials Analysis (SWGMAT)[19] suggests:

Spectra can be compared by overlaying them on a light box or by plotting them on the same graph. Mean-value spectra, generated from several scans of each sample and bracketed by curves showing +/− one or more standard deviations of the mean sets, also are suitable for comparison. Three standard deviation units will cover approximately 99 percent of the expected sample variation under normal distribution conditions.

Any procedure based upon visual comparison is arbitrary in its choice of uncertainty and gives no idea of the nature, or degree of similarity beyond some coarse grained verbal scale and it is difficult to translate into a form of evidential value.

A more complete interpretation entails an evaluation of the physicochemical observations, that is the evidence, denoted E, in the context of a proposition implied by the prosecution case, H_p, and a proposition implied by the defence case, H_d. The objective being to estimate the conditional probabilities $\Pr(E|H_p)$ and $\Pr(E|H_d)$. The ratio of these two conditional probabilities is termed the likelihood ratio, and is taken as a measure of evidential value for the spectral observations. For instance, some document written by a specific individual using a particular pen could be compared to some ink upon a questioned document to provide evidence of fraud. In this instance the source level propositions might be:

$H_p \equiv$ the ink from the recovered item and the ink from the control item are from the same pen,

$H_d \equiv$ the ink from the recovered item is from some pen, from the relevant population of pens, other than that pen which left the ink on the control item.

The evidential value of these spectra requires that some attention is given to the following points:

(a) similarity of the features observed for those objects being compared,

(b) the possible sources of uncertainty, which include at least:

1. the variation of measurements within objects,

2. the variation of measurements between those objects in the relevant population,

(c) the rarity of the observed physicochemical features,

(d) existing correlation between the features in the case of multi-dimensional data.

In contrast with simpler proximity based measures of evidential value as exemplified by the Scientific Working Group for Materials Analysis[19] approach above, the likelihood ratio for the physicochemical data requires some knowledge about the rarity of the measured physicochemical properties in the relevant population, which in this case is the population of blue pen inks. The value of the evidence in support of the proposition that the compared samples have a common origin is greater when the determined values are similar and rare in the relevant population than when the physicochemical values are equally similar but common in the same population.

A likelihood ratio (LR) approach allows the forensic scientist to include all these factors in a single, easily comprehended calculation. A likelihood ratio is a widely accepted measure of evidence value in the forensic sciences, frequently being employed to evaluate the observations of forensic scientists.[20]

A likelihood ratio can be calculated as:

$$\mathrm{LR} = \frac{\Pr(E \mid H_p)}{\Pr(E \mid H_d)}. \quad (1)$$

In the case of continuous observations $\Pr(\cdot)$ are evaluated from suitable probability density functions $f(\cdot)$. Values of the likelihood ratio above unity support the prosecutor's hypothesis (H_p), while values of LR below one support the defence hypothesis (H_d). Values of the likelihood ratio equal to one support neither of the hypotheses. In general terms the higher the value of likelihood ratio, the stronger support for the prosecution proposition (H_p). The lower the value of likelihood ratio, the stronger support for the defence proposition (H_d).

Previous studies on the use of chemometric methods to the differentiation of writing materials have been conducted for black inks by Adam,[21,22] and Thanasoulias[23] for blue inks. These analyses focused not only on comparing the samples of various ballpoint pens' brands, but also contrasting those produced by the same manufacturer. Principal Component Analysis (PCA) was used for the interpretation of the analytic output, but in these studies some component representing rarity was not included, so such work does not fulfil the requirements for the evaluation of the evidential value of these physicochemical data.[24]

The similar problem occurs with the so-called *two-stage* approach. The two-stage approach uses statistical hypothesis tests, such as Student's *t*-test, and Hotelling's test, which are easily available in commercial statistical software. Although useful in the right circumstances, significance tests have the problem that they take into account only information about within-object variation and the similarity of the compared items, and can be regarded as a subset of a wider class of proximity based measures of evidential value,[25,26] with similarities to the direct comparison of drawn spectra recommended by SWGMAT. Proximity based methods offer solutions to questions of similarity and dissimilarity between compared items, if the objects are deemed dissimilar, then the analysis is stopped and it is decided to act as if the two pieces of evidence came from different sources. When the samples are deemed similar, the second stage is the assessment of the rarity of the evidence. This stage is usually

undertaken in a rather subjective way. Other problems with classical approaches of evidence evaluation for forensic purposes have been described in the literature.[25,27]

Lindley's[28] detailed exposition clearly shows that the evidential value for observations made on the continuous scale of measurement, within a framework of source level propositions, depends not only upon the proximity between the items, but critically, also the rarity of those proximities. Developments of these likelihood ratio ideas extend to where the observed features may be multivariate, and the distribution between items not necessarily normal,[29] where the observations may contain high numbers of structural zeros,[30,31] more than two levels of variation within the observations[32] and means of dealing with relatively high dimensional observations.[33] Berger[34] employed a bivariate likelihood ratio to compare inks from pens. The ink colours were described by the red, green, and blue colour components (RGB) from images from a high resolution scanner. This technique of colour analysis is extremely interesting, however it does not allow the forensic scientist to analyse those very small areas of ink which are frequently met in forensic practice.

