Aitken CGG, Lucy D, Zadora G, Curran JM. Evaluation of trace evidence for three-level multivariate data with the use of graphical models. Computational Statistics and Data Analysis. 2006; 50; 2571-2588.

Evaluation of transfer evidence for three-level multivariate data with the use of graphical models

C.G.G. Aitken*and D. Lucy,
School of Mathematics,
and Joseph Bell Centre for Forensic Statistics and Legal Reasoning,
The University of Edinburgh;

G. Zadora,
Institute of Forensic Research,
Cracow, Poland;

J.M. Curran,
Department of Statistics,
University of Waikato, New Zealand.

July 28, 2006

Running title: Three-level multivariate data and graphical models

*School of Mathematics, The King's Buildings, The University of Edinburgh, Mayfield Road, Edinburgh, EH9 3JZ; e-mail: c.g.g.aitken@ed.ac.uk, tel: +44 (0) 131 650 4877, fax: +44 (0) 131 650 6553

Abstract

The evaluation of measurements on characteristics of trace evidence found at a crime scene and on a suspect is an important part of forensic science. There are commonly three levels of variation in such evidence. First, there is measurement error on the individual items. Then, individual items are gathered in groups and there is variation within and between groups. There are also commonly many variables which can be measured on the items, such as elemental or chemical composition. There are usually inadequate data to enable proper estimation of a full parametric model to be made. A method is described here for evaluating the evidence by means of a likelihood ratio. The likelihood ratio compares the probability of the measurements on the evidence assuming a common source for evidence from the crime scene and evidence associated with the suspect with the probability of the measurements on the evidence assuming different sources for the crime scene and suspect evidence. It is a well-documented measure of the value of the evidence. A three-level model for multivariate normal data is described. The structure of the data is determined through consideration of the inverse of the covariance matrix from which a graphical model may be determined. This enables a considerable reduction in the parameterisation from the full model whilst retaining a credible dependence structure, not recognised in a model which assumes full independence. The model for the structure of the data thus obtained provides a novel solution to a problem in forensic science where full independence is often assumed for multivariate data. The performances of the derived models are investigated on a data set provided by a European forensic science laboratory.

Keywords: Covariance matrix, evaluation of evidence, forensic science, graphical models, likelihood ratio, multivariate data.

1 Introduction

The primary focus of forensic science is the collection and interpretation of evidence for identification purposes. Over the last thirty years there have been substantial advances in analytical laboratory instruments and computing facilities which have lead to an increased ability to collect and store data from forensic evidence. Correspondingly, there has been an increase in methods for numerically (statistically) evaluating evidence associated with particular crimes. Evidence types which can be reliably summarized numerically, and for which there exist reasonable background information, are well suited to methods of Bayesian evidence interpretation.

Statistical methods are often used to infer identity of a common source for an evidentiary sample taken from a crime scene and a sample taken from a suspect. Unfortunately, the methods of choice have been those of classical hypothesis testing. As Robertson and Vignaux (1995) point out, these methods answer a pre-data question, namely 'What it the probability that these two samples match by chance alone?' rather than the post-data question 'How much more likely does this evidence make it that the accused was in contact with the crime scene rather than some-one else?'. It is this latter question, of course, that the court is interested in. As is well documented elsewhere (for example, Aitken and Taroni, 2004), the answer to this question can be well addressed by the likelihood ratio. The likelihood ratio considers a particular case and answers the post-data question as to how the evidence in the case alters the odds in favour of a given proposition (for example that the defendant was in contact with the crime scene). This paper describes a method for evaluating the likelihood ratio for multivariate evidence. The method is an extension of the two-level random effects models for normally distributed data proposed by Aitken and Lucy (2004). The models described there are best used for data with few dimensions because of the curse of dimensionality. A procedure for overcoming this curse using graphical models is described here. This procedure enables a data set of high dimension to be considered, without loss of information, as a product of mutually independent sets of low dimension. The example used for illustration is that of the elemental analysis of glass fragments. The method can, however, be applied to other evidential types of the same structure or in other areas, apart from forensic science, where source comparison is desired.

Many forensic laboratories have access to equipment (*e.g.*, scanning electron microscope coupled with energy dispersive X-ray detectors) which can quickly and accurately return elemental concentration information. Such machines have lead to an abundance of multivariate data. The methods described in this paper are applicable to three-level multivariate random effects data. In the forensic context, the levels correspond to measurement error (error relating to the precision of the instrument), within source error (variation of measurements made on the same object) and between source variation (differences between measurements made on different objects of the same evidential type). The results are compared using data from the analysis of the elemental composition of glass fragments. The composition of glass consists of major, minor and trace elements. The data used were not collected for evidential purposes, but rather to provide appropriate data for the testing of the analytical methods.

The SEM-EDX method does not give information on trace elements. Only major and minor elements are recorded. Many laboratories do not have access to equipment which will give information on trace elements. Many forensic scientists use elemental compositions to discriminate amongst glass samples in an intuitive way (Koons and Buscaglia, 2002). Thus, the example described here is one in which the data are from major and minor elements only. The method may be used by those laboratories which do not have information on trace elements.

Some laboratories, however, are able to measure the concentrations of trace elements in addition to the concentrations of major and minor elements. When concentrations of trace elements are available the method should be applied to the data set consisting of concentrations of major, minor and trace elements. The method can also be applied to the subset of the data consisting solely of the concentrations of major and minor elements. The results obtained from these two applications can then be compared. The results of the comparison will then provide an indication of the importance of the trace elements in the evaluation of glass elemental composition data.

1.1 Control and recovered data

A number, n_1 (≥ 1), of replicate measurements are taken of n_c items found at a crime scene which are assumed to come from a known source. These measurements are referred to as control data as the source, P_1, of the measurements is known. A number, n_2 (≥ 1 and not necessarily equal to n_1), of replicate measurements are taken of n_s items from a source P_2 of unknown origin. These measurements are referred to as recovered data. Often they are associated with a suspect and hence the subscript s will be used in reference to these data. The prosecution proposition, H_p, is that P_1 and P_2 are the same source and, therefore, the suspect is associated with the crime scene. The defence proposition, H_d, is that P_1 and P_2 are not the same source. Several characteristics are measured for each item.

1.2 Population database

Consider s groups with t members of each group, with $N = st$. The data used for the database in the comparison consist of q replicate measurements on each of the t items in each of the s groups. The data consist of p variables (characteristics), which are measured q times on each of t items on each of s groups.

