Forensic Science International 204 (2011) 58–66

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Implementing statistical learning methods through Bayesian networks (Part 2): Bayesian evaluations for results of black toner analyses in forensic document examination

A. Biedermann [a,*], F. Taroni [a], S. Bozza [b], W.D. Mazzella [a]

[a] The University of Lausanne, Ecole des Sciences Criminelles, Institut de Police Scientifique, le Batochime, 1015 Lausanne-Dorigny, Switzerland
[b] University Ca' Foscari, Department of Statistics, Venice, Italy

ARTICLE INFO

Article history:
Received 16 July 2009
Received in revised form 21 April 2010
Accepted 4 May 2010
Available online 2 June 2010

Keywords:
Bayesian networks
Statistical learning methods
Bayesian parameter estimation
Forensic document examination
Black toner

ABSTRACT

This paper presents and discusses the use of Bayesian procedures – introduced through the use of Bayesian networks in Part I of this series of papers – for 'learning' probabilities from data. The discussion will relate to a set of real data on characteristics of black toners commonly used in printing and copying devices. Particular attention is drawn to the incorporation of the proposed procedures as an integral part in probabilistic inference schemes (notably in the form of Bayesian networks) that are intended to address uncertainties related to particular propositions of interest (e.g., whether or not a sample originates from a particular source). The conceptual tenets of the proposed methodologies are presented along with aspects of their practical implementation using currently available Bayesian network software.

1. Introduction

Network fragments developed in Part I of this series of papers are reused here for constructing models whose purpose is to assist forensic scientists in the assessment of the evidential value of analytical results that pertain to the area of forensic document examination. The analyses will concentrate on selected characteristics of black toners that are typically encountered on printed or photocopied documents.

The purpose of the proposed Bayesian network models is to help scientists construct probabilistic arguments in order to approach situations of the following kind:

- Suppose selected characteristics of a black toner sample from a questioned document correspond to (or, are found to be indistinguishable from) those of reference documents produced by a particular xerographic printing machine (seized, for instance, in the context of a criminal investigation). What can be inferred about the proposition according to which the questioned document was printed by that machine (some other machine)?
- When black toner samples from two questioned documents are found to have indistinguishable analytical characteristics, what can be inferred about the proposition according to which these two documents have a common (two distinct) source(s) (i.e., they have been printed by the same machine)?

In the latter case, the term 'common source' refers to the proposition according to which the two documents have been printed with the same machine. The scenario is one in which there are two questioned documents but no potential source (i.e., a control item). The former setting is one in which exactly one control item (i.e., a potential source) and one questioned item are available.

The discussion will include value-of-evidence calculations using Bayesian networks that aggregate prior beliefs and that offer an interface to sets of real data. These data are briefly outlined in Section 2. The structure and the properties of the proposed Bayesian networks are described in Section 3. The paper ends with Section 4 which contains a discussion of the methodology and a conclusion.

Bayesian networks are sometimes viewed cautiously because they represent a rather compact approach with calculations that can be largely automated and performed invisible to users (with the aid of appropriate software tools). In order to illustrate that Bayesian network output is not arbitrary and agrees with general results known from existing forensic literature, both Bayesian networks and the corresponding Bayesian algebraic analyses will be presented throughout this paper.

* Corresponding author. Tel.: +41 21 692 46 00.
E-mail addresses: alex.biedermann@unil.ch (A. Biedermann), franco.taroni@unil.ch (F. Taroni), silvia.bozza@unive.it (S. Bozza).

0379-0738/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.forsciint.2010.05.001

This paper thus intends to continue and extend the discussion of forensic uses of Bayesian networks as initiated by publications in this [1,2, e.g.] and other journals [3–5, e.g.]. In particular, it focuses on a Bayesian approach to the evaluation of scientific evidence for a forensic domain (questioned document examination) in which, so far, statistical approaches are not as widely used as in other disciplines, such as DNA evidence, for example.

2. Data on analytical characteristics of black toner

2.1. Black toner samples

For the purpose of the current study, a total of 100 addressed envelopes were considered. These envelopes represent a random collection of letters received at the author's institution during the period 2003–2004. The address on each envelope was printed with a black dry toner. The manufacturer and the model of the machines used for printing the individual addresses were unknown.

The black toner present on these documents was analysed by an optoelectronic system (Section 2.2.1) and by Fourier Transform Infrared Spectroscopy (Section 2.2.2).

2.2. Methods

2.2.1. Optoelectronic systems

Single-component toner powders contain magnetic material. When affixed to paper, that material is incorporated into the toner particles. Toners of this kind typically exhibit magnetic properties similar to other forms of magnetic printing. In the present study, the presence (absence) of magnetic properties was tested by means of a Mag optoelectronic system (Vildis, Russia). These analyses can be used to differentiate magnetic single-component toners from non-magnetic bi-component toners.

2.2.2. Fourier transform infrared (FTIR) spectroscopy

Polymer resins contained in dry black printer toners were analysed by microscopic attenuated total reflectance (ATR) with an internal reflection element (Germanium crystal) using a Digilab® Excalibur spectrometer. FTIR spectra were acquired from 4000 to 650 cm^{-1} with a resolution of 4 cm^{-1}, using 64 scans that were co-added for each spectrum. Microscopical ATR by IR was shown to be a fast, reproducible and non-destructive technique for analysing the kind of polymer resins encountered in black toner samples.

2.3. Results

Table 1 summarises the results obtained following the analyses of, respectively, the component type (magnetism) and the polymer resins (IR-category) of each of 100 samples of black toner. Notice that the summary values chosen to represent the results of each of the two analyses are – according to experience – fully reproducible (i.e., there are no 'no data' results). The resin classification 'other' is a placeholder for all results of resin analyses different from those of the (more common) categories 1–6. The observations are comparable to those reported in existing literature in the field [6,7, e.g.].

Table 1
Results obtained following the analyses of, respectively, the component type (magnetism) and the polymer resins (IR-category) of each of 100 samples of black toner (taken from 100 addressed envelopes received at the author's institution during the period 2003–2004).