Within most legal jurisdictions there is some requirement to evaluate the performance of each analytical and statistical approach.[35] This is not only a need for the measurement of the discriminating power of the methods, as represented by false positive and false negative rates, but for the information that the model provides to the inference process in evidence evaluation. This cannot yet be made for significance tests, other proximity based measures of evidential value, but it can be made for likelihood ratios by the calculation of the empirical cross-entropy (ECE)[36–38] (see also Section 2.2.3).

This paper describes a likelihood ratio approach for MSP-Vis spectra to be used to compare items in a quantitative and reproducible way. Empirical cross-entropy has been used to provide an assessment of the relative performances of each of the three colour systems examined here.

2. Materials and methods

2.1. Sample preparation and the ink database

40 blue inks (36 ballpoint and 4 gel) were analysed. They came either from the Polish market, or were gifts presented to the Institute of Forensic Research. Lines were made by drawing on white printing paper (80 g m^{-2}, A4). A fragment of the paper was then cut and fixed to a microscope base slide, and placed on the stage of the microscope with the MSP instrument.

Measurements were made upon blue inks using a microspectrophotometer (MSP) Zeiss Axioplan 2 with a J&M Tidas Diode Array Detector (DAD; MCS/16 1024/100-1, Germany), which was configured for the VIS range (380–800 nm) analyses. The inks were measured in reflection mode using an integration time of 2.5 seconds, the magnification being 400×. For each ink ten measurements were made using a diaphragm to select each area. The paper type was not thought to affect the spectra a great deal since very few of the wavelengths of interest were below 420 nm.[39] Each MSP spectra was parametrised in terms of a co-ordinate in a three colour system, an approach which has been applied to other MSP spectra of forensic interest.[40,41]

The software used was Spectralys version 1.82 from J&M Tidas, which provided the tristimulous values CIE-*XYZ*, chromaticity coordinates CIE-*xyz*, and CIE-*Lab* values for the 2° standard observer. **R**[42] was used to calculate both the likelihood ratio, and ECE values.

2.2. Numerical methods

2.2.1. Likelihood ratios. Values of the relevant parameters for each system considered in this paper were measured for m objects, each measured n times, in the form of p-dimensional vectors, where p is the number of observed CIE features. This can be notated as: $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^{\mathrm{T}}$, where $i = 1, \ldots, m$ and $j = 1, \ldots, n$; where $\overline{\mathbf{x}}_i = \frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_{ij}$.

If the observations from the control ink item can be denoted as $\mathbf{y}_{1j} = (y_{1j1}, \ldots, y_{1jp})^{\mathrm{T}}$, and the observations from the recovered item as $\mathbf{y}_{2j} = (y_{2j1}, \ldots, y_{2jp})^{T}$, then let $\overline{\mathbf{y}}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1}\mathbf{y}_{1j}$ be the mean of n_1 observations from the control item, and $\overline{\mathbf{y}}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2}\mathbf{y}_{2j}$ be the mean of the set of n_2 measurements from the recovered item.

The numerator from eqn (1), which implies that both means for recovered and control objects are equal, is given as eqn (2). Assuming that the between ink data distribution is normal,[29,31,41] the denominator from eqn (1), under H_{d} which implies that the means differ, is given as eqn (3).

Numerator:

$$\begin{aligned} f(\overline{\mathbf{y}}_1, \overline{\mathbf{y}}_2 | \mathbf{U}, \mathbf{C}, \overline{\mathbf{x}}, H_{\mathrm{p}}) &= f(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2, \overline{\mathbf{y}}^* | \mathbf{U}, \mathbf{C}, \overline{\mathbf{x}}, H_{\mathrm{p}}) \\ &= (2\pi)^{-p}\left|\frac{\mathbf{U}}{n_1} + \frac{\mathbf{U}}{n_2}\right|^{-\frac{1}{2}} \\ &\times \exp\left\{-\frac{1}{2}(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2)^T\left(\frac{\mathbf{U}}{n_1} + \frac{\mathbf{U}}{n_2}\right)^{-1}(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2)\right\} \\ &\times \left|\frac{\mathbf{U}}{n_1 + n_2} + \mathbf{C}\right|^{-\frac{1}{2}} \\ &\times \exp\left\{-\frac{1}{2}(\overline{\mathbf{y}}^* - \overline{\mathbf{x}})^T\left(\frac{\mathbf{U}}{n_1 + n_2} + \mathbf{C}\right)^{-1}(\overline{\mathbf{y}}^* - \overline{\mathbf{x}})\right\}, \end{aligned} \quad (2)$$