1.3 Illustrative data

The data set used for illustration of the method consists of 130 ($s = 130$) groups of float glass. Float glass is composed of many elements. Those considered for analysis are the elements sodium (Na), magnesium (Mg), aluminium (Al), silicon (Si), calcium (Ca) and oxygen(O), none of which is a trace element. The procedures described here for the elemental analysis of glass fragments assume the measurements have a multivariate normal distribution. Descriptive data analysis shows that logarithms of the ratio of the concentration of oxygen to that of each of the other elements may be modelled reasonably by a normal distribution. These ratios are $\log_{10}(Na/O), \log_{10}(Mg/O), \log_{10}(Al/O), \log_{10}(Si/O)$ and $\log_{10}(Ca/O)$ and these are the $(p =)$ 5 characteristics used later.

2 Models

Let Ψ denote a population of p characteristics of items of a particular evidential type. Background data are available of measurements of these characteristics on a random sample of $N = st$ members from Ψ with q (≥ 2) independent replicate measurements on each of the N members. The background data are denoted as $\mathbf{x}_{ijk} = (x_{ijk1}, \ldots, x_{ijkp})^T$; $i = 1, \ldots, s$; $j = 1, \ldots, t$; $k = 1, \ldots, q$; with $\bar{\mathbf{x}}_{ij.} = \frac{1}{q}\sum_{k=1}^{q} \mathbf{x}_{ijk}$; $\bar{\mathbf{x}}_{i..} = \frac{1}{t}\sum_{j=1}^{t} \bar{\mathbf{x}}_{ij.}$; $\bar{\mathbf{x}}_{...} = \frac{1}{s}\sum_{i=1}^{s} \bar{\mathbf{x}}_{i..}$. The background data $\mathbf{x}$ are used to estimate various parameters of Ψ.

As introduced in Section 1.1, measurements are available from items found at a crime scene. These measurements are known as control data and are denoted $\mathbf{y}_1$. Measurements are available from items of an unknown origin. These measurements are known as recovered data and are denoted $\mathbf{y}_2$. For the control data, there are n_1 replicate measurements on each of n_c items. For the recovered or suspect data, there are n_2 replicate measurements on each of n_s measurements.

Define a subscript l which takes one of two values corresponding to whether the data are control $(l = 1)$ or recovered $(l = 2)$. The control and recovered measurements $\mathbf{y}_1, \mathbf{y}_2$ are vectors with elements $(\mathbf{y}_{ljk}, k = 1, \ldots, n_l;\ \ j = 1, \ldots, n_z;\ \ l = 1, 2;\ \ z = c \text{ if } l = 1, z = s \text{ if } l = 2)$ where $\mathbf{y}_{ljk} = (y_{ljk1}, \ldots, y_{ljkp})^T$. Let $n_0 = (n_1 n_c + n_2 n_s)$, $\bar{\mathbf{y}}_{lj} = \frac{1}{n_l} \sum_{k=1}^{n_l} \mathbf{y}_{ljk}$ and $\bar{\mathbf{y}}_l = \frac{1}{n_z} \sum_{j=1}^{n_z} \bar{\mathbf{y}}_{lj}$ and $\bar{\mathbf{y}} = (n_1 n_c \bar{\mathbf{y}}_1 + n_2 n_s \bar{\mathbf{y}}_2)/n_0$. The individual variable means over n_l measurements may be denoted as $\bar{\mathbf{y}}_{lj.r}$ for $j = 1, \ldots, n_z,\ \ r = (1, \ldots, p)$.

The model assumes three sources of variation, that of measurement error (replicate measurements on the same item), that between items within the same group (known as within-group variation) and that between groups (known as between-group variation). It is assumed that the variation at all three levels is normally distributed.

Replication: Denote the mean vector within item j in group i as θ_{ij} and the covariance matrix of replicate variability as U, constant over all items and groups. Then, given θ_{ij} and U, the distribution of $\mathbf{X}_{ijk}$ is taken to be normal with $(\mathbf{X}_{ijk} \mid \theta_{ij}, U) \sim N(\theta_{ij}, U), ;\ \ i = 1, \ldots, s;\ \ j = 1, \ldots, t;\ \ k = 1, \ldots, q$.

Within-group: Denote the mean vector within group i by μ_i and the within-group covariance matrix by V. Then, given $\mu_{\mathbf{i}}$ and V, the distribution of θ_{ij} is taken to be normal with $(\theta_{ij} \mid \mu_i, V) \sim N(\mu_i, V), ;\ \ i = 1, \ldots, s;\ \ j = 1, \ldots, t$.

Between-group: Denote the mean vector between groups by ϕ and the between-group covariance matrix by W. The distribution of the μ_i, as a measure of between-source variability, is taken to be normal with $(\mu_i \mid \phi, W) \sim N(\phi, W);\ \ i = 1, \ldots, s$.

Consider the control data $\mathbf{y}_{1jk}, k = 1, \ldots, n_1, j = 1, \ldots, n_c$. Then

$$
\begin{aligned}
(\mathbf{Y}_{1jk} \mid \theta_{i_1 j}, U) &\sim N(\theta_{i_1 j}, U), \\
(\bar{\mathbf{Y}}_{1j} \mid \theta_{i_1 j}, U) &\sim N(\theta_{i_1 j}, n_1^{-1} U), \\
(\bar{\mathbf{Y}}_{1j} \mid \mu_{i_1}, U, V) &\sim N(\mu_{i_1}, n_1^{-1} U + V), \\
(\bar{\mathbf{Y}}_{1} \mid \mu, U, V) &\sim N(\mu, (n_1 n_c)^{-1} U + n_c^{-1} V), \\
(\bar{\mathbf{Y}} \mid \phi, U, V, W) &\sim N(\phi, (n_1 n_c)^{-1} U + n_c^{-1} V + W),
\end{aligned}
$$

where $\theta_{i_1 j}$ is the mean of the replicate measurements on the j-th member of the control group. The control group is indicated i_1.

Analogous results follow for the recovered data with i_2 denoting the recovered group so that, for example, μ_{i_2} replaces μ_{i_1}. The means μ_{i_1} and μ_{i_2} are the means of the measurements for the groups (sources) from which $\mathbf{y}_1$ (P_1) and $\mathbf{y}_2$ (P_2) are taken. These groups are assumed to be members of Ψ but are not necessarily any of the groups which contribute to the background data $\mathbf{x}$. It is one of the important assumptions in the applications of statistics to forensic science for the evaluation of evidence that the population from which the background data are obtained is the same as that from which the control and recovered data are taken. Thus μ_{i_1} and μ_{i_2} may not be one of the group means $(\mu_i, i = 1, \ldots, s)$ for the background data $\mathbf{x}$. If P_1 and P_2 are the same source, then $\mu_{i_1} = \mu_{i_2}$.