Resin group (No., name)	Counts among single component toners	Counts among bi-component toners
1. Styrene-co-acrylate	69	14
2. Epoxy A	8	3
3. Epoxy B	0	2
4. Epoxy C	0	1
5. Epoxy D	0	1
6. Polystyrene	0	1
7. Other	0	1
Total	77	23

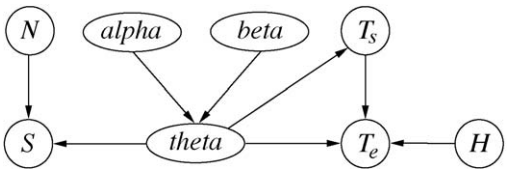

Fig. 1. Bayesian network for evaluating results obtained from the analysis of samples of black toner. The definition of the nodes is as given in Table 2.

3. Bayesian networks for evaluating results of black toner analyses in forensic document examination

3.1. One evidential document, one potential source

3.1.1. Results of optoelectronic analyses

Imagine a case that involves a single document of questioned origin. That document carries black bi-component toner impressions. A toner of the same type is used by a particular machine that could have served for printing the questioned document. A question that may be of interest in such a setting is how this analytical information ought to affect one's belief in the proposition according to which the questioned document was printed with the particular machine at hand—a so-called potential source.[1]

Such a proposition can be implemented as part of a Bayesian network, an example of which is proposed in Fig. 1. Here, the target propositions, that are

- H_p: the questioned document was printed with the potential source;
- H_d: the questioned document was not printed with the potential source;

are modeled in terms of a Boolean node H with states $t(true)$ and $f(false)$. For convenience, $H = t$ and $H = f$ will be abbreviated by H_p and H_d, respectively. Following standard literature in the field, the subscripts p and d are taken to reflect possible competing standpoints, such as those of the prosecution and the defence. The other nodes depicted in Fig. 1 are defined in Table 2.

The network fragment involving the nodes $N, S, alpha, beta$ and *theta* is analogous to the Bayesian network shown in Fig. 3 of Part I of this series of papers [10]. The node[2] *theta* represents the proportion of single-component toners. Uncertainty about that variable is modelled in terms of a beta distribution whose parameters are given by the numbered states of, respectively, the nodes *alpha* and *beta*. The nodes N and S provide a means to update the beta distribution by a binomial sample of size n, with s items found to be single-component toners.

Although reference is made here to a database that covers 100 entries (Table 1), the variable N is defined as one that covers states numbered from 96 to 110. This allows one to remain flexible with respect to situations with other database sizes. For the currently considered setting, for example, further cases may be added to the database up to 110.

[1] Conceptually speaking, the question of interest here is one of traditional forensic identification [8,9].

[2] For technical reasons, as explained in Part I, a discrete node with intervals is used to approximate the continuous variable θ. An alternative – beyond the scope of this paper – would consist in modelling the parameter outside the network through the use of an interface, such as RHugin (http://rhugin.r-forge.r-project.org/).

Table 2
Definition of the nodes used in the Bayesian network shown in Fig. 1.

Node	Description	States
alpha, *beta*	Parameters of the beta distributed variable θ (node *theta*)	$0.1, 0.5, 1, 2, 3, \ldots, 10$
H	The potential source (machine) was used to print the questioned document	$t(true)$, $f(false)$
N	Database size	$n = 96, 97, \ldots, 110$
theta	Proportion θ of single-component toners	$0 - 0.05, \ldots, 0.95 - 1$
S	Number of single-component toner samples counted in a database of size N	$s = 0, 1, 2, \ldots, 110$
T_e	A single-component toner is present on the questioned document	$t(true)$, $f(false)$
T_s	The potential source (printer) is a single-component toner	$t(true)$, $f(false)$

The node *theta* (the population proportion) acts as a conditioning for the Boolean node T_s, which models uncertainty about the potential source[3] being of type S (i.e., a single-component toner). That is, the probability of the potential source being of type S depends directly on the population proportion θ. The probabilities of the states *true* and *false* in the table of the node T_s are thus given by[4]:

```
Distribution(theta, 1-theta).
```
(1)

This is in analogy to the nodes d_i discussed in Part I [10] that model the outcome of items randomly drawn from a target population.

Let us notice that, while working with the Bayesian network, a state of the node T_s (*true* or *false*) will usually be set to 'known' (that is, the node will be instantiated). The directed link between the nodes *theta* and T_s will allow the distribution for the former node to be updated by knowledge about the truthstate of latter. The updated distribution for *theta* will subsequently serve the purpose of assessing the probability of T_e (i.e., the probability of the document in question bearing a single-component toner) in cases where H_d holds. Further details on this are given below.

The variable T_e is directly dependent upon three variables, that is H, T_s and *theta*. This structural specification allows one to express the following:

- If the document in question does truly come from the potential source (H_p), then necessarily the evidential toner is of the same type as that of the potential source:
$$Pr(T_e = t|H_p, T_s = t) = Pr(T_e = f|H_p, T_s = f) = 1. \quad (2)$$
This reflects the assumption according to which the target characteristic of a toner from a given machine will be stable and that this characteristic can be determined reliably by the scientist (e.g., there is no observational error).[5]
- If the document in question does truly come from another source (H_d), then the probability of the evidential toner being of type S is given by the appropriate population proportion of toners that are of type S (node *theta*): $Pr(T_e = t|H_d, \theta) = \theta$. Again, this is an instance where Expression (1) may be used.

The Bayesian network defined so far allows one to evaluate the ingredients of a likelihood ratio which provides a measure of the capacity of evidence T_e to discriminate between the competing states of the proposition H. Formally, that ratio writes:

$$LR = \frac{Pr(T_e|T_s, H_p)}{Pr(T_e|T_s, H_d)}. \quad (3)$$

This formula is considered here for settings where $T_e = T_s$. Following Eq. (2), the numerator $Pr(T_e|T_s, H_p)$ is 1.

