Inference of identity of source using univariate and bivariate methods

Charles E.H. Berger

Netherlands Forensic Institute, P.O. Box 24044, 2490 AA The Hague, The Netherlands.

In this study we explore the inference of identity of source using a two-dimensional feature vector. As an example, we study the use of the Bayesian framework for the estimation of the value of evidence of color measurements for identity of source of blue ballpoint pen inks. Univariate as well as bivariate analyses are carried out for color data that was acquired with a flatbed scanner. While this might not be the best method to discriminate inks, we will use it as an example to estimate what the value of the evidence is, however low or high it may be. It is hoped that this exercise is instructional, as a similar approach can readily be applied in other situations.

1 Introduction

The study in this paper is about an example of an inference of identity of source [1] using a Bayesian approach [2]. Generally, discrimination or comparison methods give a score which is a measure for either similarity or difference. It seems most logical to have the score increase with increasing difference, because there is a finite limit on one side (score zero for no difference), while the difference has no limit on the other side (features can always be more different). But even if the opposite type of score has been chosen, the same univariate approach for the inference of identity of source remains applicable.

For the present work, we will consider an example with a two-dimensional feature vector that characterizes the color of an ink. Color difference is defined as the (Euclidian) distance between two colors in this two-dimensional feature space.

We will carry out univariate analyses for the estimation of the evidential value for identity of source based on the comparison scores (color differences) of multiple ink traces from a single pen and single traces of multiple pens. We will also calculate the evidential value using bivariate methods, where we work with probability density functions (PDFs) for measuring the observed colors under either competing hypothesis regarding the source, thus bypassing the reduction to one dimension through a comparison score.

The inference of identity of source is the main topic of this study; non-destructive methods for the comparison of inks have been described elsewhere [3-6].

2 Methods

2.1 Reference ballpoint collection

For this study, 262 blue ballpoint pens from the collection of the Netherlands Forensic Institute (NFI) were used. Since the issue is whether or not the same specific instance of a pen was used, the collection should truly reflect the population of all pens, with each type of pen represented in proportion to its frequency amongst pen users. For the present study we will assume the collection to be representative of the relevant population. Though this is not sure for the present collection, it is not a limitation of the principle of our analysis.

2.2 Preparation and imaging of the samples

Samples were prepared by writing lines with all ballpoint pens on a single sheet of paper. To get an impression of the intra-source (or within source) variation for a single ballpoint pen, 100 samples were written with the same ballpoint pen on the same sheet. The imaging was done by scanning all samples in one large, high resolution scan (1270 dpi, or pixels of 20×20 μm), with a high quality scanner (CreoScitex Eversmart Jazz). After acquiring the image, it was sliced into a collection of images of all the separate samples.

2.3 Defining the feature vector

The analysis of the colors is analogous to that in Ref. [7], and is based on the three-dimensional color histogram (see Figure 1). This histogram shows the distribution of all colors present in an image in the RGB (red, green, and blue component) color space.

Note how the colors of one and the same source vary wildly, extending from the spherical cloud of colors associated with the paper background (*P*) to the color of the pure ink. This is due to differences in ink coverage in the pixels in and along the edge of the ink line. For the purpose of discrimination however, the chosen feature vector should vary as little as possible for the same source, and as much as possible for different sources.

The spatial angle of the vectors from the background paper color (*P*) to the varying ink colors, varies much less than those colors themselves. Because of this minimized variation and the fact that different colored inks will give different spatial angles, these spatial angles were chosen as the feature vector. Defining the feature vector as the direction of the elongated cloud of colors associated with an ink makes the analysis much less sensitive to ink coverage and more sensitive to different inks.

To determine this direction we first determine the average RGB values for the paper in each image. This is done by finding the position of the peak associated with the paper color in the red component histogram of the image. All pixels with a red component within 2% of the peak position are averaged to give the average RGB value of the paper color. By averaging all pixels with a red component between 50 and 130 (on a scale of 0 to 255), we also obtain an average RGB value associated with the ink.

Our feature vector is defined by the spatial angles (*x*, *y*) of the vector from the average paper RGB value to that associated with the ink, relative to the RGB axes.

3 Results

From here on we will often refer to traces from a source as *colors*, defined by their feature vector rather than by their RGB values. Figure 2 shows the intra-source (or within source) variation of the feature vector (ballpoint pen ink color) for one of the pens in a cluster of solid dots, while the open dots represent the inter-source (or

between source) variation of the feature vector. It is assumed that the intra-source variation is similar for all sources.

We define the relevant hypotheses for an ink comparison as follows:

H_s : Colors $\vec{a}$ and $\vec{b}$ represent samples that come from the *same* specific instance of a blue ballpoint.

H_d : Color $\vec{b}$ represents a sample that comes from a random blue ballpoint, *different* from the ballpoint that led to $\vec{a}$.