Denominator:

$$\begin{aligned} f(\overline{\mathbf{y}}_1, \overline{\mathbf{y}}_2 | \mathbf{U}, \mathbf{C}, \overline{\mathbf{x}}, H_{\mathrm{d}}) &= f(\overline{\mathbf{y}}_1 | \mathbf{U}, \mathbf{C}, \overline{\mathbf{x}}, H_{\mathrm{d}}) f(\overline{\mathbf{y}}_2 | \mathbf{U}, \mathbf{C}, \overline{\mathbf{x}}, H_{\mathrm{d}}) \\ &= (2\pi)^{-p}\left|\frac{\mathbf{U}}{n_1} + \mathbf{C}\right|^{-\frac{1}{2}} \\ &\times \exp\left\{-\frac{1}{2}(\overline{\mathbf{y}}_1 - \overline{\mathbf{x}})^T\left(\frac{\mathbf{U}}{n_1} + \mathbf{C}\right)^{-1}(\overline{\mathbf{y}}_1 - \overline{\mathbf{x}})\right\} \\ &\times \left|\frac{\mathbf{U}}{n_2} + \mathbf{C}\right|^{-\frac{1}{2}} \\ &\times \exp\left\{-\frac{1}{2}(\overline{\mathbf{y}}_2 - \overline{\mathbf{x}})^T\left(\frac{\mathbf{U}}{n_2} + \mathbf{C}\right)^{-1}(\overline{\mathbf{y}}_2 - \overline{\mathbf{x}})\right\}, \end{aligned} \quad (3)$$

where $\bar{\mathbf{x}}$ is a vector of the overall means of p variables estimated using n measurements for m objects from the database:

$$\bar{\mathbf{x}} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{m}\mathbf{x}_{ij}.$$

The weighted mean $\bar{\mathbf{y}}^*$ is given as: $\bar{\mathbf{y}}^* = \frac{n_1\bar{\mathbf{y}}_1 + n_2\bar{\mathbf{y}}_2}{n_1 + n_2}$. More details about likelihood ratio calculations used here can be found in Section 3.

U and **C** are the within- and between-object variance-covariance matrices calculated from eqn (4) and (5).

$$\mathbf{U} = \frac{\mathbf{S}_w}{m(n-1)}, \quad (4)$$

where:

$$\mathbf{S}_w = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i\right)^T\left(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i\right),$$

and

$$\mathbf{C} = \frac{\mathbf{S}^*}{m-1} - \frac{\mathbf{S}_w}{nm(n-1)}, \quad (5)$$

where:

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

In the case of univariate observations ($p = 1$), the variance-covariance matrices become scalar variances, and vectors $\bar{\mathbf{x}}$ and $\bar{\mathbf{x}}_i$ become scalars $\bar{x}$ and $\bar{x}_i$.

2.2.2. Experimental protocol. For both the CIE-*XYZ* and CIE-*Lab* parametrisation systems there are three independent variables, however, for the CIE-*xyz* system a degree of freedom has been removed by the summation process, so the CIE-*xyz* system can be considered as having only two independently varying quantities. The calculation of likelihood ratios for a bivariate comparison[29] requires the estimation of five coefficients for the background population; that is: two means, two variances and one covariance, and is only just possible from a database having 40 samples. A trivariate likelihood ratio needs some nine coefficients estimated; three means, three variances and three covariances, which is not really possible from a sample of 40, so it was decided to restrict the dimensionality to uni-($p = 1$) and bivariate ($p = 2$) cases.

To prevent the under-estimation of the proportion of wrong decisions due to the inclusion of a set of observations in the training set from items to be compared, a jack-knife resampling strategy was employed,[43] where the populational parameters for each comparison were calculated excluding all those observations from the items to be compared.

Two experiments were conducted to establish the percentage of false positive and false negative answers. As each ink had 10 replicate observations, a likelihood ratio for the pairwise comparison between observations known to be from exactly the same pen (H_p) was calculated between the first five, and the second five replicates. As, in this instance, it is known that H_p is true, the expected likelihood ratio was greater than one. Each value of the likelihood ratio less than one for a known same source comparison was considered to be a false negative. Pairwise comparisons between pens known to be different (H_d) were made using all ten replicates. The total sample was of 40 pens, so there were 40 pairwise comparisons between inks known to be from the same source, and 780 pairwise comparisons $\left(\binom{40}{2} = \frac{40!}{2!38!} = \frac{39\times 40}{2} = 780\right)$ between inks known to be from different pens.