3 Methods

3.1 Parameter estimation

The overall mean μ is estimated by $\bar{\mathbf{x}}$, the mean vector over all groups.

The measurement error (replicate error) covariance matrix U is estimated from the background data $\{\mathbf{x}_{ijk}\}$ by

$$\hat{U} = \frac{S_U}{\{st(q-1)\}} \tag{1}$$

where $S_U = \sum_{i=1}^{s} \sum_{j=1}^{t} \sum_{k=1}^{q} (\mathbf{x}_{ijk} - \bar{\mathbf{x}}_{ij.})(\mathbf{x}_{ijk} - \bar{\mathbf{x}}_{ij.})^T$.

The within-group covariance matrix V is estimated from the background data $\{\mathbf{x}_{ijk}\}$ by

$$\hat{V} = \frac{S_W}{\{s(t-1)\}} - \frac{\hat{U}}{q} \tag{2}$$

where $S_W = \sum_{i=1}^{s} \sum_{j=1}^{t} (\bar{\mathbf{x}}_{ij.} - \bar{\mathbf{x}}_{i..})(\bar{\mathbf{x}}_{ij.} - \bar{\mathbf{x}}_{i..})^T$.

The between-group covariance matrix W is estimated from the background data $\{\mathbf{x}_{ijk}\}$ by

$$\hat{W} = \frac{S_B}{(s-1)} - \frac{\hat{V}}{t} - \frac{\hat{U}}{tq}, \tag{3}$$

where $S_B = \sum_{i=1}^{s} (\bar{\mathbf{x}}_{i..} - \bar{\mathbf{x}}_{...})(\bar{\mathbf{x}}_{i..} - \bar{\mathbf{x}}_{...})^T$.

The estimate s_k of the pooled within-group standard deviation for variable k is the square root of the $k-$th term on the leading diagonal of $\hat{V}$; $(s_1, \ldots, s_p)^T$ is denoted $\mathbf{s}$.

3.2 Likelihood ratio using a multivariate random effects model and assumptions of normality

The value of the evidence $\mathbf{y}_1$ and $\mathbf{y}_2$ is the ratio of two probability density functions of the form $f(\mathbf{y}_1, \mathbf{y}_2 \mid \phi, U, V, W)$, one for the numerator, where H_p is assumed true, and one for the denominator, where H_d is assumed true. In the numerator the source means θ_1 and θ_2 are assumed equal (to θ, say) but unknown. In the denominator it is assumed that the source means θ_1 and θ_2 need not be equal.

In the numerator denote the probability density function by $f_0(\mathbf{y}_1, \mathbf{y}_2 \mid \phi, U, V, W)$. It is given by

$$\int_\mu \int_\theta f(\mathbf{y}_1 \mid \theta, U) f(\mathbf{y}_2 \mid \theta, U) f(\theta \mid \mu, V) f(\mu \mid \phi, W) d\mu d\theta,$$

where the four probability density functions are multivariate normal. The integral can then be shown to be equal to

$$f_0(\mathbf{y}_1, \mathbf{y}_2 \mid \phi, U, V, W) =$$

$$\mid 2\pi U \mid^{-\frac{1}{2}(n_1 n_c + n_2 n_s)} \mid 2\pi(V+W) \mid^{-1/2} \mid 2\pi\{(n_1 n_c + n_2 n_s)U^{-1} + (V+W)^{-1}\}^{-1} \mid^{1/2}$$
$$\exp\left\{ -\frac{1}{2}\sum_{j=1}^{n_c} tr(S_{1j}U^{-1}) - \frac{1}{2}\sum_{j=1}^{n_s} tr(S_{2j}U^{-1}) - \frac{1}{2}n_1 tr(S_1 U^{-1}) - \frac{1}{2}n_2 tr(S_2 U^{-1}) \right\}$$
$$\exp\left\{ -\frac{1}{2}(H_1 + H_2) \right\} \qquad (4)$$

where

$$\begin{aligned} H_1 &= (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)^T \left(\frac{n_1 n_c n_2 n_s U^{-1}}{n_0} \right) (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2), \\ H_2 &= (\bar{\mathbf{y}} - \phi)^T \big(\frac{U}{n_0} + (V + W) \big)^{-1} (\bar{\mathbf{y}} - \phi), \end{aligned}$$

$$
\begin{aligned}
S_l &= \sum_{j=1}^{n_z} (\bar{\mathbf{y}}_{lj} - \bar{\mathbf{y}}_l)(\bar{\mathbf{y}}_{lj} - \bar{\mathbf{y}}_l)^T; \quad (l = 1, 2), \\
S_{lj} &= \sum_{k=1}^{n_l} (\mathbf{y}_{ljk} - \bar{\mathbf{y}}_{lj})(\mathbf{y}_{ljk} - \bar{\mathbf{y}}_{lj})^T; \\
& \quad j = 1, \ldots, n_z; \quad (l = 1, 2).
\end{aligned}
$$

Further details are given in the Appendix.

The exponential term is a combination of two terms, H_1 which accounts for the difference $(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$ between the means of the measurements on the control and recovered items and H_2 which accounts for their rarity (as measured by the distance of the mean weighted by sample sizes from ϕ).

In the denominator, the probability density function, denoted $f_1(\mathbf{y}_1, \mathbf{y}_2 \mid \phi, U, V, W)$, is given by

$$
\begin{aligned}
& \int_\mu \int_\theta f(\mathbf{y_1} \mid \theta, U) f(\theta \mid \mu, V) f(\mu \mid \phi, W) d\mu d\theta \\
\times & \int_\mu \int_\theta f(\mathbf{y_2} \mid \theta, U) f(\theta \mid \mu, V) f(\mu \mid \phi, W) d\mu d\theta,
\end{aligned}
$$

where $\mathbf{y}_1$ and $\mathbf{y}_2$ are taken to be independent as the data are assumed to be from different sources. The integral $\int_\mu \int_\theta f(\mathbf{y}_1 \mid \theta, U) f(\theta \mid \mu, V) f(\mu \mid \phi, W) d\mu d\theta$ can be shown to be equal to

$$
f(\mathbf{y}_1 \mid \phi, U, V, W) =
$$

$$
\begin{gathered}
\mid 2\pi U \mid^{-\frac{1}{2} n_1 n_c} \mid 2\pi (V + W) \mid^{-1/2} \mid 2\pi [(n_1 n_c) U^{-1} + (V + W)^{-1}] \mid^{-1/2} \\
\exp \left\{ -\frac{1}{2} \sum_{j=1}^{n_c} tr(S_{1j} U^{-1}) - \frac{1}{2} n_1 tr(S_1 U^{-1}) \right\} \\
\exp \left\{ -\frac{1}{2} (\bar{\mathbf{y}}_1 - \phi)^T \left(\frac{U}{n_1 n_c} + (V + W) \right)^{-1} (\bar{\mathbf{y}}_1 - \phi) \right\}
\end{gathered} \tag{5}
$$

with an analogous result for $\int_\mu \int_\theta f(\mathbf{y}_2 \mid \theta, U) f(\theta \mid \mu, V) f(\mu \mid \phi, W) d\mu d\theta$.