3.1 Univariate approaches

We define color differences as Euclidian distances d (a comparison score) between the colors of two samples in the two-dimensional feature space. We can make histograms of the distances d between all possible pairs of colors for the intra-source and the inter-source measurements. Strictly speaking we have only 50 independent pairs for the 100 intra-source samples (and an analogous limitation exists for the inter-source samples), but we have chosen to use this approximation. From these histograms follow the probability density functions for measuring certain color differences, for intra-source as well as inter-source samples (see Figure 3a). By dividing the two PDFs we obtain the *LR* (likelihood ratio) associated with our hypotheses, as a function of the color difference between $\vec{a}$ and $\vec{b}$. For reasons of symmetry and presentation we will work with the *LLR* (log likelihood ratio), which for evidence pointing in opposite directions has opposite signs and for neutral evidence equals zero (see Figure 3b, solid dots).

If we assume our intra-source feature vectors to be normally distributed (with a standard deviation σ), the PDF for color differences given H_s (same source) becomes a Rayleigh distribution of the form

$$f(d|\sigma) = \frac{d}{\sigma^2} e^{-\frac{d^2}{2\sigma^2}}, \qquad (1)$$

that we can fit to the data. Using this fit we can extrapolate to some extent to find the *LLR* for color differences for which no intra-source data were measured (avoiding division by zero). The result can be seen in Figure 3b (open dots).

Kernel density estimation (KDE) can be used to obtain a continuous, smooth curve (see solid curves in Figure 3) for the inter-source data. With KDE, the PDF is

constructed as a sum of Gaussians around every inter-source color difference μ_i, leading to a smoother PDF:

$$f(d|\mu,\kappa) = \sum_{i=1}^{n} \frac{1}{\kappa\sqrt{2\pi}} e^{-\frac{(d-\mu_i)^2}{2\kappa^2}}, \qquad (2)$$

where κ is a smoothing constant which was chosen as 0.003, which is close to the standard deviation σ of the intra-source feature vectors in either dimension.

In reality of course, the *LLR* should not only depend on the color difference *d*, but also on the actually involved colors $\vec{a}$ and $\vec{b}$, because the value of the evidence should depend on whether those colors are rare or common. To incorporate this information we derive a different PDF for the color differences given H_d (different source). This concept is usually referred to as anchoring, while the previous approach is called non-anchored. The PDF is now based on the distances from $\vec{a}$ to all inter-source feature vectors (anchoring to $\vec{a}$).

To obtain meaningful PDFs from this smaller set of color differences, we apply kernel density estimation again. Figure 4a shows the probability density functions for the inter-source measurements (using KDE) based on three different histograms: that of all possible distances (solid line), that of all distances to a common $\vec{a}$ (dashed line), and that of all distances to a rare $\vec{a}$ (dotted line). The associated log likelihood ratios (*LLR*) are shown in Figure 4b.

We will now calculate a two-dimensional *LLR* distribution as a function of measurement $\vec{b}$ given a certain measurement $\vec{a}$. That way we can look at the univariate results in two dimensions, which will also allow us to compare with the results of the bivariate approach later.

Figure 5 shows the value of the *LLR* as a function of measurements $\vec{b}$ with given $\vec{a}$, in 6 different situations. From top to bottom 3 univariate approaches are used. The results on the left side are for a common $\vec{a}$ while those on the right are for a rare $\vec{a}$. The graphs show lines of equal *LLR*, and cross sections along the vertical axes.

Figure 5a and 5b show the non-anchored univariate results using all inter-source distances (compare Figure 3b). In Figure 5c and 5d, the univariate *LLR* results using all distances to $\vec{a}$ ($\vec{a}$-anchored) are shown (compare Figure 4b).

In a third univariate approach (not mentioned before), the distances between $\vec{b}$ and all inter-source measurements are used to create the probability density function. The results of this $\vec{b}$ -anchored approach are shown in Figure 5e and 5f.

3.2 Bivariate approaches

We will also calculate a two-dimensional distribution of *LLR* values as a function of $\vec{b}$ given a certain a in two bivariate approaches. The evidence consists of the feature vectors of the first and second color measurement a and $\vec{b}$: $E = (a, \vec{b})$. For continuous measurements the probabilities are replaced by probability density functions *f* so that

$$LR = \frac{f\left(\vec{a}, \vec{b} \middle| H_{same}, I\right)}{f\left(\vec{a}, \vec{b} \middle| H_{different}, I\right)}. \tag{3}$$

Applying the rules of conditional probability we can write (for a derivation see the Appendix):

$$LR = \frac{\int f\left(\vec{a} \middle| \vec{\theta}\right) f\left(\vec{\theta}\right) d\vec{\theta} \int f\left(\vec{b} \middle| \vec{\theta}\right) f\left(\vec{\theta}\right) d\vec{\theta}}{\int f\left(\vec{b} \middle| \vec{\theta}\right) f\left(\vec{a} \middle| \vec{\theta}\right) f\left(\vec{\theta}\right) d\vec{\theta}}, \tag{4}$$

where we denote the true mean of the measurement on a source by $\vec{\theta}$, and we omit the symbols for the hypotheses and background information.

This equation is evaluated numerically to find the *LR* for a given a as a function of $\vec{b}$. The probability density functions are represented by arrays that can easily be integrated numerically and displayed graphically.