2.2.3. Empirical cross-entropy. The likelihood ratio is the final term in the odds form of the Bayes theorem in eqn (6):

$$\frac{\Pr(H_p \mid E)}{\Pr(H_d \mid E)} = \frac{\Pr(H_p)}{\Pr(H_d)}\cdot\frac{\Pr(E \mid H_p)}{\Pr(E \mid H_d)}. \quad (6)$$

$\Pr(H_p)$ and $\Pr(H_d)$ are called *prior probabilities* and their quotient is the *prior odds*. In the judicial process their estimation lies within the competence of the fact finders (judge, prosecutor, or police) expressing their opinions about the considered hypotheses before the evidence is considered. It is the duty of a fact finder, that is the police, or other members of the court, to determine whether the objects which form the evidence are deemed to stem from the same source, or from different sources. This decision is represented by the conditional probabilities $\Pr(H_p|E)$ and $\Pr(H_d|E)$, which are the *posterior probabilities*, their quotient being the *posterior odds*. The *posterior odds* are formed from the *prior odds* and the information delivered by the forensic expert in the form of the likelihood ratio. Given the importance of the likelihood ratio it is important that the method used for the evaluation of evidence delivers strong support for the correct proposition whichever that proposition might be; that is $LR \gg 1$ when H_p is correct, and $LR \ll 1$ when H_d is correct. Additionally, it might be considered desirable that were an incorrect proposition supported by the likelihood ratio then that likelihood ratio value should be close to 1, providing only weak misleading evidence for an incorrect proposition. A measure of this ability of an evidence evaluation method to strongly support correct propositions, and weakly support incorrect propositions is the empirical cross-entropy. The empirical cross-entropy is a more nuanced measure of performance than false positives and false negatives, and comes from information theory.[38,44] It has been used previously to evaluate likelihood ratios for forensic glass interpretation,[36,37,45] and previously employed in automated speaker recognition.[46] ECE is a measure of "loss of information" for any set of likelihood ratios, thus lower values of cross-entropy are considered better.

The empirical cross-entropy is bound up with the notion of the strictly proper scoring rule. The strictly proper scoring rule can be expressed:

(a) if H_p is true: $-\log_2 \Pr(H_p|E)$,

(b) if H_d is true: $-\log_2 \Pr(H_d|E)$.

Substituting where there are $N_s = 40$ likelihood ratios known to be from the same source, and $N_d = 780$ likelihood ratios known to be from different sources the ECE can be computed from eqn (7):

$$\mathrm{ECE} = \frac{\Pr(H_\mathrm{p})}{N_\mathrm{s}} \sum_{i=1}^{N_\mathrm{s}} \log_2\left[1 + \frac{\Pr(H_\mathrm{d})}{\mathrm{LR}_i \Pr(H_\mathrm{p})}\right] + \frac{\Pr(H_\mathrm{d})}{N_\mathrm{d}} \sum_{j=1}^{N_\mathrm{d}} \log_2\left[1 + \frac{\mathrm{LR}_j \Pr(H_\mathrm{p})}{\Pr(H_\mathrm{d})}\right]. \quad (7)$$

The ECE (eqn (7)) calculations are represented in Fig. 2. The ECE can only be calculated if some prior odds are known, so here we have used a reasonably large range of prior odds. In practice a limited, but credible range of ECE values can be calculated for any given method of calculating a likelihood ratio as in the court process the estimation of the prior odds lies within the competence of the fact finders, and that range would be selected by them.

Each plot in Fig. 2 has three lines. These correspond to:

1. the black dotted line is the ECE for the *null* likelihood ratio, that is where the observations give no information about the source of the ink, and that the likelihood ratio is always equal to one.

2. The blue dashed line is the ECE for the calibrated set. This represents a theoretical best set of ECEs for the feature set were there are no losses of information due to calibration. This is denoted as $C_{\mathrm{llr}}^{\mathrm{min}}$ for prior log odds equal to zero, and represents the likelihood ratio value sets of the best performance of all other likelihood ratio sets offering the same discriminating power, and is calculated using the Pool Adjacent Violators (PAV) algorithm.[47,48]

3. Finally, the red solid line is the ECE for the observed likelihood ratios, and represents the loss of information about the ink source from the features that have actually been observed, and has been denoted as C_{llr} for prior log odds equal to zero.

The differences between ECEs calculated from the calibrated set, that is $C_{\mathrm{llr}}^{\mathrm{min}}$ for prior log odds equal to zero (Fig. 2), and the value of the ECE for the experimental likelihood ratios, that is C_{llr} for prior log odds equal to zero, are considered to be due to the problems with the calibration of the evidence evaluation method.[38,44] If the calculated likelihood ratio values for the same and for the different source comparisons are entirely separated, the values for the calibrated set of ECEs will tend to zero, and will be some "infallible" method of assigning the ink source.

3. Results and discussion

The correlation between the features for each CIE system is quite high, this is particularly noticeable from the CIE-*XYZ* system, and illustrated by the scatterplots in Fig. 1. The *X* and *Y* components of the tristimulous CIE-*XYZ* values are especially highly correlated ($r \approx 0.91$), whereas the *X* and *Z* components, and the *Y* and *Z* components, are moderately correlated at $r \approx 0.57$ and $r \approx 0.67$ respectively. Plots of CIE-*Lab* and CIE-*xyz* colour systems are not given here as these two colour systems are simple transformations of the more basic tristimulous (CIE-*XYZ*) colour system.

