The interpretation of elemental composition measurements from forensic glass evidence: II

JM CURRAN*, CM TRIGGS*

Department of Statistics, University of Auckland, Private Bag 90219, Auckland, New Zealand
e-mail: curran@stat.auckland.ac.nz or triggs@stat.auckland.ac.nz

JR ALMIRALL

Metro-Dade Crime Laboratory, Miami, Florida 33172, USA
e-mail: almirall@fiu.edu

and

JS BUCKLETON, KAJ WALSH

ESR, Private Bag 92-021, Auckland, New Zealand
e-mail: john.buckleton@esr.cri.nz or kevan.walsh@esr.cri.nz

Science & Justice 1997; 37: 245–249

Received 26 February 1997; accepted 26 June 1997

Previously the authors suggested the use of Hotelling's T^2 statistic, a multivariate equivalent of Student's t-test, for determining a 'match' between glass fragments recovered from a suspect and a 'control' sample of glass fragments from the crime scene [1]. The use of Hotelling's T^2 test was suggested as a replacement to either range overlap methods or '3 sigma' methods. While Hotelling's T^2 test is certainly a better approach than any '3 sigma' or range overlap method, it is still subject to the weaknesses inherent in any hypothesis testing approach. This paper will introduce a continuous Bayesian method based on Hotelling's T^2 test that overcomes these weaknesses.

In einer kürzlich erschienenen Arbeit haben die Autoren vorgeschlagen, für die Prüfung der Übereinstimmung von Glasspuren von einem Tatverdächtigen mit inkriminierten Glasspuren den Hotelling's T^2 Test anzuwenden, einem multivariaten Äquivalent zum Student's t-Test [1]. Es wurde vorgeschlagen, den Hotelling's T^2 Test anstatt der Überlappungsbereichmethoden oder der 3 Sigma-Methoden anzuwenden. Obwohl der Hotelling's T^2 Test sicherlich zu einer besseren Annäherung führt als jeder 3 Sigma-Test oder jede Überlappungsbereichmethode, ist dieser Test immer noch der Schwachpunkt, der jedem hypothetischen Näherungsverfahren innewohnt. Die Arbeit stellt eine weiterentwickelte Bayes Methode vor, die auf dem Hotelling's T^2 Test aufbaut und diesen Schwachpunkt eliminiert.

Les auteurs ont précédemment proposé l'utilisation des statistiques T^2 de Hotelling, un équivalent multivarié du test t de Student, pour déterminer une correspondance entre des fragments de verre retrouvés sur un suspect et un échantillon de contrôle provenant de la scène de crime [1]. L'utilisation du test T^2 de Hotelling a été suggéré en remplacement des méthodes de superposition des domaines ou méthodes dites '3 sigma'. Alors que le test T^2 de Hotelling représente certainement une meilleure approche que toute méthode '3 sigma' ou superposition des domaines, il souffre toujours des faiblesses inhérentes à toute approche de test d'hypothèse. Cet article introduit une méthode Bayesienne continue, basée sur le test T^2 de Hotelling, qui surmonte ces faiblesses.

Previamente los autores sugirieron el uso de la estadística T^2 de Hotelling, un equivalente multivariado del test de la t de Student, para determinar el grado de correlación entre los fragmentos de vidrio recogidos de un sospechoso y una muestra control de fragmentos de vidrio de la escena del crimen [1]. Se sugirió el uso del test de la T^2 de Hotelling para reemplazar tanto al método de rango de solapamiento como a los métodos de 3 sigma. Aunque este método del test de la T^2 de Hotelling es mejor que los anteriores, tiene una cierta debilidad inherente a cualquier método de enfoque con hipótesis. Este trabajo introduce un método bayesiano continuo sobre el test de la T^2 de Hotelling para salvar esta debilidad.

Key Words: Forensic science; Interpretation of evidence; Statistics; Glass; Hotelling's T^2 test.

*Corresponding authors

Introduction

If a window is broken tiny fragments of glass may be transferred to the clothing of the person breaking the window. If a crime is committed in the breaking of a window, these fragments may be used in evidence. Recent work in the characterisation of glass evidence by its elemental composition, as opposed to traditional refractive index (RI) based methods, has required a framework for the evaluation of this data.