We first calculate $f\left(\vec{\theta}\right)$ by constructing the probability density function from three-dimensional Gaussian peaks associated with the feature vectors $\vec{p}_i$ for every ink in our collection, using a smoothing constant κ = 0.01 for kernel density estimation:

$$f\left(\vec{\theta}\right) = \frac{1}{2n\pi\kappa^2} \sum_{i=1}^{n} e^{-\frac{\left|\vec{\theta} - \vec{p}_i\right|}{2\kappa^2}}. \tag{5}$$

This distribution is graphically represented in Figure 6a. It corresponds to our expectation to measure a certain color if the source is randomly chosen. It is an approximation in the sense that for this study every $\vec{p}_i$ is a single measurement, and not a true mean.

Next we evaluate $f\left(\vec{a}\middle|\vec{\theta}\right)$. The Gaussian that determines it has the standard deviation σ that was derived from the fit to the Rayleigh distribution earlier (Equation 1). The resulting distribution is evaluated numerically and an example can be found in Figure 6b.

We can now calculate $f\left(\vec{b}\middle|\vec{\theta}\right)$ for every $\vec{b}$ and with that and $f\left(\vec{a}\middle|\vec{\theta}\right)$ and $f\left(\vec{\theta}\right)$ we can integrate according to Eq. 4 to calculate the *LLR* as a function of $\vec{b}$.

If the inter-source variability is approximated by a normal distribution instead of a kernel density estimate, the calculation can be simplified, and numerical integration of Equation 4 is not needed [8].

Figure 7 shows the results for both bivariate approaches. The inter-source variability is modeled by a bivariate normal distribution (see Figure 7a and 7b), or by a kernel density estimate (see Figure 7c and 7d).

4 Discussion

The first univariate approach did not take into account that the *LLR* should not only depend on the color difference, but also on the actual values of $\vec{a}$ and $\vec{b}$ themselves. Therefore, the results in Figure 5a and 5b are essentially the same but centered around a different $\vec{a}$.

In the second univariate approach the inter-source variation depends on $\vec{a}$ but not on $\vec{b}$, which makes Figure 5c different from 5d, while both still consist of concentric circles (since $\vec{a}$ is not varied).

For the more rare colors, the same color difference leads to a higher *LLR* for the colors coming from the same source, because there are fewer alternative sources that resemble the true source. For the more common colors those alternative candidates *are* available, which leads to an increase of the support for the hypothesis that the colors originated from different ballpoint pens (lowering the *LLR*). The cross sections given with Figure 5a to 5d are the same as the graphs shown earlier in Figure 3b and 4b.

The third univariate approach is like the second one, but due to $\vec{b}$-anchoring, the inter-source variation now depends on $\vec{b}$ and not on $\vec{a}$. The effect of the rarity of the colors is seen again in Figure 5e and 5f, but since the *LLR* is given as a function of

$\vec{b}$, we do not have concentric circles anymore. Please note that for this type of casework it does not matter which color you measure first, a or $\vec{b}$. This symmetry is reflected in the results for common colors (Figure 5c and 5e), which are very similar. For rare colors however (with less data), the results can differ significantly, as seen in Figure 5d and 5f. The *LLR* in those graphs also seems overly sensitive with respect to the color difference *d*.

Two bivariate approaches were used to obtain Figure 7, with the inter-source variability modeled by a normal distribution (Figure 7a and 7b) and by a kernel density estimate (Figure 7c and 7d), respectively. The effect of the rarity of ink color a is visible again, but without the hypersensitivity seen in Figure 5d and 5f.

To compare the performance of the various methods we can also look at the *LLR* values for intra-source comparisons (true pairs) and inter-source comparisons (false pairs). Ideally, the histogram for the true pair *LLR* values would be well-separated from that for the false pair values. For the present study, we will plot instead a cumulative derivative of the histogram called the Tippett plot [9]. The Tippett plot gives the proportion of the *LLR* values greater than a value *s*, for cases corresponding to either hypothesis.

The bivariate approaches perform better than the univariate ones, though differences do not seem to be very large. For the false pairs the curves for the univariate approaches overlap, and so do the curves for the bivariate approaches. Differences can be seen in the inset of Figure 8. The bivariate method with normally distributed between-source variability performs better than the bivariate KDE method for the true pairs, but the latter performs better for the false pairs and overall as well.

5 Conclusion

In many forensic fields, comparisons lead to one dimensional scores and univariate methods for calculating *LLR* values. While this simplifies calculations, it will not lead to the most accurate determination of the *LLR*, in particular for the less common feature vectors. The example given here of univariate and bivariate Bayesian analysis for the inference of identity of source for color measurements is meant to illustrate the different approaches. The bivariate methods were shown to have a better performance than the univariate methods.

In a follow-up of this study we will explore more methods of judging the performance of *LLR* calculation methods and calibrating [10] their outcome.

Appendix

This derivation of Equation (4) follows pages 319 to 321 in Ref. [11].