The observations from the three colour systems are continuous in their nature, and in this case the distribution between items for all univariate features from the CIE-*XYZ*, CIE-*xyz*, and CIE-*Lab* systems were found to be marginally normal (Shapiro test), with the exception of that of the *a* element of the CIE-*Lab* parametrisation, for which there was a moderate lack of support for marginal between items normality. Despite this it was decided that the lack of support was small enough for normality to be a good approximation to the density of the *a* element in the CIE-*Lab* system.

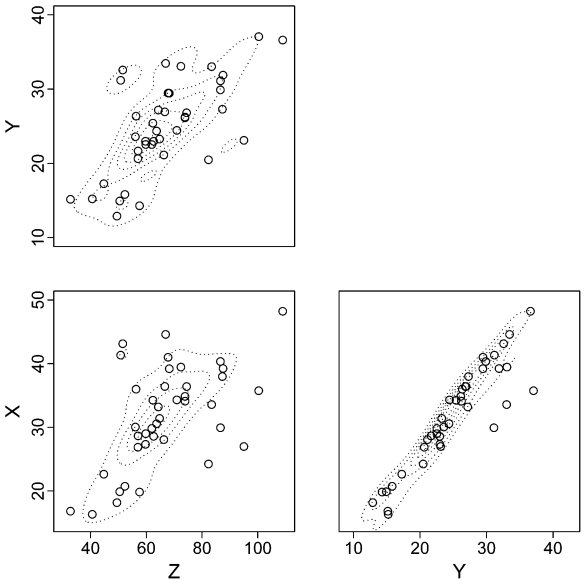

Fig. 1 Pairwise plots of the tristimulous values for the CIE-*XYZ* colour system. The points represent the mean values for the forty inks and the contours (dashed lines) are taken from the 10 replicated values for each of the forty inks.

All 820 possible pairwise comparisons (780 under H_d and 40 under H_p) were made for each possible univariate CIE element, and each of the nine bivariate cases conducted for the possible pairs of CIE elements within each of the colour systems was under consideration here. The pairs were: *ab*, *aL*, *bL* from the CIE *Lab* system, *xy*, *xz*, *yz* from the CIE-*xyz* and *XY*, *XZ*, *YZ* from the CIE-*XYZ* system. Initially the likelihood ratios were evaluated in terms of the number of false positives, and the number of false negatives, for each of the 18 sets of features. These are given in Table 1.

In Table 1 a false positive is considered to be a likelihood ratio greater than one calculated for the pairwise comparison where both sets of observations are known to be from different

Table 1 Illustration of percentage of false positives (false +ves), and percentage of false negatives (false −ves) for every possible pairwise comparison

CIE-*abL*			CIE-*xyz*			CIE-*XYZ*		
	False +ves	False −ves		False +ves	False −ves		False +ves	False −ves
a	16.3	7.5	*x*	12.9	2.5	*X*	15.5	12.5
b	19.2	2.5	*y*	20.8	5.0	*Y*	17.1	15.0
L	16.2	10.0	*z*	18.7	2.5	*Z*	19.4	12.5
ab	3.5	2.5	*xy*	4.6	2.5	*XY*	6.8	10.0
aL	6.7	7.5	*xz*	4.6	2.5	*XZ*	5.4	7.5
bL	6.5	10.0	*yz*	4.6	2.5	*YZ*	6.3	7.5

inks. This represents a misleading support for the prosecutor's proposition. Likewise, a false negative is considered to be an estimated likelihood ratio less than one calculated for the pairwise comparison where the two sets of observations are known to come from the same ink. The lower both these rates are the better a system can be said to perform.

As can be seen from Table 1 most of the colour systems considered here reveal relatively limited rates of false positive and false negative answers and on the whole, bivariate colour parameterisations have a much lower proportion of false positives than univariate parameterisations.

For most of the colour components considered separately the false positive and false negative rates are comparable for *L*, *a*, *b* and *X*, *Y*, *Z*, their rates range from *ca.* 16% to *ca.* 19%. For the *x*, *y*, *z* components considered separately the rates range from *ca.* 13% to *ca.* 21%.

It is not possible to assess which set of CIE features delivers the overall best results based upon false positives and false negatives alone as they are all fairly similar.

However, there is one bi-variate combination within the CIE-*Lab* set (*ab*) that shows the lowest false rates of *ca.* 3%, however the remaining two, *bL* and *aL*, behave substantially worse getting nearly 7% of incorrect responses. A similar phenomenon is observed for the CIE-*XYZ* set, for which *XZ* behaves poorly in *ca.* 5% of cases, whereas *YZ* and *XY* extend this value to *ca.* 7%. Worth noting is the observation that the CIE-*xyz* set delivers equally low rates of false positive answers, none exceeding 5% for each of the possible bivariate combinations. Therefore chromaticity coordinates (CIE-*xyz*) would appear to create a parametrisation of choice for the ink evidence evaluation based on spectral data, since it is the combination of colour coordinates providing the least misleading information overall.