Further details are given in the Appendix.

The value of the evidence is the ratio of (4) to the product of (5) and the analogous result for $\bar{\mathbf{y}}_2$. This is equal to the ratio of

$$\mid (V+W) \mid^{\frac{1}{2}} \mid \left[(n_1 n_c + n_2 n_s)U^{-1} + (V+W)^{-1}\right] \mid^{-\frac{1}{2}} \exp\{-\frac{1}{2}(H_1+H_2)\} \quad (6)$$

to

$$\mid \left(n_1 n_c U^{-1} + (V+W)^{-1}\right) \mid^{-\frac{1}{2}} \mid \left(n_2 n_s U^{-1} + (V+W)^{-1}\right) \mid^{-\frac{1}{2}} \exp\{-\frac{1}{2}(H_3+H_4)\} \quad (7)$$

where

$$\begin{aligned} H_3 &= (\bar{\mathbf{y}}_1 - \phi)^T \left[(n_1 n_c)^{-1} U + (V+W)\right]^{-1} (\bar{\mathbf{y}}_1 - \phi), \\ H_4 &= (\bar{\mathbf{y}}_2 - \phi)^T \left[(n_2 n_s)^{-1} U + (V+W)\right]^{-1} (\bar{\mathbf{y}}_2 - \phi). \end{aligned}$$

4 Partial correlation matrices and graphical models

The background data $\mathbf{x}_{ijk}$ from which the mean ϕ and covariance matrices U, V and W are estimated are multivariate. One of the criticisms levelled against the use of multivariate statistical techniques in forensic science is the lack of background data from which to estimate the parameters for these first and second order moments. In many cases it is assumed that the variables are independent in order to lead to a reduction in the number of parameters to be estimated (Koons and Buscaglia, 2002). An alternative approach is to use principal component analysis (Curran *et al.*, 1997a, b) but this leads to problems of interpretability for a legal audience. Indeed, any approach which leads to the omission of data will need careful justification in order to avoid the challenge of suppression of evidence.

An approach is described here in which the model for use with the numerator and denominator of the likelihood ratio is derived from a graphical model. All the data are retained, a structure is determined which is easily described and justified (unlike the independence assumption) and there are no linear combinations of the variables (unlike a principal component analysis). The model retains the advantage of interpretability whilst reducing the number of parameters which have to be estimated. The graphical model is determined from consideration of the inverse of the covariance matrix (see, for example, Whittaker, 1990). The inverse covariance matrix is scaled to give a new matrix Ω which has unit entries on the diagonal. The off-diagonal elements of Ω are the negatives of the partial correlation coefficients between the corresponding pairs of variables, given the remaining variables.

There are three covariance matrices U, V and W in the models considered here. The first, U, refers to measurement error variance. The other two only appear in the expression for the likelihood ratio (the ratio of (6) to (7)) in the form $(V + W)$. In the examples described later, it is the correlation matrix derived from $(V + W)$ which is inverted and scaled to derive Ω. The matrix $(V+W)$ is the sum of the within-group matrix V and the between-group matrix W. Matrix U represents the variability in the measurement error and so is under the control of the laboratory in which the measurements are made. Matrix

$(V + W)$ represents the natural variability of items and groups and is external to the laboratory. Matrix Ω is estimated from the background data $\mathbf{x}$ only.

Consideration of the possible values of the partial correlation coefficients, $(r_{ab}, a, b = 1, \ldots p, a \neq b$, say) enables consideration of various independence structures. For example, consider a new matrix Ω_{r_0}, in which all the absolute values of the off-diagonal terms (r_{ab}) in Ω less than a certain value, r_0 say, are set to zero $(0 \leq r_0 \leq 1)$. The matrix Ω_{r_0} then gives a model for the data.

1. $r_0 = 0$, $\Omega_0 = \Omega$ and the full data structure of the model is retained. All parameters in ϕ, U, V and W need to be estimated.

2. $r_0 = 1$, $\Omega_0 = I$. This corresponds to an assumption of full independence and is the model used by many forensic scientists.

3. $0 < r_0 < 1$. Such values of r_0 enable different independence structures with different numbers of parameters to be estimated.

These structures may be represented by a graphical model. The p variables are represented by nodes. Pairs of variables in which the partial correlation matrix is greater than r_0 are joined by edges from one node (variable) to the other. The resultant graph is directed following the principles described in Lauritzen and Spiegelhalter (1988) in order to fit a conditional probability model.

Note, it is perhaps fortunate that in the examples described here the correlation matrices derived from $(V + W)$ are positive definite and hence invertible. If this were not the case, then some reparameterisation may be needed (*e.g.*, Hayes and Hill, 1980) though such a process would be subject to the same criticism made earlier of reparameterisation.

5 Analysis

The data used for testing the method originated from 130 groups ($s = 130$) of float glass. There are four members ($t = 4$) in each group and four replicate measurements ($q = 4$) for each member. The original data of elemental concentrations are transformed into the following variables $\log_{10}(Na/O)$, $\log_{10}(Mg/O)$, $\log_{10}(Al/O)$, $\log_{10}(Si/O)$, $\log_{10}(Ca/O)$ for normality.