The value of the evidence is defined as the likelihood ratio *LR*, with the two competing propositions denoted by H_s (same source) and H_d (different source) and the background information by *I*.

$$LR = \frac{\Pr(E|H_s, I)}{\Pr(E|H_d, I)}. \tag{1}$$

The evidence consists of the feature vectors of the first and second color measurement $\vec{a}$ and $\vec{b}$:

$$E = (\vec{a}, \vec{b}). \tag{2}$$

For continuous measurements the probabilities are replaced by probability density functions *f* so that

$$LR = \frac{f(\vec{a}, \vec{b}|H_s, I)}{f(\vec{a}, \vec{b}|H_d, I)}. \tag{3}$$

Applying the rules of conditional probability we can write:

$$LR = \frac{f(\vec{a}, \vec{b}|H_s, I)}{f(\vec{a}, \vec{b}|H_d, I)} = \frac{f(\vec{b}|\vec{a}, H_s, I)}{f(\vec{b}|\vec{a}, H_d, I)} \times \frac{f(\vec{a}|H_s, I)}{f(\vec{a}|H_d, I)} = \frac{f(\vec{b}|\vec{a}, H_s, I)}{f(\vec{b}|\vec{a}, H_d, I)}, \tag{4}$$

because the probability density function for $\vec{a}$ is independent of whether H_s or H_d is true:

$$f(\vec{a}|H_s, I) = f(\vec{a}|H_d, I). \tag{5}$$

If H_d is true, then the first and second measurement $\vec{a}$ and $\vec{b}$ are independent:

$$f(\vec{b}|\vec{a}, H_d, I) = f(\vec{b}|H_d, I), \tag{6}$$

and so

$$LR = \frac{f(\vec{b}|\vec{a}, H_s, I)}{f(\vec{b}|H_d, I)}. \tag{7}$$

In the following, we denote the true mean of the measurement on a source by $\vec{\theta}$, and we omit the symbols for the hypotheses and background information.

The numerator can be rewritten using the rules of total probability as before and Bayes' theorem:

$$f(\vec{b}|\vec{a}) = \int f(\vec{b}|\vec{\theta}) f(\vec{\theta}|\vec{a}) d\vec{\theta} \tag{8}$$

$$= \frac{\int f(\vec{b}|\vec{\theta}) f(\vec{a}|\vec{\theta}) f(\vec{\theta}) d\vec{\theta}}{f(\vec{a})} \tag{9}$$

$$= \frac{\int f(\vec{b}|\vec{\theta}) f(\vec{a}|\vec{\theta}) f(\vec{\theta}) d\vec{\theta}}{\int f(\vec{a}|\theta) f(\vec{\theta}) d\vec{\theta}}. \tag{10}$$

Applying the law of total probability and replacing summation with integration, we can write the denominator as:

$$f(\vec{b}) = \int f(\vec{b}|\vec{\theta}) f(\vec{\theta}) d\vec{\theta}. \tag{11}$$

We finally obtain

$$LR = \frac{\int f(\vec{b}|\vec{\theta}) f(\vec{a}|\vec{\theta}) f(\vec{\theta}) d\vec{\theta}}{\int f(\vec{a}|\vec{\theta}) f(\vec{\theta}) d\vec{\theta} \int f(\vec{b}|\vec{\theta}) f(\vec{\theta}) d\vec{\theta}}. \tag{12}$$

References

[1.] Q.Y. Kwan, Inference of Identity of Source, PhD Thesis, University of California, Berkeley, 1977.

[2.] I.W. Evett, Towards a Uniform Framework for Reporting Opinions in Forensic Science Casework. Science & Justice 38 (1998) 198-202.

[3.] J. Zięba-Palus and M. Kunicki, Application of the micro-FTIR spectroscopy, Raman spectroscopy and XRF method examination of inks. Forensic Science International 158 (2006) 164-172.

[4.] C. Roux, M. Novotny, I. Evans and C. Lennard, A study to investigate the evidential value of blue and black ballpoint pen inks in Australia. Forensic Science International 101 (1999) 167-176.

[5.] N.C. Thanasoulias, N.A. Parisis and N.P. Evmiridis, Multivariate chemometrics for the forensic discrimination of blue ball-point pen inks based on their Vis spectra. Forensic Science International 138 (2003) 75-84.

[6.] H.S. Chen, H.H. Meng and K.C. Cheng, A survey of methods used for the identification and characterization of inks. Forensic Science Journal 1 (2002) 1-14.

[7.] C.E.H. Berger, J.A. de Koeijer, W. Glas and H.T. Madhuizen, Color Separation in Forensic Image Processing. Journal of Forensic Sciences 51 (2006) 100–102.

[8.] C.G.G. Aitken and D. Lucy, Evaluation of trace evidence in the form of multivariate data. Applied Statistics 53 (2004) 109-122.

[9.] I.W. Evett and J.S. Buckleton, Statistical Analysis of STR Data: Advances in Forensic Haemogenetics, Springer-Verlag, Heidelberg, 1996.

[10.] D. Ramos, Forensic evaluation of the evidence using automatic speaker recognition systems, Ph.D. thesis, Universidad Autonoma de Madrid, 2007.

[11.] C.G.G. Aitken and F. Taroni, Statistics and the Evaluation of Evidence for Forensic Scientists, John Wiley & Sons, Chichester, 2004.