Recall that the probability of a sample containing a single-component toner depends on the relevant population proportion θ of that toner type, and assume the uncertainty about θ is modeled through a beta probability density $Be(\alpha, \beta)$, with parameters augmented by a n-sized binomial sample among which s units are found to be single-component toners, that is $Be(\alpha + s, \beta + n - s)$. For a setting in which the questioned document and the potential source have single-component toners, the denominator of the likelihood ratio can be computed as

$$\begin{aligned} Pr(T_e = t|T_s = t, H_d) &= \int_0^1 Pr(T_e = t|T_s = t, H_d, \theta)\pi(\theta|T_s = t)d\theta \\ &= \int_0^1 \theta\pi(\theta|T_s = t)d\theta = E(\theta|T_s = t), \end{aligned} \quad (4)$$

where $\pi(\theta|T_s = t)$ is a beta posterior distribution with updated parameters $\alpha + s + 1, \beta + n - s$, because of the conditioning on the observation of the toner type of the potential source, which is of type S.

Analogously, for a setting in which the questioned document and the potential source have bi-component toners, the denominator is given by:

$$\begin{aligned} Pr(T_e = f|T_s = f, H_d) &= \int_0^1 (1 - \theta)\pi(\theta|T_s = f)d\theta \\ &= 1 - E(\theta|T_s = f), \end{aligned} \quad (5)$$

where $\pi(\theta|T_s = f) = Be(\alpha + s, \beta + n - s + 1)$.

In summary, thus, the denominator for the likelihood ratio for evidence on toner type is:

$$Pr(T_e|T_s, H_d) = \begin{cases} \dfrac{\alpha + s + 1}{\alpha + \beta + n + 1}, & T_e = T_s = t, \\ 1 - \dfrac{\alpha + s}{\alpha + \beta + n + 1}, & T_e = T_s = f. \end{cases} \quad (6)$$

For the purpose of illustration, consider a $Beta(1, 1)$ prior distribution for the parameter θ and a sample with $n = 100$ and $s = 77$ (Table 1). In such a setting, the following likelihood ratios are obtained:

$$LR = \begin{cases} \dfrac{1}{(\alpha + s + 1)/(\alpha + \beta + n + 1)} \approx 1.3, & T_e = T_s = t, \\ \dfrac{1}{1 - ((\alpha + s)/(\alpha + \beta + n + 1))} \approx 4.1, & T_e = T_s = f. \end{cases} \quad (7)$$

Fig. 2 provides a graphical illustration of the use of a Bayesian network for evaluating the denominator of the likelihood ratio in a case where the sample from the questioned document and the potential source are both found to be of type S.

3.1.2. A note on classification error

The discussion so far assumed that a sample can be classified as single or bi-component toner without any uncertainty in the classification. The reason for this is that the analyses of the target analytical characteristics generally lead to unambiguous results that can clearly and reliably be distinguished, at least on a technical level. Notwithstanding, false classifications may still occur (e.g., through a reporting or typing error) even if the generally 'clear'

[3] For the remainder of this paper, the expression 'potential source' will be used for short to designate the machine that is suspected of having been used to print the document in question.

[4] Throughout this paper, all expressions in Courier will refer to *Hugin* language (http://www.hugin.com/).

[5] It is part of a general understanding in that particular area of forensic examination (questioned document examination) that the analytical technique used by scientist can sufficiently be trusted, that is to work reliably and give accurate results. For settings in which it is desirable to relax this assumption, Bayesian networks can actually be extended so as to allow for uncertainty about the reliability of analytical results. This has been described, for instance, in [11] in the context of attribute sampling (see also Section 3.1.2).

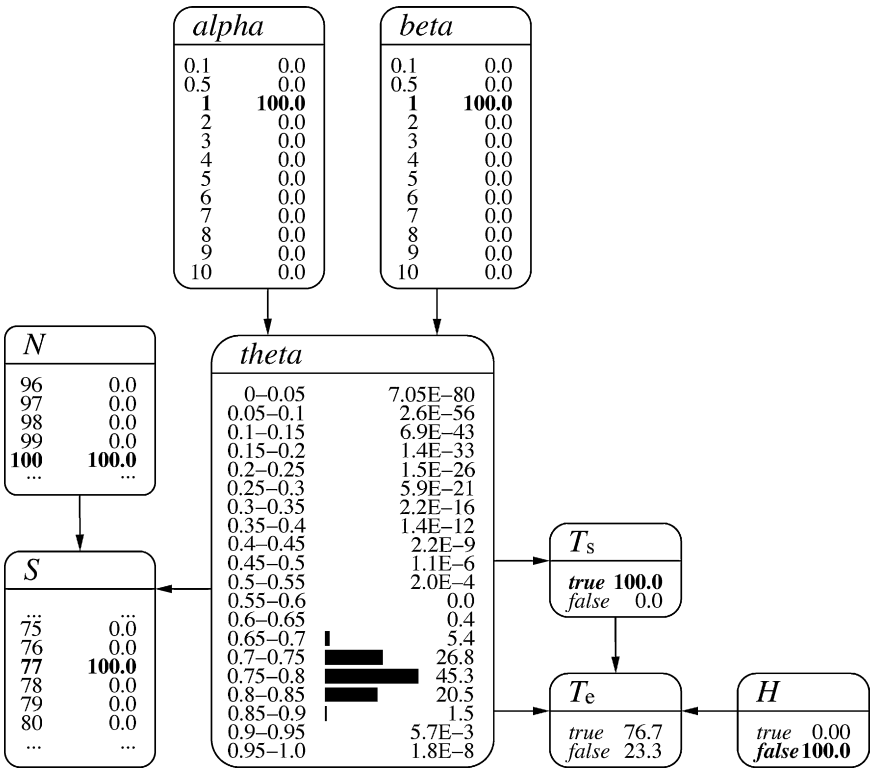

Fig. 2. Bayesian network for evaluating results of optoelectronic analyses. The definition of the nodes is as given in Table 2. Instantiations are shown in bold. The state *true* of the node T_e displays the value of the denominator of the likelihood ratio, that is $Pr(T_e|T_s, H_d)$, for a case in which the toner on a questioned document and a potential source are found to share the same characteristic (i.e., single-component toner).

results should not be 'missed' by a competent examiner. According to the case at hand, it may thus be desirable to incorporate a means to account for this source of uncertainty. One way to account for possible false classifications would be to introduce an additional hierarchy of nodes that distinguish between the *true* presence of an analytical characteristic (as assumed, for instance, by the nodes T_e and T_s) and the *observed* or *reported* analytical characteristic (represented by, for instance, nodes O_e and O_s). Then, with a structure of the kind $T_{e(s)} \rightarrow O_{e(s)}$, probabilities for false positives and false negatives can be used as entries for the conditional probability table of the node $O_{e(s)}$.[6] An example of this has been described, for instance, in the context of attribute sampling using Bayesian networks [11].