5.1 Summary statistics

Variables $\log_{10}(Na/O)$, $\log_{10}(Mg/O)$, $\log_{10}(Al/O)$, $\log_{10}(Si/O)$, $\log_{10}(Ca/O)$ are considered. The corresponding vector of observations from the training data is denoted $\mathbf{x}$. The overall mean, $\bar{\mathbf{x}}$, which is used as an estimate of ϕ, is

$$\bar{\mathbf{x}} = (-0.703, -1.364, -2.162, -0.162, -0.918)^T.$$

The measurement-error covariance matrix U is estimated by $\hat{U}$, from equation (1), as:

$$\hat{U} = \begin{pmatrix} 3.118 \times 10^{-5} & 1.803 \times 10^{-4} & -5.756 \times 10^{-6} & 1.450 \times 10^{-5} & -1.570 \times 10^{-5} \\ 1.803 \times 10^{-4} & 1.035 \times 10^{-2} & 3.573 \times 10^{-4} & -5.100 \times 10^{-5} & -1.173 \times 10^{-3} \\ -5.756 \times 10^{-6} & 3.573 \times 10^{-4} & 4.398 \times 10^{-2} & 6.309 \times 10^{-5} & 8.255 \times 10^{-5} \\ 1.450 \times 10^{-5} & -5.100 \times 10^{-5} & 6.309 \times 10^{-5} & 2.065 \times 10^{-4} & 3.206 \times 10^{-4} \\ -1.570 \times 10^{-5} & -1.173 \times 10^{-3} & 8.255 \times 10^{-5} & 3.206 \times 10^{-4} & 2.001 \times 10^{-3} \end{pmatrix}.$$

The within-group covariance matrix V is estimated by $\hat{V}$, from equation (2), as:

$$\hat{V} = \begin{pmatrix} 7.880 \times 10^{-5} & 5.049 \times 10^{-5} & -2.474 \times 10^{-5} & 2.252 \times 10^{-5} & 3.022 \times 10^{-5} \\ 5.049 \times 10^{-5} & 2.390 \times 10^{-4} & 4.861 \times 10^{-4} & 1.553 \times 10^{-4} & 2.369 \times 10^{-4} \\ -2.474 \times 10^{-5} & 4.861 \times 10^{-4} & 2.264 \times 10^{-3} & 3.463 \times 10^{-4} & 6.339 \times 10^{-4} \\ 2.252 \times 10^{-5} & 1.553 \times 10^{-4} & 3.463 \times 10^{-4} & 9.370 \times 10^{-4} & 1.672 \times 10^{-3} \\ 3.022 \times 10^{-5} & 2.369 \times 10^{-4} & 6.339 \times 10^{-4} & 1.672 \times 10^{-3} & 3.159 \times 10^{-3} \end{pmatrix}.$$

The between-group covariance matrix W is estimated by $\hat{W}$, from equation (3), as:

$$\hat{W} = \begin{pmatrix} 4.102 \times 10^{-4} & -1.205 \times 10^{-4} & -3.340 \times 10^{-4} & 5.155 \times 10^{-4} & 5.525 \times 10^{-4} \\ -1.205 \times 10^{-4} & 2.636 \times 10^{-2} & -7.508 \times 10^{-3} & 3.904 \times 10^{-4} & 8.568 \times 10^{-4} \\ -3.340 \times 10^{-4} & -7.508 \times 10^{-3} & 5.959 \times 10^{-2} & -9.019 \times 10^{-4} & 8.255 \times 10^{-5} \\ 5.155 \times 10^{-4} & 3.904 \times 10^{-4} & -9.019 \times 10^{-4} & 1.513 \times 10^{-3} & 2.242 \times 10^{-3} \\ 5.525 \times 10^{-4} & 8.568 \times 10^{-4} & -3.101 \times 10^{-3} & 2.242 \times 10^{-3} & 5.752 \times 10^{-3} \end{pmatrix}.$$

5.2 Determination of structure

¿From Section 5.1 the inverse matrix $(\hat{V} + \hat{W})^{-1}$

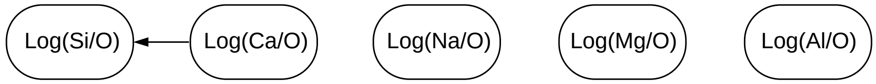
$$\begin{pmatrix} 2962.96 & 25.41 & 23.07 & -1161.69 & 319.67 \\ 25.41 & 39.48 & 5.20 & -17.08 & 2.43 \\ 23.07 & 5.20 & 17.27 & -21.83 & 12.22 \\ -1161.69 & -17.08 & -21.83 & 1832.43 & -732.72 \\ 319.67 & 2.43 & 12.22 & -732.72 & 416.18 \end{pmatrix},$$

and the scaled inverse matrix $\hat{\Omega}$

$$\begin{pmatrix} 1.000 & 0.074 & 0.102 & -0.499 & 0.288 \\ 0.074 & 1.000 & 0.199 & -0.064 & 0.019 \\ 0.102 & 0.199 & 1.000 & -0.123 & 0.144 \\ -0.499 & -0.064 & -0.123 & 1.000 & -0.839 \\ 0.288 & 0.019 & 0.144 & -0.839 & 1.000 \end{pmatrix}$$

are calculated in order to consider the structure for the relationship of the five variables. For example, the partial correlation coefficient of $\log_{10}(Ca/O)$ and $\log_{10}(Si/O)$ given $\log_{10}(Na/O)$, $\log_{10}(Mg/O)$ and $\log_{10}(Al/O)$ is 0.839. Let r denote a partial correlation coefficient. If it is decided that a pair of variables may be considered conditionally independent when the corresponding value of

Figure 1: A graphical model of data structure when $\mid r \mid < 0.5$.



r is such that $\mid r \mid < 0.5$ then the graphical model presented in Figure 1 represents the dependence structure. The corresponding scaled inverse matrix $\hat{\Omega}_{0.5}$ is represented as follows, with $*$ located in the cells for which the corresponding pairs of variables have a partial correlation coefficient less than 0.5 in absolute value.

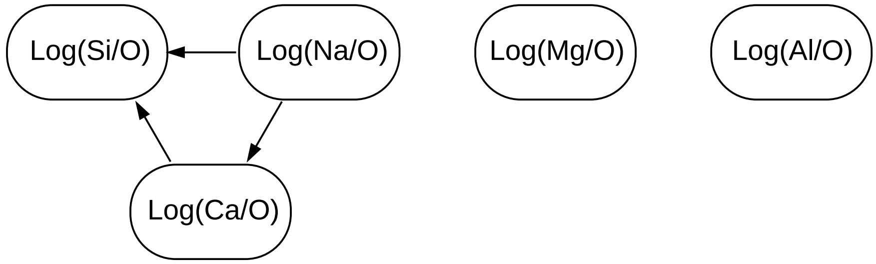
$$\hat{\Omega}_{0.5} = \begin{pmatrix} 1.000 & * & * & * & * \\ * & 1.000 & * & * & * \\ * & * & 1.000 & * & * \\ * & * & * & 1.000 & -0.839 \\ * & * & * & -0.839 & 1.000 \end{pmatrix} .$$

For ease of notation in the argument of density functions, the logarithm to base 10 of the ratio of an elemental concentration to the concentration of oxygen is just replaced by the name of the element, *e.g.*, the probability density function of $log_{10}(Si/O)$ is represented by $f(Si)$. Given the structure of Figure 1, the joint density function of $\log_{10}(Na/O)$, $\log_{10}(Mg/O)$, $\log_{10}(Al/O)$, $\log_{10}(Si/O)$ and $\log_{10}(Ca/O)$ may be represented as

$$f(Si, Ca) \cdot f(Na) \cdot f(Mg) \cdot f(Al) \ . \tag{8}$$

The structure for $\mid r \mid < 0.2$ can be obtained in a similar manner and is illustrated in Figure 2.