The analysis of glass evidence consists of comparing the physical and chemical properties of a fragment retrieved from a suspect to a possible source of the glass and then assessing the value of that association. In the case where the fragments are sufficiently large, coincidental edges may be found or density, colour and thickness comparisons can be attempted. The typical glass transfer case, however, produces very small recovered fragments and the only analyses that can be performed are refractive index (RI) and elemental composition comparisons.

In a previous publication [1] the authors suggested the use of Hotelling's T^2 test, a multivariate equivalent of Student's t-test for determining a 'match' between glass fragments recovered from a suspect and a 'control' sample of glass fragments from the crime scene. Hotelling's T^2 test is for use with multivariate data consisting of concentrations of several different elements, rather than the single datum, the RI of a sample fragment. The use of Hotelling's T^2 test was suggested as a replacement to either range overlap methods or '3 sigma' methods.

A typical range overlap method compares each recovered measurement to the range of the control sample for each element. If any of the recovered measurements fall outside the control range, then the fragment is deemed to not match.

A typical '3 sigma' approach to the analysis of the data has been to compare the intervals defined by adding and subtracting three times the standard error (or three times sigma) of an element concentration to the mean concentration for the control and recovered samples respectively. If the intervals overlap for every discriminating element then samples are said to match. However, if any *one* interval does not overlap then the samples are said to not match.

While Hotelling's T^2 test, or any formal hypothesis test, is certainly a better approach than any '3 sigma' or range overlap method, it is still subject to the weaknesses inherent in hypothesis testing. Use of hypothesis tests in forensic science suffers from three main problems. The first is that a hypothesis test fails to incorporate relevant evidence, such as the relative frequency of the recovered glass and indeed the mere presence of glass fragments on the suspect. The second problem is what K Smalldon termed the 'fall off the cliff' effect. It seems illogical that if a probability of 0.989999 (1 – P-value, i.e. 1– 0.010001) is returned from a hypothesis test then the samples should be deemed to 'match' when a probability of 0.990001 (1 – P-value, i.e. 1–0.0099999) would be a 'non-match', particularly when these probabilities are calculated under distributional assumptions which (almost certainly) do not hold. The third problem is that hypothesis testing does not answer the question that the court is interested in. Robertson and Vignaux [2] argue that presentation of a probability answers the pre-data question 'What is the probability of a match if I carry out this procedure' rather than the post-data question 'How much does this evidence increase the likelihood that the suspect is guilty.' It is, of course, the latter that the court is interested in.

The suggestion of taking a Bayesian approach to these problems is by no means a novel approach [3]. For refractive index based data see Walsh et al. [4]. This paper will extend the continuous approach for multivariate (elemental composition) data.

The continuous likelihood ratio for elemental observations

Walsh et al. [4] discuss a case where a pharmacy window is broken. Fragments of glass were retrieved from two suspects, and compared with a sample of fragments from the crime scene on the basis of mean refractive index. Both recovered samples just fail a t-test and that is where the matter might end as a conventional approach would deem these to be a 'non-match'. However, there are a number of aspects contradictory to the conclusion that the fragments did not come from the crime scene window. Firstly, both offenders had a large number of fragments of glass on their person. Studies [5–6, J McQuillan and S McCrossan Personal Communication] have shown that finding large groups of glass fragments on clothing is a reasonably rare event on people unassociated with a crime. Pearson et al. [7] found more than three fragments of glass on 16 out of 100 suits taken from a local dry cleaner, however, no attempt was made to group the fragments into common sources. Lambert et al. [6] examined glass from 589 individuals involved in 405 cases. In the cases where matching glass was found, 36.6% of the suspects had more than three fragments of glass from one source. In the cases where no matching glass was found, 5.6% of the suspects had more than three fragments of glass from one source.

Laboratory examination of the pharmacy window suggested that the recovered fragments came from a flat float glass object – again a relatively rare event (somewhere between 1 and 3% [6]). In addition, paint flakes recovered from one of the suspects were unable to be distinguished from the paint in the window frame at the crime scene.