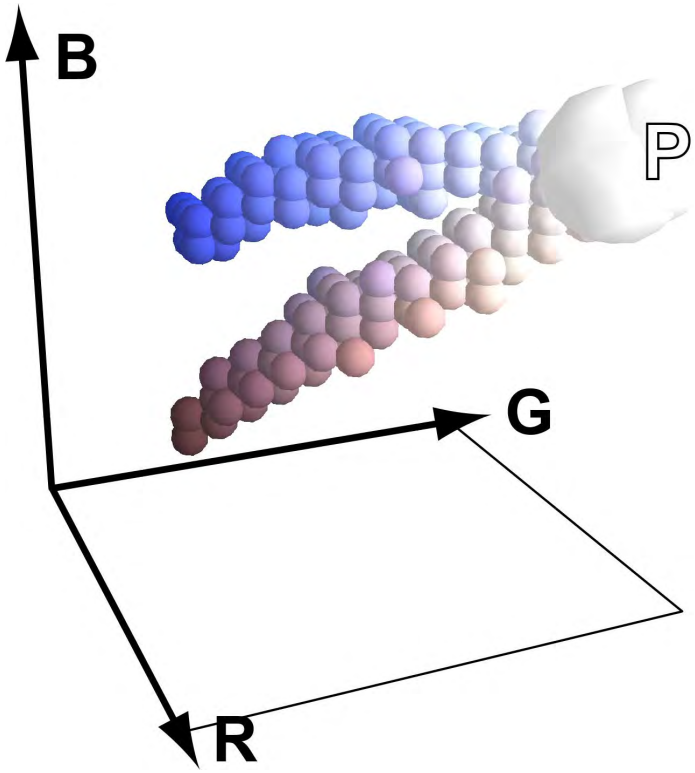

Figure 1

Three-dimensional color histogram, showing the distribution of colors present in an image in the RGB (red, green and blue component) color space. The frequency of pixels in every histogram bin is indicated by the radius of the corresponding sphere. In this case, the colors come from an image of black and blue ballpoint pen ink on white paper.

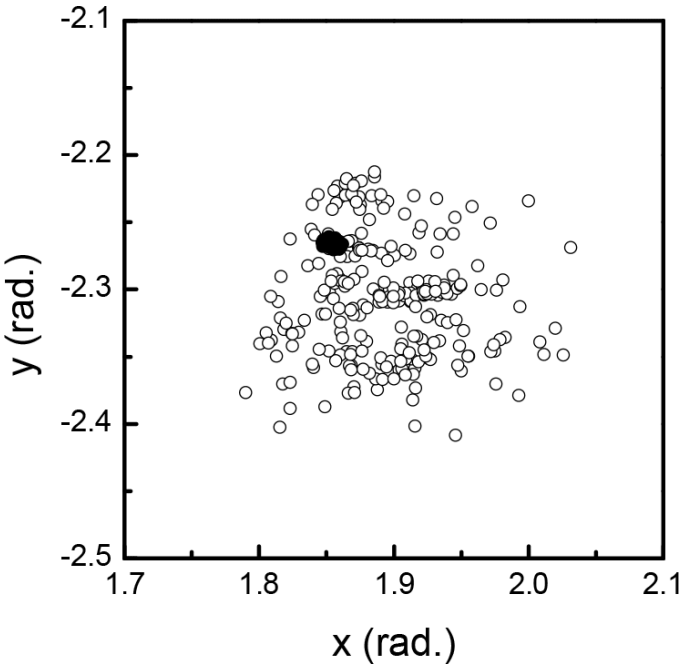

Figure 2

The inter-source (or between source) variation of the feature vector (ballpoint pen ink color), shown with open dots. The cluster of solid dots represents the intra-source (or within source) variation of the feature vector for one of the pens.

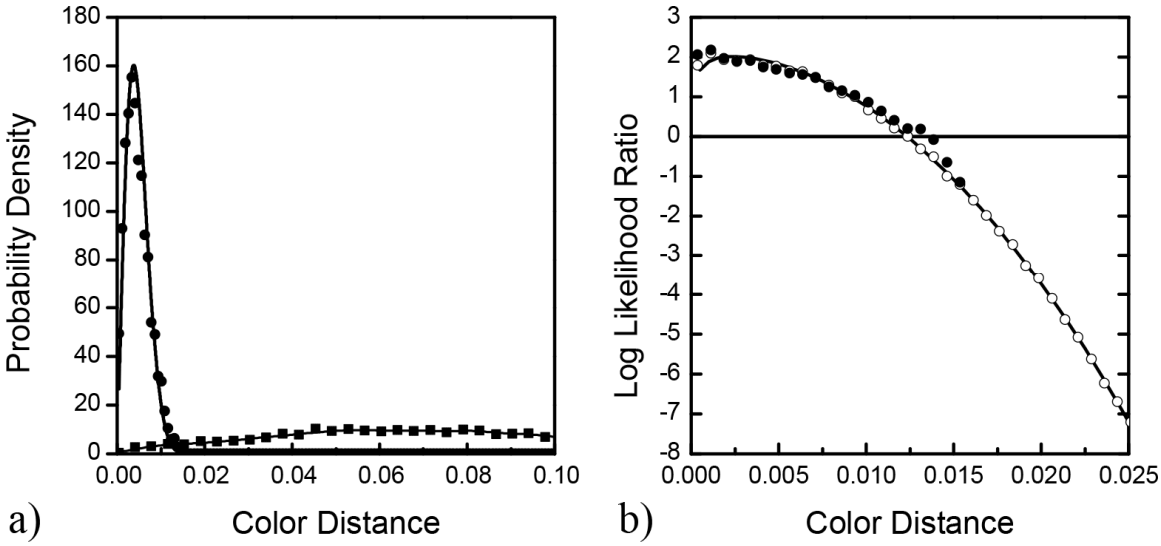

Figure 3a

Probability density functions for color differences, which were derived from histograms of the distances between all possible pairs of colors for the intra-source (solid dots) and the inter-source measurements (solid squares). The solid lines represent the Rayleigh fit for the intra-source data, and the kernel density estimation result for the inter-source data.