Figure 2: A graphical model of data structure when $| r |< 0.2$.



The matrix $\Omega_{0.2}$ is

$$\hat{\Omega}_{0.2} = \begin{pmatrix} 1.000 & * & * & -0.499 & 0.288 \\ * & 1.000 & * & * & * \\ * & * & 1.000 & * & * \\ -0.499 & * & * & 1.000 & -0.839 \\ 0.288 & * & * & -0.839 & 1.000 \end{pmatrix}.$$

The joint density function is

$$f(Na, Si, Ca) \cdot f(Mg) \cdot f(Al) \ . \tag{9}$$

The apparently arbitrary choice of a cut-off criterion may be criticised. However, some justification may be claimed when the resultant graphical model makes contextual sense. Thus, Figure 1 has no chemical justification whereas Figure 2 does. In Figure 2, the elements *Si, Na*, and *Ca* which are linked are major elements and the elements *Mg* and *Al* which are not linked are not major elements but modifiers (a component of the glass which changes the physico-chemical property of the object).

5.3 Experiment to consider the proposition that the fragments have the same source

The control and recovered data are determined from the training set in the following way to assess the performance of the method. All groups have four members ($t = 4$). One group is chosen. This group has four members on which replicated measurements were performed. The recovered sample is taken to be one of these four members ($n_s = 1$), here the first in the group, and the remaining three members are taken to the control group ($n_c = 3$). (The experiment was repeated with the second, the third and the fourth member of the group chosen as the recovered sample with similar results to those reported here.) There are replicate measurements on each member of the control and recovered groups. For the recovered group, there are $n_2 = 4$ replicate measurements on the one member of the group; for the control group there are $n_1 = 4$ replicate measurements also on each of the three members. There are 130 groups. Comparisons of control and recovered data are performed on the basis of the structures obtained for the correlation coefficients r with $| r |< 0.5$ and with $| r |< 0.2$ as well as for $r = 0$ (complete dependence of variables) and $r = 1$ (complete mutual independence of variables). Results are given in Table 1.

5.4 Experiment to consider the proposition that the fragments have different sources

One of the 130 groups is treated as the control with n_c items (fragments) and another group is treated as the recovered with n_s items (fragments). The particular comparison is that one item from the recovered group is compared with the set of all four items from the control group. (As for the experiment for fragments from the same source, the experiment was repeated with the other three members of the group chosen as the recovered sample with similar results to those reported here.) There are $129 \times 130/2$ (8385) comparisons.

Table 1: Comparison of the four models (structures) of evaluation of continuous multivariate data when control and recovered data come from the same source. First piece is recovered, second to fourth pieces are control.

Likelihood ratio	$\mid r \mid \leq 1$	$\mid r \mid < 0.5$	$\mid r \mid < 0.2$	$r = 0$
< 1	61	54	60	63
$1 - 10^1$	17	20	18	16
$10^1 - 10^2$	11	14	15	16
$10^2 - 10^3$	27	32	33	31
$10^3 - 10^4$	10	7	2	2
$10^4 - 10^5$	2	1	0	0
$10^5 - 10^6$	0	0	0	0
$10^6 - 10^7$	0	0	0	0
$10^7 - 10^8$	0	0	1	1
$10^8 - 10^9$	1	1	0	0
$10^9 - 10^{10}$	0	0	0	0
$> 10^{10}$	1	1	1	1
Total	130	130	130	130
True positive (%)	53.1	58.5	53.9	51.5
False negative (%)	46.9	41.5	46.1	48.5

The approach used simulates the practical situation when glass fragments n_s are found as debris collected from the examination of the clothes of a suspect. The experiment is designed so that the two sets of fragments did not come from the same source. They were chosen to originate from different sources. The control material n_c is more readily available in large amounts so several items (glass fragments) can be analyzed ($n_c > 1$). Comparisons of control and recovered data are performed on the basis of structures obtained when $\mid r \mid < 0.5$ and $\mid r \mid < 0.2$ as well as for $r = 0$ and $\mid r \mid \leq 1$. Results are given in Table 2.

Table 2: Comparison of the four models (structures) of evaluation of continues multivariate data when control and recovered data come from different sources. First piece from one group is treated as the recovered fragment, four pieces from a different group are treated as the control.

Likelihood ratio	$\mid r \mid \leq 1$	$\mid r \mid < 0.5$	$\mid r \mid < 0.2$	$r = 0$
> 1	618	590	557	539
$1 - 10^{-1}$	220	252	232	227
$10^{-1} - 10^{-2}$	214	222	226	206
$10^{-2} - 10^{-3}$	191	215	223	225
$10^{-3} - 10^{-4}$	182	199	222	224
$10^{-4} - 10^{-5}$	203	236	199	216
$< 10^{-5}$	6757	6671	6726	6748
Total	8385	8385	8385	8385
True negative (%)	92.6	93.0	93.4	93.6
False positive (%)	7.4	7.0	6.6	6.4

The method was also applied to a small test set of 10 groups, independent of the training set. There were four members in four of the groups, five members in four of the groups and six members in two of the groups. For comparisons within the same group, one member was taken as the recovered member and the remainder (three, four or five) were taken as the control members. There were ten such within-group comparisons. For comparisons between different groups, one member was taken as the recovered member and all members of the different group were taken as the control group. There were forty-five between-group comparisons.

Results on the ten within-group comparisons showed true positive rates of 80% when $\mid r \mid = 1$ (full independence) and of 90% for the other models ($r = 0.5, 0.2, 0$).

6 Conclusions

There is a large proportion of false negatives. Almost 50% of comparisons of fragments known to come from the same source have a likelihood ratio less than 1 and are classified as coming from different sources.

There is a small proportion of false positives. Less than 8% of comparisons of fragments known to have come from different sources have a likelihood ratio greater than 1 and are classified as coming from the same source.