Thus the weight of the evidence supports the suggestion that both suspects were at the crime scene when the window was broken. This is a classic example of Lindley's paradox

[8], although Lindley himself refers to it as Jefferys' paradox. Although the samples fail the t-test, the results are still more likely if they had come from the same source than from different sources.

Walsh et al. [4] propose an extension to the ideas put forward by Evett and Buckleton [9] which retains grouping information while dropping the match/non-match approach. Evidence in this case represents the refractive index measurements made on the fragments recovered from the suspect and taken from the crime scene. They suggest that

$$\frac{\Pr(\text{Evidence}|\text{Contact})}{\Pr(\text{Evidence}|\overline{\text{Contact}})} = LR = T_0 + \frac{T_L P_0}{P_1 S_L} . lr_{cont} \quad (1)$$

where

$$lr_{cont} = \frac{f(\bar{X} - \bar{Y} \mid S_X, S_Y)}{\hat{g}(\bar{Y})} \quad (2)$$

and

T_0 = the probability that none of the glass fragments transferred have persisted and been recovered

T_L = the probability that three or more of the glass fragments transferred, have persisted and been recovered

P_0 = the probability of a person having no glass on their clothing

P_1 = the probability of a person having one group of glass on their clothing

S_L = the probability that a group of glass on clothing contains three or more fragments

$\hat{g}(\bar{Y})$ = the value of the probability density for float glass at the mean RI of the recovered sample, usually obtained from a density estimate

$f(\bar{X} - \bar{Y} \mid S_X, S_Y)$ = the value of the probability density for the difference of two sample means. This is simply a rescaled t-distribution using Welch's modification to Student's t-test. Welch's modification is used because it is more robust to departures from the assumptions of the t-test.

$\bar{X}$, $\bar{Y}$, S_X and S_Y are sample quantities defined in Appendix A.

Equation (1) can be rewritten as

$$LR = \frac{T_L P_0}{P_1 S_L} . lr_{cont} \quad (3)$$

as the term T_0 is small and generally unlikely to affect the final interpretation.

From Equation (2) it can be seen that lr_{cont} depends in two quantities: $\hat{g}(\bar{Y})$, and $f(\bar{X} - \bar{Y} \mid S_X, S_Y)$. What are these quantities? The motivation for a continuous approach rests on two main factors: (1) some evidence is contributed by the difference between the mean RI of the recovered sample and the mean RI of the control sample, and (2) additional evidence is given by the relative rarity (or frequency) of the recovered RI mean. The equation $f(\bar{X} - \bar{Y} \mid S_X, S_Y)$ is the height of a rescaled Student's t-distribution. The standard (or scaled) Student's t-distribution has been scaled by the standard error of the difference between the means so that the difference is measured in unit-free standard errors not RI. To remove this scaling, normally one would multiply the value by the standard error. However, we wish to put the data on a 1/RI scale, the reason being that $\hat{g}(\bar{Y})$, which is a data-based estimate of the frequency (in fact, it is the height of the density estimate at the recovered mean), is on the same scale. This rescaling is necessary to ensure that the resulting likelihood ratio is unit-free.

In a case where the glass evidence is quantified by elemental composition rather than by refractive index, the only change in evaluating the likelihood ratio is the method for evaluating lr_{cont}. It should be noted that lr_{cont} is an approximation to the true ratio of the densities, and the basic derivation is given in Walsh et al. [4].

Hotelling's T^2 is a multivariate analogue of the t-test, which examines the standardised squared distance between two points in n-dimensional space. These two points, of course, are given by the estimated mean concentration of the discriminating elements in both samples. It seems logical that the multivariate form of lr_{cont} should replace $f(\bar{X} - \bar{Y} \mid S_X, S_Y)$ with the rescaled probability density function for the distribution of T^2. It is necessary to ensure that the resulting likelihood ratio is unit-free. This, however, is not quite as simple as it sounds.