3.1.3. Results extended to FTIR analyses

Consider now a situation in which an evaluation needs to be extended to results of FTIR analyses. For this purpose, the Bayesian network shown in Fig. 1 has been extended to include two additional nodes, R_e and R_s (Fig. 3). These nodes have states numbered $i = 1, 2, \ldots, 7$, $j = 1, 2, \ldots, 7$ which represent the different resin groups defined earlier in Table 1 (see also Table 3).

Whether or not a toner contains a particular resin group can be generally thought of under the conditioning that the substance at hand is a single- or a bi-component toner. Consider this with regards to a potential source, for instance, that is a setting for which the Bayesian network shown in Fig. 3 allows one to specify the following conditional distributions:

$$Pr(R_s|T_s = t) = Dirichlet(\alpha_{s1} + x_{s1}, \ldots, \alpha_{s7} + x_{s7}) \text{ and}$$
$$Pr(R_s|T_s = f) = Dirichlet(\alpha_{b1} + x_{b1}, \ldots, \alpha_{b7} + x_{b7}),$$

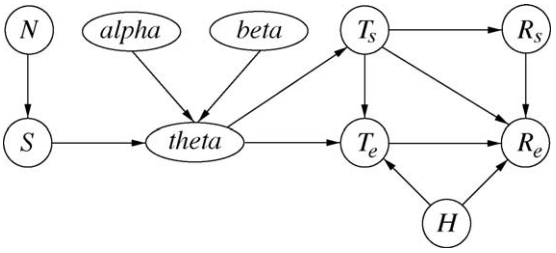

Fig. 3. Bayesian network for evaluating results obtained from the analysis of samples of black toner. The definition of the nodes is as given in Tables 2 and 3.

where α_{sj} and α_{bj} are the prior Dirichlet parameters ('prior observation counts') for the resin group j among, respectively, the single- and bi-component toners. The x_{sj} and x_{bj} denote the counts of toner samples of resin group j in a n-sized sample (in which there are s single-component toners and $n - s$ are bi-component toners). Thus, the entries j of the probability table of the node R_s are given by the mean of a Dirichlet distribution[7]:

$$Pr(R_s = j|T_s) = \frac{\alpha_{\cdot j} + x_{\cdot j}}{\sum_{j=1}^{7} \alpha_{\cdot j} + \sum_{j=1}^{7} x_{\cdot j}} \quad (8)$$

where $\alpha_{\cdot j}$ and $x_{\cdot j}$ refer to the prior Dirichlet parameters and the counts of toner samples of resin type j among, respectively, the single- ($T_s = t$) and bi-component ($T_s = f$) toners.

Consider now the probabilities that need to be specified for the table of R_e, the node which refers to the characteristics of the document in question. Clearly, if H_p holds, then the type of resin

[6] The underlying idea of this is analogous to a situation in which one accounts for the accuracy of a diagnostic test by specifying the sensitivity and the specificity [12–15, e.g.].

[7] See also Table 2 in Section 4.3 of Part I [10] for an example on how the mean of a Dirichlet distribution is calculated.

Table 3
Definition of the additional nodes R_e and R_s used in the Bayesian network shown in Fig. 3.

Node	Description	States
R_e	Resin group contained in the toner present on the questioned document	$i = 1, 2, \ldots, 7$
R_s	Resin group contained in the toner used by the potential source (machine)	$j = 1, 2, \ldots, 7$

contained in the toner on the document in question corresponds to that of the potential source, therefore (for $i, j = 1, 2, \ldots, 7$):

$$Pr(R_e = i|H_p, R_s = j) = \begin{cases} 1, & i = j \\ 0, & i \neq j. \end{cases}$$

If H_d holds (the evidential toner comes from another source), then the type of resin in the toner on the questioned document (node R_e) depends on whether that toner is a single- or a bi-component toner. For cases in which the type of the questioned toner's true source is different from that of the potential source, that is $T_e \neq T_s$, the probability table of R_e contains the following values (for $i = 1, 2, \ldots, 7$):

$$Pr(R_e = i|H_d, T_e, T_s) = \frac{\alpha_{\cdot i} + x_{\cdot i}}{\sum_{i=1}^{7} \alpha_{\cdot i} + \sum_{i=1}^{7} x_{\cdot i}}$$

where $\alpha_{\cdot i}$ and $x_{\cdot i}$ refer to the prior Dirichlet parameters and the counts of toner samples of resin type i among, respectively, the single- ($T_s = t$) and bi-component ($T_s = f$) toners.

Note that these probabilities do not depend on the resin type of potential source, R_s.

When the true source and the potential source are of the same toner type (i.e., single- or bi-component toners), then the following probabilities apply (for $i = 1, 2, \ldots, 7$):

$$Pr(R_e = i|H_d, R_s = j, T_e = T_s) = \begin{cases} \dfrac{\alpha_{\cdot i} + x_{\cdot i} + 1}{\sum_{i=1}^{7} \alpha_{\cdot i} + \sum_{i=1}^{7} x_{\cdot i} + 1}, & i = j, \\ \dfrac{\alpha_{\cdot i} + x_{\cdot i}}{\sum_{i=1}^{7} \alpha_{\cdot i} + \sum_{i=1}^{7} x_{\cdot i} + 1}, & i \neq j. \end{cases}$$

In summary, thus, the likelihood ratio for corresponding resin groups ($i = j$), conditional on knowing the type of toner involved (i.e., single- or bi-component) writes:

$$LR = \frac{Pr(R_e = i|H_p, R_s = i, T_e, T_s)}{Pr(R_e = i|H_d, R_s = i, T_e, T_s)} = \frac{1}{(\alpha_{\cdot i} + x_{\cdot i} + 1)/(\sum_{i=1}^{7} \alpha_{\cdot i} + \sum_{i=1}^{7} x_{\cdot i} + 1)}. \quad (9)$$

The Bayesian network proposed in Fig. 3 provides a convenient means to implement the individual components of Eq. (9)

Table 4
Likelihood ratios (values rounded) for results of FTIR analyses (analysis of a toner's resin component, where $i = j$) in cases involving toner samples from a single questioned document and a single potential source, conditional on knowing that these share the same general characteristics (i.e., both are either single- or bi-component toners, $T_e = T_s$). The values are obtained from Eq. (9) using a uniform Dirichlet prior distribution and data from Table 1.