The proportion of true positives is disappointingly low (about 50%) and the proportion of false negatives is correspondingly high. However, the proportion of false positives is also low. It is much more important that the proportion of false positives be low than that the proportion of false negatives be low. A likelihood ratio greater than 1 supports the proposition that the control and recovered fragments come from the same source and thus support the proposition that the person on whom the recovered fragments were found is associated with the crime scene. A likelihood ratio less than 1 supports the proposition that the control and recovered fragments come from different sources and thus support the proposition that the person on whom the recovered fragments were found is not associated with the crime scene. A false positive implies that a person on whom the recovered fragments have been found will be associated, incorrectly, with the crime scene. A false negative implies that a person on whom the recovered fragments have been found will be disassociated, incorrectly, with the crime scene. The burden of proof that the suspect is associated with the crime scene lies with the prosecution and thus it is more important for a scientific method to have a low false positive rate than a low false negative rate.

The high false negative rate is a reflection of the similar orders of magnitude for the estimates of the variances and covariances of the matrices U, V and W. These results are in contrast to the estimates of the within- and between-group matrices for the two-level models described in Aitken and Lucy (2004) where, in general, the entries in the estimate of the between-group matrix are two or three orders of magnitude greater than the entries in the estimate of the within-group matrix.

In practice, these data cannot be treated as fully independent ($\mid r \mid= 1$). It is not possible in glass for the levels of silicon, calcium and sodium to be independent of each other. The relationships amongst these three elements is clearly shown in the structure derived from $\mid r \mid< 0.2$, illustrated in Figure 2. The three relationships are not shown in Figure 1, derived from $\mid r \mid< 0.5$. The structure shown in Figure 1 is very similar to the full independence model $\mid r \mid= 1$ in which no nodes are linked. For these data it is therefore recommended that the structure derived from $\mid r \mid< 0.2$ be used in evaluating evidence from new cases.

The suggested value of 0.2 for $\mid r \mid$ is supported by another application of the two-level model with a nonparametric distribution described in Aitken and Lucy (2004). In this application, there were data on 200 groups of glass, each with four members. No account was taken of measurement error. The graphical model derived from $| r |< 0.2$ provided the closest representation to the chemical relationships amongst elements present in glass. The graphical model derived from $| r |< 0.5$ was again very similar to a full independence model.

The method is illustrated with five elements from glass analysis. There has been no analysis of data including trace elements. For those forensic science laboratories that are able to measure trace elements then further analyses to investigate the effects of their inclusion are recommended.

A method for the evaluation of evidence in the form of a three-level random effects model has been described. The method can be used for high-dimensional data through consideration of a structure determined by a graphical model obtained from a scaled inverse covariance matrix. A subjective choice based on chemical considerations has been used for the upper limit on the sample correlation coefficient, a limit below which the corresponding variables will be taken to be conditionally independent. The issue of multiple testing within a graphical model which may help with an objective choice of the cut-off for the sample correlation coefficient is discussed in Drton and Perlman (2004). The method provides a low proportion of false positives, acceptable in an application with legal ramifications as a false positive result corresponds to evidence against an innocent person. The method is offered as a credible alternative to the assumption by forensic scientists for many evidential types that the variables are independent.

7 Appendix

Some of the intervening mathematics used to derive expressions (4) and (5) is given here.

Completion of the square

The fundamental result is that of the completion of squares. The integrals evaluated are of the Normal probability density form and thus the algebra required for their solution is that of the completion of squares for multivariate data.

Consider vectors $\theta, \mathbf{a}$ and $\mathbf{b}$ and matrices A and B of appropriate order for the multiplications. Then the general result is that

$$(\theta - \mathbf{a})^T A(\theta - \mathbf{a}) + (\theta - \mathbf{b})^T B(\theta - \mathbf{b}) =$$
$$(\theta - \theta^*)^T (A + B)(\theta - \theta^*) + (\mathbf{a} - \mathbf{b})^T (A^{-1} + B^{-1})^{-1}(\mathbf{a} - \mathbf{b})$$

where

$$\theta^* = (A + B)^{-1}(A\mathbf{a} + B\mathbf{b}).$$

Derivation of the numerator of the likelihood ratio

The notation used here is the same notation as used in the body of the paper.

The numerator is the integral expression

$$\int_{\mu}\int_{\theta} f(\mathbf{y}_1 \mid \theta, U) f(\mathbf{y}_2 \mid \theta, U) f(\theta \mid \mu, V) f(\mu \mid \phi, W) d\mu d\theta, \tag{10}$$

First,

$$\int_{\mu} f(\theta \mid \mu, V) f(\mu \mid \phi, W) d\mu d\theta,$$

is the convolution of two multivariate normal distributions and is equal to

$$\mid 2\pi(V+W) \mid^{-1/2} \exp\big\{ -\frac{1}{2}(\theta-\phi)^T (V+W)^{-1}(\theta-\phi)\big\}.$$

The product

$$f(\mathbf{y}_1 \mid \theta, U) f(\mathbf{y}_2 \mid \theta, U)$$

is the product of the joint density functions of two sets of independent observations, namely $(\mathbf{y}_{1j}, j = 1, \ldots, n_c)$ and $(\mathbf{y}_{2j}, j = 1, \ldots, n_s)$, and can be written as

$$\prod_{j=1}^{n_c} f(\mathbf{y}_{1j} \mid \theta, U) \prod_{j=1}^{n_s} f(\mathbf{y}_{2j} \mid \theta, U).$$

The first set of products may then be written as

$$\prod_{j=1}^{n_c} \Big[\mid 2\pi U \mid^{-\frac{1}{2}n_1} \exp\big\{ -\frac{1}{2} tr(S_{1j}U^{-1}) - \frac{1}{2} n_1 (\theta - \bar{\mathbf{y}}_{1j})^T U^{-1} (\theta - \bar{\mathbf{y}}_{1j})\big\}\Big]$$

$$= \mid 2\pi U \mid^{-\frac{1}{2}n_1 n_c} \exp\left\{ -\frac{1}{2}\sum_{j=1}^{n_c} tr(S_{1j}U^{-1}) - \frac{1}{2} n_1 tr(S_1 U^{-1}) - \frac{1}{2} n_1 n_c (\bar{\mathbf{y}}_1 - \theta)^T U^{-1} (\bar{\mathbf{y}}_1 - \theta) \right\}.$$