This conclusion that the CIE-*xyz* system is somehow optimal is re-enforced by the low levels of false negatives obtained for the feature combinations from it. For the bivariate combinations a 3% level of false negative answers is never exceeded. All the remaining univariate and bivariate combinations deliver similar (*b*, *x*, *z*, and *ab*) or higher (*a*, *L*, *y*, *X*, *Y*, *Z*, *aL*, *bL*, *XZ*, *YZ*, and *XY*) rates of false negative answers ranging from 2.5% to 15% for single features, and to 10% for the bivariate ones. The worst set of features providing the highest rates of incorrect responses is the *XYZ* set. It is obvious that this set should not be used further for evidence evaluation.

Of some concern is the fact that some variables, such as these considered to be the best *xy*, *xz*, *yz* and *ab*, and all the univariate ones, introduce more false positives than false negatives. This is a rather undesirable effect which may lead to serious legal consequences in real cases. One should imagine a situation when a likelihood ratio suggests two inks come from the same source, where in reality they do not. Then the defendant may be convicted, despite being truly innocent. The converse situation is slightly less serious, when the rates of false negative answers are relatively high, since in the worst case it will result in a true offender not being convicted.

False positive and false negative rates give some relative idea of the overall performance, however, as discussed above, a more complete picture is given by the empirical cross-entropy. A contrasting pair of sets of ECEs is given in Fig. 2. Fig. 2(a) are ECEs from the *XY* variable pair from the CIE-*XYZ* colour system. This feature set would be a poor choice for providing evidence for the source of an ink as the loss of information is high, and there are high losses due to calibration. By way of contrast Fig. 2(b) would be a good choice of features to examine as there is little loss of information, and few calibration problems. These two plots are given across a range of the log to base ten of the prior odds on H_p. As ECE is a measure of loss of information then lower is usually better. In Fig. 2(a) the observed ECEs (solid red line) are very high, in fact so high that at points for log odds on $H_p > 0.8$ they go above the line representing the case where the features give no information about proposition, and the features could, for these regions, be seen as giving misleading information. It is also evident that the ECE values (solid line) are some way above the theoretical calibrated set line. From this one may infer that the features are, in an information sense, poorly calibrated. In all, the *XY* feature set from the CIE-*XYZ* colour system would be a poor choice of feature set from which to draw conclusions about the source of ink. In contrast Fig. 2(b) shows that the observed ECEs (solid line) are quite low. This

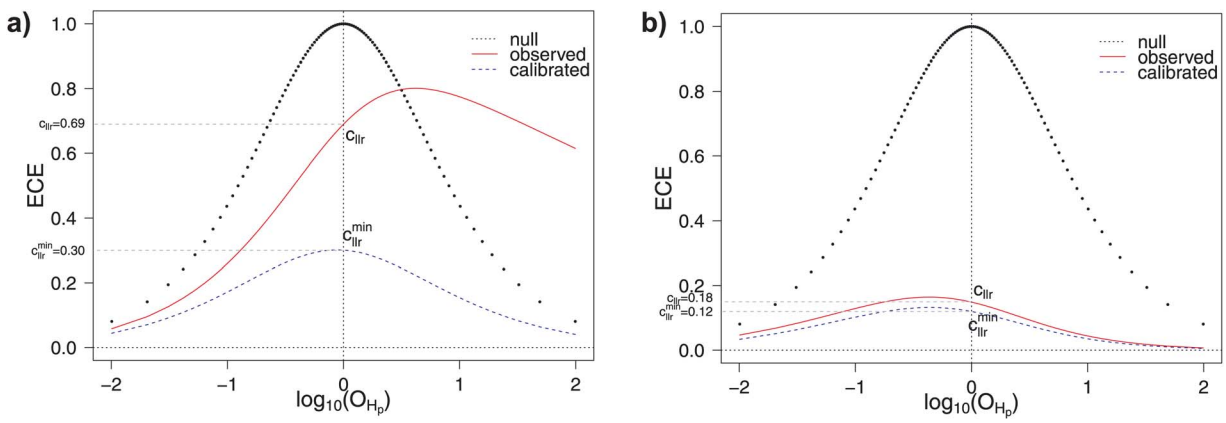

Fig. 2 Specimen empirical cross-entropy (ECE) plot for the ECEs from two contrasting colour components. Plot (a) is the bivariate *XY* component from the CIE-*XYZ* system, and (b) is for the ECEs from the bivariate *xy* component from the CIE-*xyz* system. $\log_{10}(O_{H_p})$ denotes the prior log odds in favour of H_p.

means that there is little loss of information about the ink source. Also, the observed ECEs (solid line) are very close to a theoretical optimum represented by the calibrated set (dashed line). This means there is little loss of information due to calibration. This feature set (*xy* from CIE-*xyz*) would be a good set of features to look at for the evidence of the ink source.