Figure 3b

The *LLR* (log likelihood ratio) giving the increase of the support for H_s relative to H_d as a function of the color difference, as derived from dividing the probability density functions in Figure 3a (solid dots). Using the Rayleigh fit for the intra-source distribution of distances, we can extrapolate (open dots). Applying kernel density estimation for the inter-source distribution we get a continuous smooth result (line).

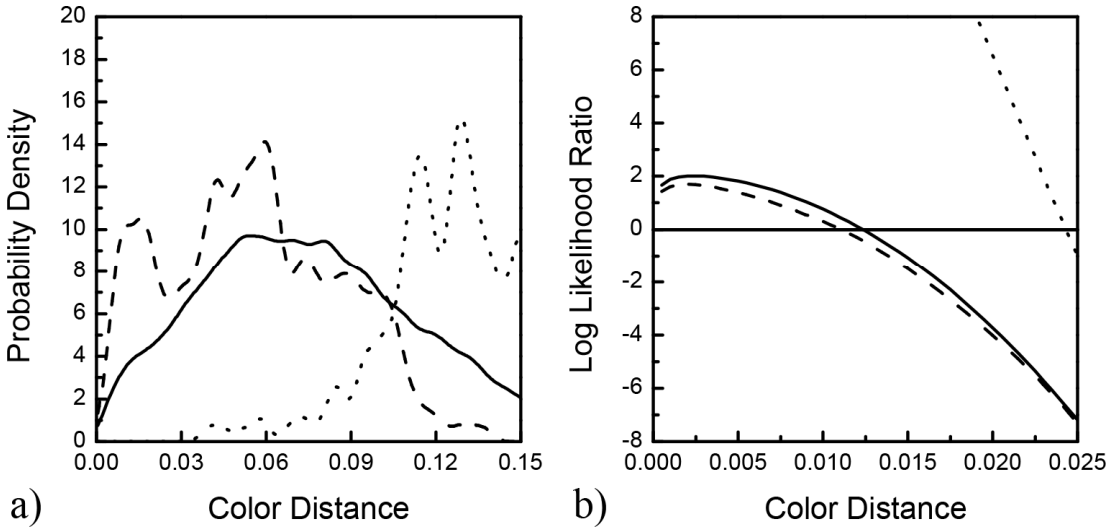

Figure 4a

The probability density functions for the inter-source measurements (using KDE) based on 3 different histograms: that of all possible distances (solid line), that of all distances to a for a common ink color (dashed line), and that of all distances to a for a rare ink color (dotted line).

Figure 4b

The *LLR* (log likelihood ratio) based on the Rayleigh fit and the three probability density functions on the left. The line types correspond to those in Figure 4a.