Corresponding resin group	Toner type	
	Single component	Bi-component
1. Styrene-co-acrylate	1.2	1.9
2. Epoxy A	8.5	6.2
3. Epoxy B	42.5	7.8
4. Epoxy C	42.5	10.3
5. Epoxy D	42.5	10.3
6. Polystyrene	42.5	10.3
7. Other	42.5	10.3

using real data. Table 4 provides results of likelihood ratio calculations for results of FTIR analyses conditional on knowing the results from Section 3.1.1. Notice that, for some categories, identical numbers of observations are made in the sample (see data summarised in Table 1). As a consequence of this, along with the fact that equal prior probabilities have been used, the posterior probabilities of those categories with identical counts will also be equal.

The joint evidential value of optoelectronic and FTIR analyses can now be obtained as the individual likelihood ratio for the former category of evidence (Eq. (3)) times the likelihood ratio for the latter (Eq. (9)), that is, using notation introduced so far (for $R_e = R_s$, that is $i = j$):

$$\begin{aligned} LR &= \frac{Pr(R_e, R_s, T_e, T_s|H_p)}{Pr(R_e, R_s, T_e, T_s|H_d)} \\ &= \frac{Pr(R_e|R_s, T_e, T_s, H_p)}{Pr(R_e|R_s, T_e, T_s, H_d)} \times \frac{Pr(T_e|T_s, H_p)}{Pr(T_e|T_s, H_d)} \\ &= \begin{cases} \dfrac{\sum_{i=1}^{7} \alpha_{si} + \sum_{i=1}^{7} x_{si} + 1}{\alpha_{si} + x_{si} + 1} \times \dfrac{\alpha + \beta + n + 1}{\alpha + s + 1}, & T_e = T_s = t, \\ \dfrac{\sum_{i=1}^{7} \alpha_{bi} + \sum_{i=1}^{7} x_{bi} + 1}{\alpha_{bi} + x_{bi} + 1} \times \left(1 - \dfrac{\alpha + s}{\alpha + \beta + n + 1}\right)^{-1}, & T_e = T_s = f. \end{cases} \end{aligned} \quad (10)$$

Table 5 provides a summary of the various likelihood ratios that are obtained from uniform beta and Dirichlet prior distributions and the data presented in Table 1.

In a Bayesian network, these results may be obtained by taking, for example, the quotient the posterior probabilities of the states of the variable H—assuming equal prior probabilities for the states of H and appropriate instantiations made in the other nodes. Fig. 4 illustrates this for a case in which a bi-component toner with a styrene-co-acrylate resin was found on

Table 5
Likelihood ratios (values rounded to one decimal) for results of optoelectronic ($T_e = T_s$) and FTIR analyses ($R_e = R_s$). The values are obtained from Eq. (10) using uniform beta and Dirichlet prior distributions and data from Table 1.

Corresponding resin group	Toner type	
	Single component	Bi-component
1. Styrene-co-acrylate	1.6	8.0
2. Epoxy A	11.1	25.5
3. Epoxy B	55.4	31.9
4. Epoxy C	55.4	42.6
5. Epoxy D	55.4	42.6
6. Polystyrene	55.4	42.6
7. Other	55.4	42.6

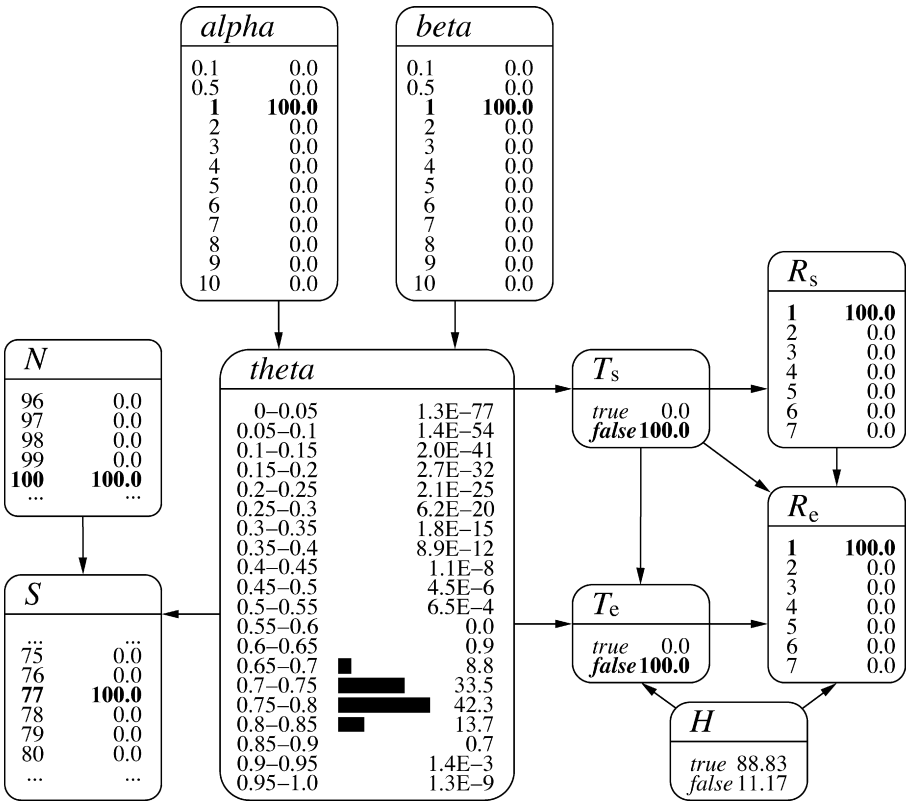

Fig. 4. Example of the use of a Bayesian network for the joint evaluation of results of optoelectronic and FTIR analyses. The definition of the nodes are as given in Tables 2 and 3. The model provides the posterior probabilities for the states of the node *H*. Instantiations are shown in bold.

a questioned document as well as in a machine that could have been used to print that document. In order to assure agreement with the results provided in Table 5, equal beta and Dirichlet prior distributions are assumed along with a database of cases as summarized in Table 1. Notice again that, for reasons explained in the previous section, some categories have equal posterior probabilities.