The full ECE function is usually evaluated for a range of prior values for the $\log(O_{H_p})$ of H_p centred around zero, however comparisons between variable sets can be made at $\log(O_{H_p}) = 0$,[37,44] simplifying considerably the presentation of many comparisons. Fig. 3 gives ECE values (C_{llr} and C_{llr}^{min}) for each of the nine univariate variables, and each of the nine bivariate pairs. These are arranged in descending order of C_{llr} and those variable sets towards the bottom of Fig. 3 are better, and are on the whole closer to performing as well as possible for some feature set with the same discriminating power.

The very much better performing variable sets are the bivariate pairs, the univariate colour elements performing worse. This follows the same pattern seen from the false negatives and positives from Table 1. Of the bivariate pairs the CIE-*xyz* system consistently gave minimal loss of information, and has the advantage that is a simple rescaling of the fundamental tristimulous values. Also a bivariate combination of *a* and *b* variables within the CIE-*Lab* colour system delivers satisfactory results and effectively reduces the information loss due to the evidence analysis.

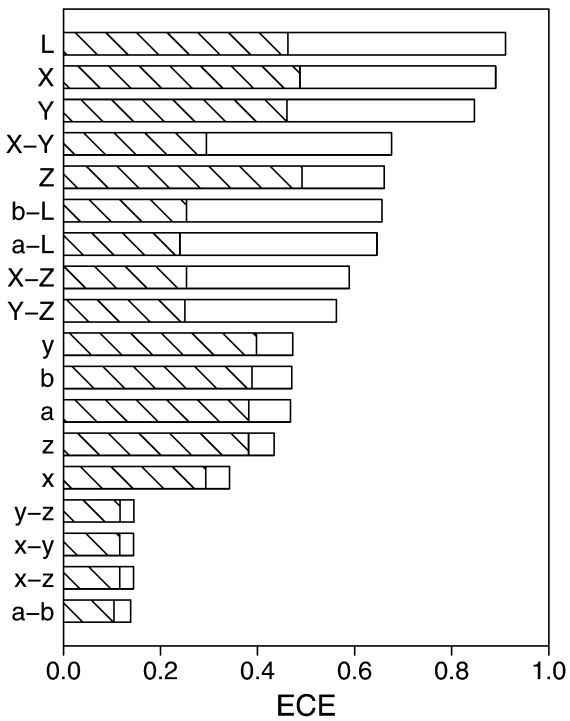

Fig. 3 Barplot of the ECE for each univariate, and pairwise combination, of colour parameterisations. Each bar is the ECE at $\log(O_{H_p}) = 0$, known as C_{llr}. The hatched regions for each bar represent the calibrated ECE, also for prior log odds zero, known as C_{llr}^{min}, for each univariate, and pairwise combination, of colour parameterisations. The bars are arranged in descending order of C_{llr} for which the values can be seen as loss of information, so lower is better. This plot summarises the ECEs for many feature sets in a more succinct way than that of Fig. 2.

4. Conclusions

LR models were employed for the evidence evaluation process of the data obtained for blue pen inks within CIE-*XYZ* (tristimulous values: *X*, *Y* and *Z*), CIE-*xyz* (chromaticity coordinates: *x*, *y* and *z*), or CIE-*Lab* colour determination systems (*a*, *b* and *L* parameters). Both univariate and bivariate LR computations were performed. The performance of the models was assessed by the empirical cross-entropy approach.

It has been found that the bivariate combinations of *xyz* set of colour features, as well as bivariate combination of *a* and *b* colour features within the CIE-*Lab* colour system achieve the best results, both in terms of false positives, false negatives, and empirical cross-entropy, whereas those derived from the *XYZ* set seem to be the most misleading. For variables *xy*, *xz* and *yz* the loss of information by evaluating the evidence is reduced from 100% to 15% and from 100% to 13% for *ab* variables.