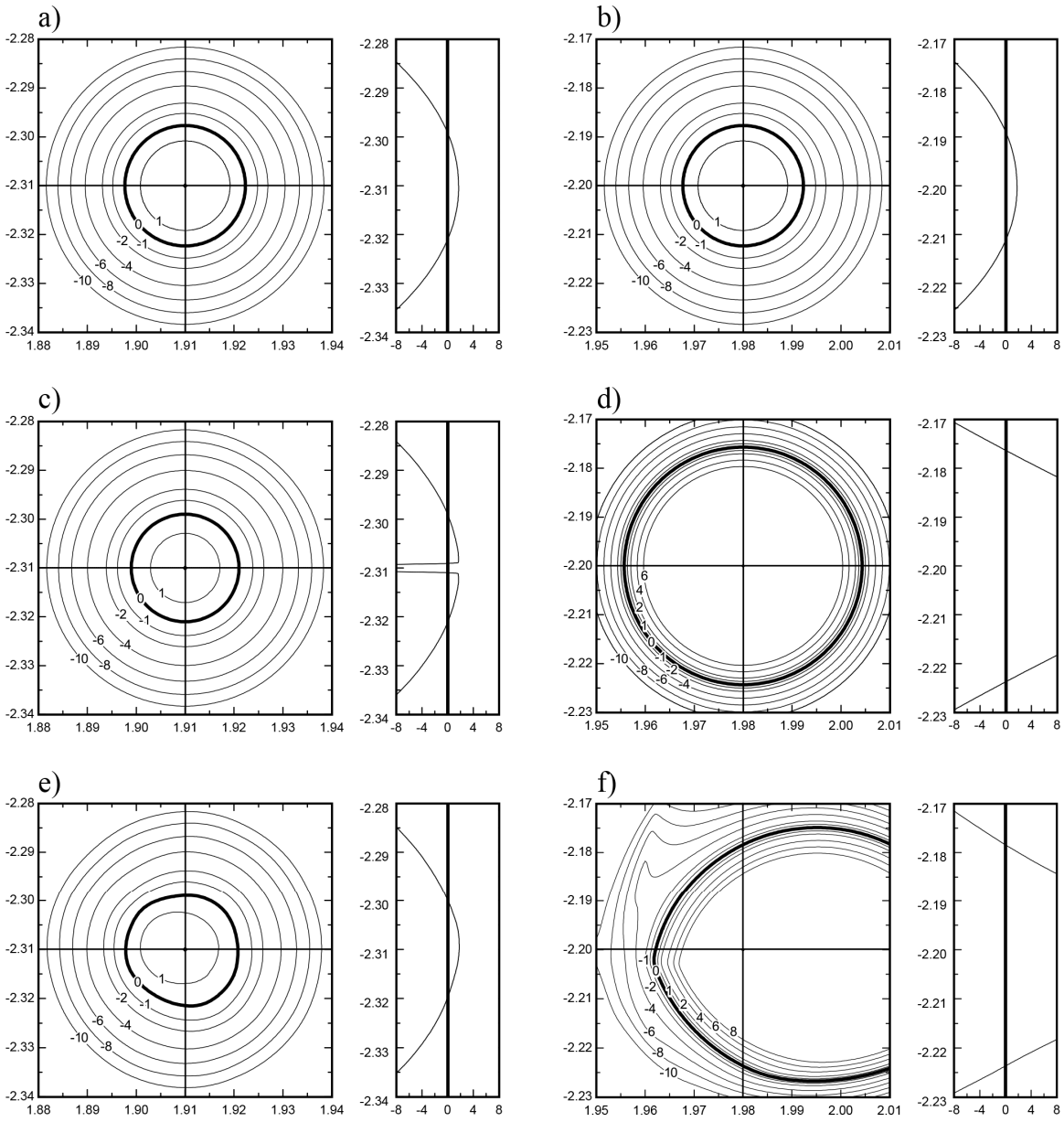

Figure 5

The *LLR* as a function of $\vec{b}$ and given $\vec{a}$, in 6 different situations. From top to bottom, three univariate approaches are used. The results on the left side are for a common a while those on the right are for a rare a. The graphs show lines of equal *LLR*, and a cross section along the vertical axis. The univariate approaches from top to bottom can be characterized as non-anchored, and anchored to a and $\vec{b}$ respectively.

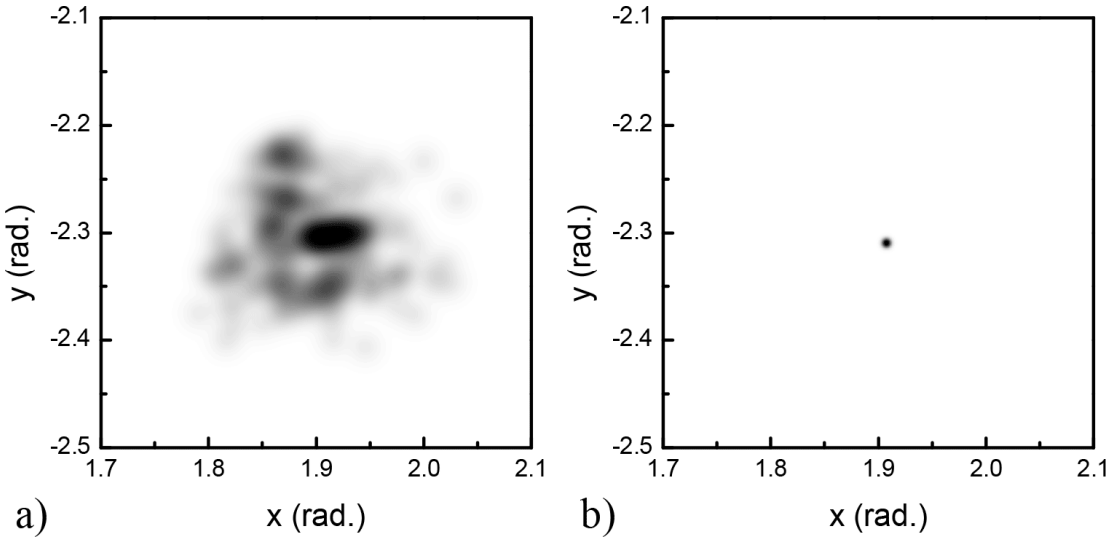

Figure 6a

The probability density $f(\vec{\theta})$ based on kernel density estimation, where the axes are the same as in Figure 2, and the intensity represents the probability density.

Figure 6b

Example of the probability density $f(\vec{a}|\vec{\theta})$ for $\vec{\theta}$ = [1.91, -2.31]. The axes are the same as in Figure 2, and the intensity represents the probability density.