Generally, the likelihood ratio values summarised in Table 5 can be interpreted as factors by which particular evidence supports a proposition versus another. For example, in a case where a bi-component toner with an Epoxy A resin is found on both a questioned document and a potential source, a likelihood ratio of about 25 (Table 5) means that the probability of the scientific evidence at hand is 25 times greater if the document was printed with the examined printing device (the potential source) than with another printing device.

3.2. Two questioned documents, no potential source

3.2.1. Results of optoelectronic analyses

Forensic scientists may also be asked to conduct comparative toner analyses in the absence of a potential source. Common situations in which this may be of interest are those that involve two questioned documents about which one is unsure whether or not they could have come from the same source. For such a scenario, the development of the individual components of a likelihood ratio differs slightly from the approach discussed in Section 3.1.1 in the sense that there is no conditioning on analytical results from a reference sample, that is from a potential source [16]. Instead, the scientist needs to consider two sets of results, those obtained from the toners present on, respectively, the first and the second questioned document.

More formally, the likelihood ratio defined earlier in Section 3.1.1, Eq. (3), may be re-written as follows:

$$LR = \frac{Pr(T_1, T_2|H_p)}{Pr(T_1, T_2|H_d)} \tag{11}$$

where T_1 and T_2 are Boolean variables[8] that represent propositions according to which the toners present on, respectively, the first and the second questioned document are single-component toners. The definition of the variable H is changed here to:

- H_p : 'the two questioned documents come from the same source';
- H_d : 'the two questioned documents come from two distinct sources'.

Let us recall from Section 3.1.1 that the numerator of the likelihood ratio (3) for $T_e = T_s$ was one, that is, given the questioned document coming from the potential source (H_p being true) and knowing the toner characteristics of that source (T_s), then - assuming constant source characteristics - surely the questioned document will have the same characteristics.

In the currently discussed scenario, one still needs this probability of one item of evidence being of a certain toner type, given knowledge of the toner type of the other item of evidence and knowing that both come from the same source. But, one needs to multiply this probability by the probability of that common source consisting of a toner of the observed type. Using notation introduced so far and probabilistic modelling

[8] In Section 3.1, the notation T_e and T_s was used.

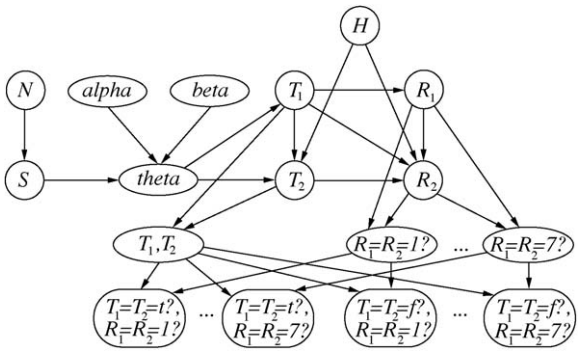

Fig. 5. Bayesian network for the evaluation of optoelectronic and FTIR analyses in scenarios with two items of evidence but no potential source. Node definitions are given in Table 6, except for the nodes $N, S, alpha, beta$ and *theta*, which have been defined in Table 2.

assumptions encoded in Fig. 5, one has, for the numerator of Eq. (11):

$$Pr(T_1, T_2|H_p) = Pr(T_2|T_1, H_p) \times Pr(T_1) = \begin{cases} 1 \times \dfrac{\alpha+s}{\alpha+\beta+n} = \dfrac{\alpha+s}{\alpha+\beta+n}, & T_1 = T_2 = t, \\ 1 \times \left(1 - \left(\dfrac{\alpha+s}{\alpha+\beta+n}\right)\right), & T_1 = T_2 = f. \end{cases} \quad (12)$$

The Bayesian network discussed earlier in Section 3.1 (Fig. 1) does not allow one to evaluate this probability directly because there are distinct nodes for T_1 and T_2. The probability of both T_1 and T_2 being true (or false) can be obtained, however, by introducing a summary node 'T_1, T_2' (as proposed in Fig. 5) with states tt, tf and ff. This allows one to account for the logical combinations of the states of the nodes T_1 and T_2.

The denominator is obtained as follows:

$$Pr(T_1, T_2|H_d) = Pr(T_2|T_1, H_d) \times Pr(T_1) = \begin{cases} \dfrac{\alpha+s+1}{\alpha+\beta+n+1} \times \dfrac{\alpha+s}{\alpha+\beta+n}, & T_1 = T_2 = t, \\ \left(1 - \dfrac{\alpha+s}{\alpha+\beta+n+1}\right) \times \left(1 - \dfrac{\alpha+s}{\alpha+\beta+n}\right), & T_1 = T_2 = f. \end{cases} \quad (13)$$

Because of the factor $Pr(T_1)$ that cancels in the numerator and the denominator, the overall value of the likelihood ratio is the same as that obtained in Section 3.1.1 (Eq. (6)). Eqs. (12) and (13) are needed, however, to find values for the individual components of the likelihood ratio.

3.2.2. Results of FTIR analyses

The likelihood ratio for results of FTIR analyses, given knowledge of the type of toner involved, is (assuming $R_1 = R_2$ and $T_1 = T_2$):

$$LR = \frac{Pr(R_1, R_2|T_1, T_2, H_p)}{Pr(R_1, R_2|T_1, T_2, H_d)} = \frac{Pr(R_2|R_1, T_1, T_2, H_p)}{Pr(R_2|R_1, T_1, T_2, H_d)} \times \frac{Pr(R_1|T_1, T_2, H_p)}{Pr(R_1|T_1, T_2, H_d)}, \quad (14)$$

where R_1 and R_2 refer to the kind of resin found in the toner from the first and the second document, respectively (in Section 3.1, the notation R_e and R_s was used).