References

1 C. Roux, M. Novotny, I. Evans and C. Lennard, *Forensic Sci. Int.*, 1999, **101**, 167–176.
2 C. Neumann, R. Ramotowski and T. Genessay, *J. Chromatogr., A*, 2011, **1218**, 2793–2811.
3 K. Banas, A. Banas, H. O. Moser, M. Bahou, W. Li, P. Yang, M. Cholewa and S. K. Lim, *Anal. Chem.*, 2010, **82**, 3038–3044.
4 J. A. Denman, W. M. Skinner, K. P. Kirkbirde and I. M. Kempson, *Appl. Surf. Sci.*, 2010, **256**, 2155–2163.
5 M. Gallidabino, C. Weyermann and R. Marquis, *Forensic Sci. Int.*, 2011, **204**, 169–177.
6 Z. Takats, J. M. Wiseman, B. Gologan and R. G. Cooks, *Science*, 2004, **306**, 471–473.
7 R. B. Cody, J. A. Laramee and H. D. Durst, *Anal. Chem.*, 2005, **77**, 2297–2302.
8 R. W. Jones, R. B. Cody and J. F. McClelland, *J. Forensic Sci.*, 2006, **51**, 915–918.
9 K. Bojko, C. Roux and B. J. Reedy, *J. Forensic Sci.*, 2008, **53**, 1458–1467.
10 A. Kher, M. Mulholland, E. Green and B. Reedy, *Vib. Spectrosc.*, 2006, **40**, 270–277.
11 J. Wang, G. Lou and S. Sun, *J. Forensic Sci.*, 2001, **46**, 1093–1097.
12 W. Dirwono, J. S. Park, M. R. Augustin-Camacho, J. Kim, H. M. Park, Y. Lee and K. B. Lee, *Forensic Sci. Int.*, 2010, **199**, 6–8.
13 R. M. Seifar, J. M. Verheul, F. Aries, U. A. T. Brinkman and C. Gooijer, *Analyst*, 2001, **126**, 1418–1422.
14 J. Zieba-Palus and M. Kunicki, *Forensic Sci. Int.*, 2006, **158**, 164–172.
15 S. K. Shevell, *The science of color*, Elsevier, Oxford, 2003.
16 V. Causin, R. Casamassima, C. Marega, P. Maida, S. Schiavone, A. Marigo and A. Villari, *J. Forensic Sci.*, 2008, **53**, 1468–1472.
17 D. Laing and M. Isaacs, *J. Forensic Sci.*, 1983, **23**, 147–154.
18 P. Pfefferli, *Forensic Sci. Int.*, 1983, **23**, 129–136.
19 SWGMAT, Standard Guide for Microspectrophotometry and Color Measurement in Forensic Paint Analysis, Scientific

working group for materials analysis (swgmat) technical report, 2007.
20 C. G. G. Aitken and F. Taroni, *Statistics and the evaluation of evidence for forensic scientists*, Wiley, 2004.
21 C. Adam, S. Sherratt and V. Zholobenko, *Forensic Sci. Int.*, 2008, **174**, 16–25.
22 C. Adam, *Forensic Sci. Int.*, 2008, **182**, 27–34.
23 N. Thanasoulias, N. Parisis and N. Evmiridis, *Forensic Sci. Int.*, 2003, **138**, 75–84.
24 Guest Editorial, *Sci. Justice*, 2011, **51**, 1–2.
25 C. Aitken, *Probl. Forensic Sci.*, 2006, **65**, 68–81.
26 J. Curran, C. Triggs, J. Almirall, J. Buckleton and K. Walsh, *Sci. Justice*, 1997, **37**, 241–244.
27 B. Robertson and G. Vignaux, *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*, Wiley, 1995.
28 D. Lindley, *Biometrika*, 1977, **64**, 207–213.
29 C. Aitken and D. Lucy, *Appl. Statist.*, 2004, **53**, 109–122.
30 G. Zadora, T. Neocleous and C. Aitken, *J. Forensic Sci.*, 2010, **55**, 371–384.
31 T. Neocleous, C. Aitken and G. Zadora, *Chemom. Intell. Lab. Syst.*, 2011, **109**, 77–85.
32 C. Aitken, D. Lucy, G. Zadora and J. Curran, *Computational Statistics and Data Analysis*, 2006, **50**, 2571–2588.
33 C. Aitken, G. Zadora and D. Lucy, *J. Forensic Sci.*, 2007, **52**, 412–419.
34 C. E. H. Berger, *Sci. Justice*, 2009, **49**, 265–271.
35 D. Risinger, M. Saks, W. Thompson and R. Rosenthal, *California Law Review*, 2002, **90**, 1–56.
36 G. Zadora and D. Ramos, *Chemom. Intell. Lab. Syst.*, 2010, **102**, 63–83.
37 D. Ramos and G. Zadora, *Anal. Chim. Acta*, 2011, **705**, 207–217.
38 D. Ramos, J. Gonzalez-Rodriguez, G. Zadora and C. Aitken, *J. Forensic Sci.*, 2012, DOI: 10.1111/1556-4029.12233.
39 A. Zeichner, N. Levin, A. Klein and Y. Novoselsky, *J. Forensic Sci.*, 1988, **33**, 1171–1181.
40 L. Olson, *J. Forensic Sci.*, 1986, **31**, 1330–1340.
41 G. Zadora, *J. Chemom.*, 2010, **24**, 346–366.
42 http://cran.r-project.org, last visited 03.09.2013.
43 B. Efron, *The jackknife, the bootstrap, and other resampling plans*, Society for Industrial and Applied Mathematics, Philadelphia, 1982.
44 D. Ramos and J. Gonzalez-Rodriguez, *Forensic Sci. Int.*, 2013, **230**, 156–159.
45 D. Lucy and G. Zadora, *Forensic Sci. Int.*, 2011, **212**, 189–197.
46 N. Brümmer and J. du Preez, *Comput. Speech Lang.*, 2006, **20**, 230–275.
47 M. Ayer, H. Brunk, G. Ewing, W. Reid and E. Silverman, *Ann. Math. Stat.*, 1955, **26**, 641–647.
48 M. Best and N. Chakravarti, *Math. Program.*, 1990, **47**, 425–439.