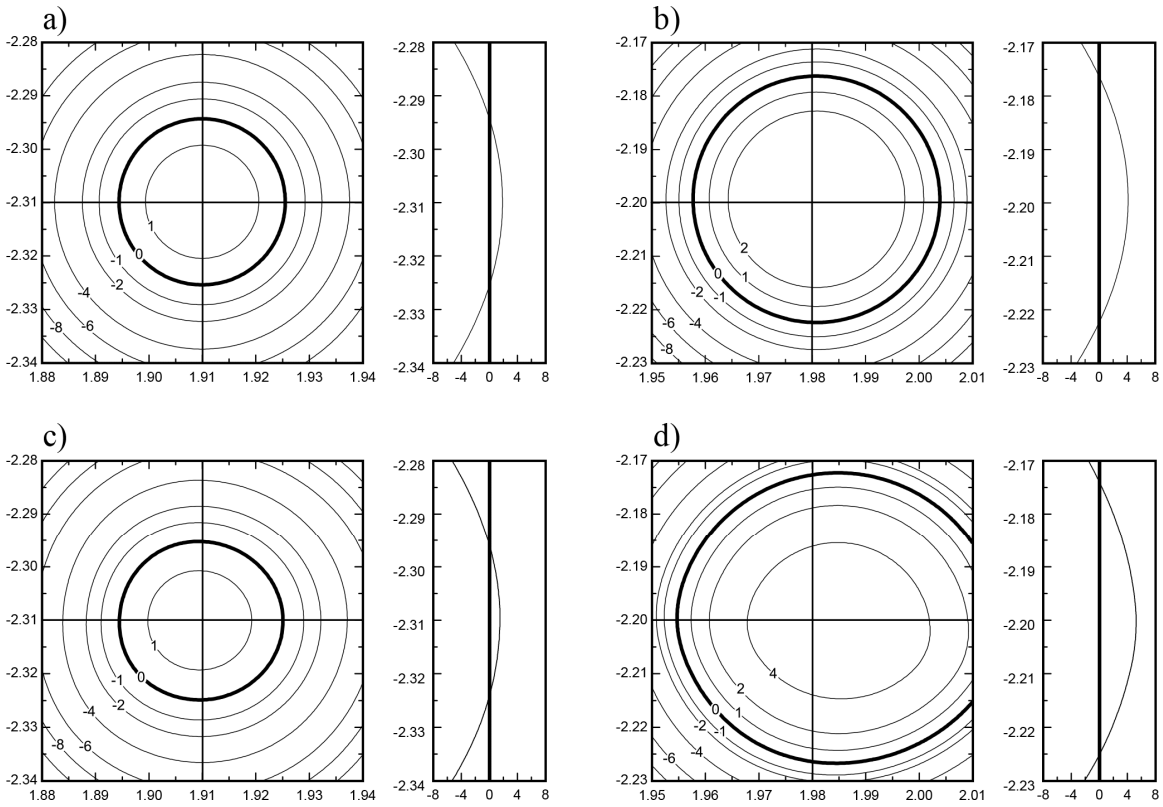

Figure 7

The *LLR* as a function of $\vec{b}$ and given $\vec{a}$, in 4 different situations. In 7a and 7b, the inter-source variability is modeled by a normal distribution, while for the graphs 7c and 7d a kernel density estimate was used. The results on the left side are for a common $\vec{a}$ while those on the right are for a rare $\vec{a}$. The graphs show lines of equal *LLR*, and a cross section along the vertical axis.

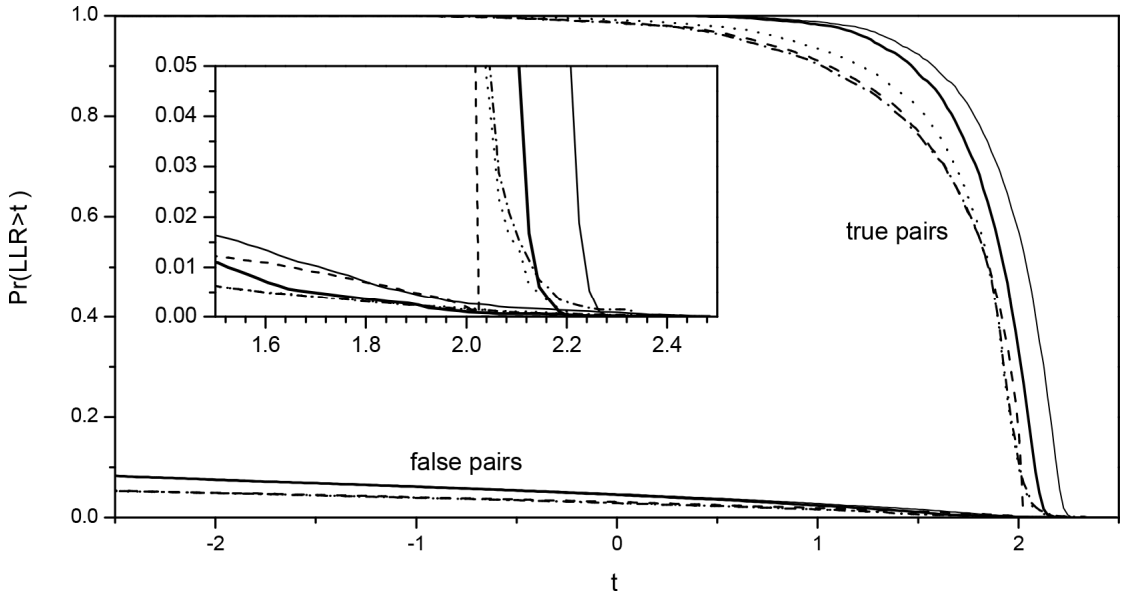

Figure 8

Tippett plots for the univariate methods: non-anchored (dashed line), a-anchored (dotted line), $\bar{b}$-anchored (dash-dotted line); and bivariate methods: with inter-source variability normally distributed (solid thin line), or approximated with KDE (solid thick line).