As may be seen, the overall value of the likelihood ratio is the same as that given by Eq. (9). This is because the second ratio in Eq. (14) is in fact 1. For the separate evaluation of the numerator and the denominator, however, a factor $Pr(R_1|T_1, T_2, H)$ needs to be accounted for (Eq. (8)). The assessment of an observation that the toner present on both questioned documents contains a resin of type i can be assisted by additional nodes as proposed in the Bayesian network shown in Fig. 5. These nodes, labelled $R_1 = R_2 = i$ with $i = 1, 2, \ldots, 7$, are logical combinations of the nodes R_1 and R_2. The table of a node $R_1 = R_2 = i$ can be defined using an expression of the form `and(R1==i,R2==i)`, where $i = 1, 2, \ldots, 7$.

Table 6
Definition of nodes used in the Bayesian network shown in Fig. 5. Definitions of nodes not contained in this table are as given earlier in Table 2. For shortness of notation, letters t and f are used for designating the states *true* and *false*.

Node	Description	States
H	The two questioned documents have a common source	t, f
R_1	Resin group contained in the toner present on the first questioned document	$i = 1, 2, \ldots, 7$
R_2	Resin group contained in the toner present on the second questioned document	$j = 1, 2, \ldots, 7$
T_1	A single-component toner is present on the first questioned document	t, f
T_2	A single-component toner is present on the second questioned document	t, f
T_1, T_2	Toner type(s) present on the two questioned documents	tt, tf, ff
$R_1 = R_2 = i?$	Resin of type i is present on both questioned documents ($i = 1, \ldots, 7$)	t, f
$T_1 = T_2 = \{t, f\}?, R_1 = R_2 = i?$	Toner type and resin group present on both questioned documents	t, f

The likelihood ratio for the joint evaluation of results of optoelectronic and FTIR analyses is given by the product of the component likelihood ratios obtained above:

$$LR = \frac{Pr(R_1, R_2, T_1, T_2|H_p)}{Pr(R_1, R_2, T_1, T_2|H_d)} = \frac{Pr(R_2|R_1, T_1, T_2, H_p)}{Pr(R_2|R_1, T_1, T_2, H_d)} \times \frac{Pr(T_2|T_1, H_p)}{Pr(T_2|T_1, H_d)}. \quad (15)$$

The reader can see that by substituting in Eq. (15) results from Eqs. (9), (12), (13), that the likelihood ratio is identical to that obtained with one evidential document and one potential source, Eq. (10).

Within a Bayesian network, the respective calculations can be made by adding a minor extension to the graphical model discussed so far. In fact, the separate summary nodes for evaluating, respectively, a sample's component and resin type,

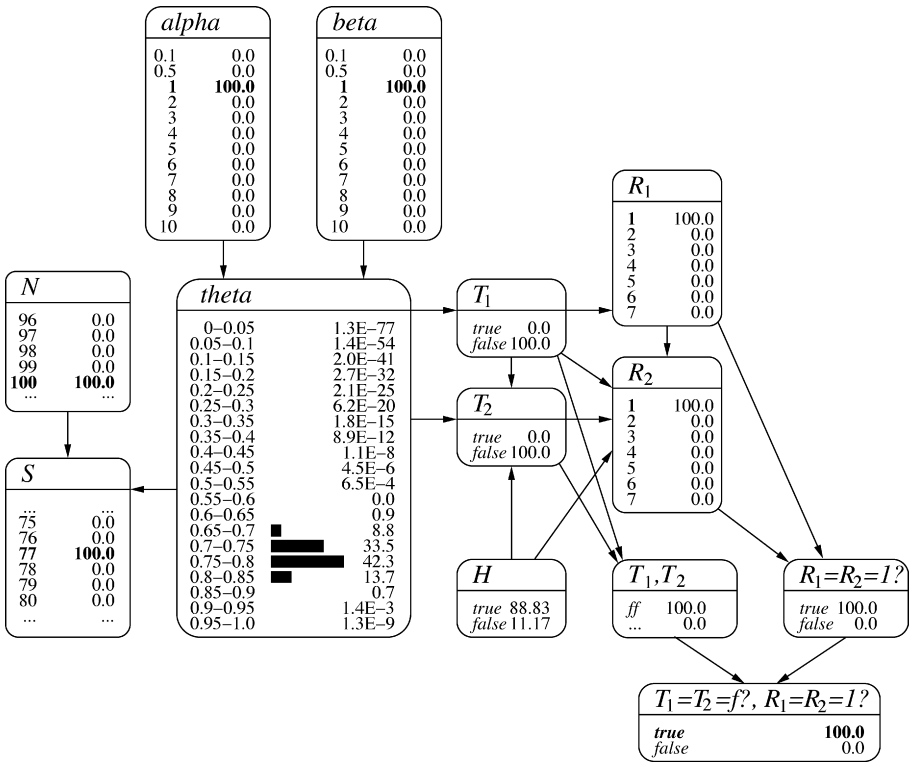

Fig. 6. Expanded representation of part of the Bayesian network shown in Fig. 5. The node *H* displays the posterior probabilities for two questioned samples originating from a common source in a scenario where the samples are found to be bi-component toners with a resin of type 1. Instantiations are shown in bold.

can be logically combined into further summary nodes. Such nodes are shown at the bottom of Fig. 5. For example, the Boolean node '$T_1 = T_2 = t?, R_1 = R_2 = 1?$' allows one to approach a setting in which the two questioned samples consist of a single-component toner that contains a resin of type 1. The node table can be completed using an expression (in Hugin language) of the form `and(T1T2==''tt'',R1R21)`. Here the term `T1T2==''tt''` expresses the condition that the node T_1, T_2 is in the state *tt* whereas the term `R1R21` specifies the truth of the node $R_1 = R_2 = 1?$ (see also Table 6).