Similarly the second set of products may be written as

$$=\mid 2\pi U \mid^{-\frac{1}{2}n_2 n_s} \exp\left\{-\frac{1}{2}\sum_{j=1}^{n_s} tr(S_{2j}U^{-1}) - \frac{1}{2}n_2 tr(S_2 U^{-1}) - \frac{1}{2}n_2 n_s(\bar{\mathbf{y}}_2-\theta)^T U^{-1}(\bar{\mathbf{y}}_2-\theta)\right\}.$$

Hence (10) may be written as

$$\int_\theta \mid 2\pi U \mid^{-\frac{1}{2}n_1 n_c} \mid 2\pi U \mid^{-\frac{1}{2}n_2 n_s} \mid 2\pi(V+W) \mid^{-1/2}$$

$$\exp\left\{-\frac{1}{2}\sum_{j=1}^{n_c} tr(S_{1j}U^{-1}) - \frac{1}{2}\sum_{j=1}^{n_s} tr(S_{2j}U^{-1}) - \frac{1}{2}n_1 tr(S_1 U^{-1}) - \frac{1}{2}n_2 tr(S_2 U^{-1})\right\}$$

$$\exp\left\{-\frac{1}{2}n_1 n_c(\bar{\mathbf{y}}_1-\theta)^T U^{-1}(\bar{\mathbf{y}}_1-\theta) - \frac{1}{2}n_2 n_s(\bar{\mathbf{y}}_2-\theta)^T U^{-1}(\bar{\mathbf{y}}_2-\theta) - \frac{1}{2}(\theta-\phi)^T(V+W)^{-1}(\theta-\phi)\right\} d\theta.$$

Let

$$\begin{aligned} G_{1c} &= n_1 n_c(\bar{\mathbf{y}}_1-\theta)^T U^{-1}(\bar{\mathbf{y}}_1-\theta) \\ G_{2s} &= n_2 n_s(\bar{\mathbf{y}}_2-\theta)^T U^{-1}(\bar{\mathbf{y}}_2-\theta) \\ G_3 &= (\theta-\phi)^T(V+W)^{-1}(\theta-\phi) \end{aligned}$$

Two operations of the general result for the completion of the square shows that

$$\int_\theta \exp\big\{-\frac{1}{2}(G_{1c}+G_{2s}+G_3)\big\}d\theta =$$

$$\mid 2\pi\{n_0 U^{-1} + (V+W)^{-1}\}^{-1} \mid^{\frac{1}{2}} \exp\{-\frac{1}{2}(\bar{\mathbf{y}}-\phi)^T\big(\frac{U}{n_0}+V+W\big)^{-1}(\bar{\mathbf{y}}-\phi) -$$
$$\frac{1}{2}(\bar{\mathbf{y}}_1-\bar{\mathbf{y}}_2)^T\big(\frac{U}{n_1 n_c}+\frac{U}{n_2 n_s}\big)^{-1}(\bar{\mathbf{y}}_1-\bar{\mathbf{y}}_2)\}.$$

The result (4) for the numerator then follows.

Derivation of the denominator of the likelihood ratio

The first integral of the denominator is

$$\int_\mu \int_\theta f(\mathbf{y}_1 \mid \theta, U) f(\theta \mid \mu, V) f(\mu \mid \phi, W) d\mu d\theta.$$

Using a similar argument to that for the numerator it can be shown that this integral may be written as

$$\mid 2\pi(V+W) \mid^{-\frac{1}{2}} \mid 2\pi U \mid^{-\frac{1}{2}n_1 n_c} \int_\theta \exp\Big\{ -\frac{1}{2}\sum_{j=1}^{n_c} tr(S_{1j}U^{-1}) - \frac{1}{2}n_1 tr(S_1 U^{-1}) \\ -\frac{1}{2}n_1 n_c(\bar{\mathbf{y}}_1 - \theta)^T U^{-1}(\bar{\mathbf{y}}_1 - \theta) - \frac{1}{2}(\theta - \phi)^T (V+W)^{-1}(\theta - \phi)\Big\} d\theta.$$

Use of the general result for the completion of the square gives the result for the first term of the denominator. The result for the second term of the denominator follows analogously. The three terms may then be put together as in the body of the paper to give the result for the likelihood ratio.

References

Aitken,C.G.G and Lucy,D., 2004. The evaluation of trace evidence in the form of multivariate data. Applied Statistics, 53, 109-122.

Aitken,C.G.G. and Taroni,F., 2004. Statistics and the evaluation of evidence for forensic scientists, (2nd edition). Wiley, Chichester.

Curran,J.M., Triggs,C.M., Almirall,J.R., Buckleton,J.S. and Walsh,K.A.J., 1997a. The interpretation of elemental composition from forensic glass evidence:

I. Science and Justice, 37, 241-244.

Curran,J.M., Triggs,C.M., Almirall,J.R., Buckleton,J.S. and Walsh,K.A.J., 1997b. The interpretation of elemental composition from forensic glass evidence: II. Science and Justice, 37, 245-249.

Drton,M. and Perlman,M.D., 2004. Model selection for Gaussian concentration graphs. Biometrika, 91, 591-602.

Hayes,J.F. and Hill,W.G., 1980. A reparameterization of a genetic selection index to locate its sampling properties. Biometrics, 36, 237-248.

Koons,R.D. and Buscaglia,J., 2002. Interpretation of glass composition measurements: the effects of match criteria on discrimination capability. Journal of Forensic Sciences, 47, 505-512.

Lauritzen,S. and Spiegelhalter,D.J., 1988. Local computations with probabilities on graphical structure and their application to expert systems, with Discussion. Journal of the Royal Statistical Society, Series B, 50, 157-224.

Robertson B. and Vignaux G.A., 1995. DNA evidence: wrong answers or wrong questions? In: B.S. Weir (Ed.), Human identification: the use of DNA markers, Kluwer Academic Publishers, Dordrecht, 145-152.

Whittaker,J., 1990. Graphical models in applied multivariate statistics. Wiley, Chichester.

Acknowledgements

We gratefully acknowledge the assistance of the Institute of Forensic Research, Cracow, for the provision of the data and helpful discussions of our models and interpretation. This research was supported by a Research Development Grant from the Scottish Higher Education Funding Council and by a visiting Leonardo da Vinci Fellowship of the European Union. Some results of elemental glass analysis were obtained within Project Number 0T00C 008 22 of the State Committee for Scientific Research, Poland, for one of us (GZ). Computer programmes for the various probability distributions referred to in the paper are available from Dr. David Lucy, e-mail d.lucy@ed.ac.uk. Comments from referees have been very helpful in improving the presentation of the paper.