Fig. 6 shows an expanded network with a selection of nodes needed for evaluating a scenario in which the two samples contain a bi-component toner along with a resin group 1. The node *H* shows the posterior probabilities for the proposition according to which the two samples come from the same source (assuming the same prior- and database-settings as in the previous Sections).

It is clear from these considerations that, once a Bayesian network model is fully specified, any combination of observations can be readily evaluated.

4. Discussion and conclusions

Laser printed or copied documents enjoy a widespread use in many parts of daily life. It is thus not uncommon to encounter them in various sorts of criminal activities such as revealing of compromising, injurious or confidential material, fraud or counterfeiting. Evaluating possible relationships between, for instance, a questioned document and a potential source, or between two questioned documents, thus represents a topic of ongoing interest. More generally, these topics also hold a relevant position among the forensic science disciplines that reunite to support current initiatives (e.g., counterterrorism) through an intelligence perspective.

There is considerable analytical literature that illustrates the potential gain that may be expected from investigating various toner components. These may be organised in computer searchable libraries and assist scientists in assigning, for example, a questioned toner sample to a particular group [6]. This may then allow to set up a list of machines whose toners have comparable analytical characteristics. Such analysis can support ongoing investigations in particular when there is a need to reduce the population of machines that could potentially be at the origin of a questioned sample.

Additional analyses and argument are necessary, however, if one wishes to address particular propositions of interest about which uncertainty exists (e.g., the proposition according to which a questioned sample originates from a particular source). Generally, uncertainty about propositions is most appropriately expressed through probability, a concept that allows one to derive rigourous evaluative procedures, both in forensic science [13, e.g.] and also in many other applications [15, e.g.]. However, the derivation of a formal procedure may involve considerable technicalities, often difficult to adjust whenever the initially stated target question or purposes change. In addition, further complication may arise if one wishes the numerical specification of a particular model to relate experimental data.

With the aim of illustrating that all of these facets of an inference problem can be approached within a unified framework, this series of papers has focused on the discussion of Bayesian networks, from both a conceptual and a practical point of view. Conceptually, the approach allows one to account for the essential ingredients of the Bayesian approach to statistics, that is:

- a parameter of interest can be considered as a random variable;
- inferences about a parameter can be made using the rules of probability;
- probability statements about a parameter can be based on (subjective) prior information which may subsequently be revised in the light of newly acquired data, using Bayes' theorem.

The proposed Bayesian networks can be used at several levels of detail. Local network components (sub-networks) may be retained for implementing probability 'learning' procedures that themselves maybe part of a larger network used for the analysis of particular case scenarios. Generally, the Bayesian network environment assures that the two aspects neatly interface and work within a coordinated whole.

From a practical point of view, a Bayesian network approach offers various useful features. In particular, it allows one:

- to avoid the technicalities of managing and manipulating different (possibly cumbersome) probabilistic formulae (instead, efforts can be concentrated on structural issues and underlying assumptions);
- encode, store, communicate and share domain knowledge by translating probabilistic inference procedures into graphical representations;
- to interface an inference model to real data and to learn target probabilities from that data;
- to readily adjust parametric settings (e.g., distributive assumptions) without the need to care about underlying calculations;
- to implement Bayesian principles for both parameter estimation and inference about propositions.

The availability of Bayesian network software packages provides further support in extending these ideas to practical applications (e.g., for analysing how the value of several distinct items of scientific evidence may be assessed). In addition, it opens opportunities to explore categories of scientific evidence from other forensic disciplines.

References

[1] P. Garbolino, F. Taroni, Evaluation of scientific evidence using Bayesian networks, Forensic Science International 125 (2002) 149–155.
[2] A.P. Dawid, J. Mortera, P. Vicard, Object-oriented Bayesian networks for complex forensic DNA profiling problems, Forensic Science International 169 (2007) 195–205.
[3] C.G.G. Aitken, A. Gammerman, Probabilistic reasoning in evidential assessment, Journal of the Forensic Science Society 29 (1989) 303–316.
[4] A.P. Dawid, I.W. Evett, Using a graphical method to assist the evaluation of complicated patterns of evidence, Journal of Forensic Sciences 42 (1997) 226–231.
[5] A.P. Dawid, J. Mortera, V.L. Pascali, D. van Boxel, Probabilistic expert systems for forensic inference from genetic markers, Scandinavian Journal of Statistics 29 (2002) 577–595.
[6] R.A. Merrill, E. Bartick, J.H. Taylor III, Forensic discrimination of photocopy and printer toners. I. The development of an infrared spectral library, Analytical and Bioanalytical Chemistry 376 (2003) 1272–1278.
[7] Info-markt, Info-markt Guide to Laserprinters, FACTS GmbH, Düsseldorf, Germany, 1995–2003.
[8] D.J. Balding, P. Donnelly, Inference in forensic identification, Journal of the Royal Statistical Society Series A 158 (1995) 21–53.
[9] A.P. Dawid, J. Mortera, Coherent analysis of forensic identification evidence, Journal of the Royal Statistical Society, Series B 58 (1996) 425–443.
[10] A. Biedermann, S. Bozza, F. Taroni, Implementing statistical learning methods through Bayesian networks (Part I): a guide to Bayesian parameter estimation using forensic science data, Forensic Science International 193 (2009) 63–71.
[11] A. Biedermann, F. Taroni, S. Bozza, C.G.G. Aitken, Analysis of sampling issues using Bayesian networks, Law, Probability and Risk 7 (2008) 35–60.
[12] D.H. Kaye, The validity of tests: Caveant omnes, Jurimetrics Journal (1987) 349–361.
[13] B. Robertson, G.A. Vignaux, Interpreting Evidence. Evaluating Forensic Science in the Courtroom, John Wiley & Sons, Chichester, 1995.
[14] D.J. Balding, Weight-of-Evidence for Forensic DNA Profiles, John Wiley & Sons, Chichester, 2005.
[15] D. Lindley, Understanding Uncertainty, John Wiley & Sons, Hoboken, NJ, 2006.
[16] F. Taroni, S. Bozza, A. Biedermann, Two items of evidence, no putative source: an inference problem in forensic intelligence, Journal of Forensic Sciences 51 (2006) 1350–1361.