If there are n_c control fragments and n_r recovered fragments to be compared on the concentration of p different elements, and $n_c + n_r > p + 1$, then T^2 has an F-distribution scaled by the sample sizes [1], i.e.

$$T^2 \sim \frac{(n_c + n_r - 2)p}{(n_c + n_r - p - 1)} F_{p, n_c + n_r - p - 1} = k \times F_{p, n_c + n_r - p - 1}$$

where

$$k = \frac{(n_c + n_r - 2)p}{(n_c + n_r - p - 1)} .$$

If the approach suggested by Walsh et al. [4] is applied here, $f(\bar{X} - \bar{Y} \mid S_X, S_Y)$ should be replaced by the value of the rescaled (in order to gain a unit-free likelihood ratio) probability density for an F-distribution on p and $n_c + n_r - p - 1$ degrees of freedom at $\frac{T^2}{k}$.

In the same way $\hat{g}(\bar{Y})$ should be replaced by the value of a multivariate probability density estimate at the recovered mean $\bar{y}$ (recall that $\bar{y}$ is now a $p \times 1$ vector). However, the appropriate value used to rescale a multivariate distribution is the inverse of the covariance matrix. Multiplying the F-distribution by the inverse of the covariance matrix is analogous to dividing the t-distribution by the standard error. As noted however, this involves a matrix, while both $f\left(\frac{T^2}{k}\right)$ and $\hat{g}(\bar{Y})$ are scalars, so lr_{cont} could be evaluated, but the result would be a matrix and have no intuitive meaning. The solution to this problem comes from the way Hotelling's T^2 test works.

The test statistic for a Student's t-test is calculated by dividing the difference between the means by the value of the standard error of the difference. This is referred to as a standardised difference. Hotelling's T^2 does the same, except in this case the difference is squared. This must be done so that the F-distribution is useable. Hotelling's T^2 finds the linear combination of the variables that maximises the squared standardised distance between the two mean vectors (this vector defines the first principal component, an axis that explains the largest proportion of the variation present in the data).

More specifically, there is some vector l ($p \times 1$) of coefficients such that $l^T(\bar{x} - \bar{y})$ maximises the difference between the populations. This is more easily understood if the geometry of the problem is explained. Imagine that each measurement lies in a multi-dimensional space, with each axis representing the measurement on a specific element. Finding l can be thought of as rotating the axes so that the two means lie on an axis where the distance between them is the greatest.

If Hotelling's T^2 rejects the null hypothesis of no difference then $l^T(\bar{x} - \bar{y})$ will have a non-zero mean [11]. This fact provides the solution. The vector l that maximises the difference is used to derive a new statistic, t_l^2 (Appendix A), which has the same distribution as T^2 but has a scaling factor that is not a matrix [11]. The numerator of lr_{cont}, therefore becomes $f(l^T(\bar{X} - \bar{Y})\,|\,s_l)$. The numerator is the height of a probability density for an F-distribution on p and $n_c + n_r - p - 1$ degrees of freedom at $\frac{t_l^2}{k}$ transformed back to the scale of the linear combination. The denominator must be on the same scale, thus it becomes the value of a univariate probability estimate density at $l^T\bar{y}$.

Examples

The data in the following examples come from two distinct sources, one green bottle and one colourless bottle taken from the same plant at the same time. Ten fragments were taken from each bottle and the concentrations of aluminium, calcium, barium, iron, and magnesium ($p = 5$) were determined by ICP/AES. The quantities, T_L, P_0, P_l and S_L are taken to be those given in [4], so that

$$LR \approx 8.lr_{cont}$$

The first example uses five fragments from the green bottle as a control sample ($n_c = 5$) and five fragments from the same bottle as a recovered sample ($n_r = 5$) so that the population means are truly equal. Thus, $lr_{cont} \approx 2{,}600$, so in this case the evidence would be 20,800 times more likely if the suspect was at the crime scene than if he wasn't.

The second example takes the ten fragments from the green bottle as the control sample ($n_c = 10$) and the ten fragments from the colourless bottle as the recovered sample ($n_r = 10$), so the null hypothesis is false, i.e. the population means are truly different.

$$lr_{cont} \approx 4 \times 10^{-10}$$

Because lr_{cont} is so small, the term T_0 in equation (1) now determines the LR. If T_0 is taken at a typical value of around 0.1, then, in this case the evidence would be 10 times less likely if the suspect was at the crime scene than if he wasn't, i.e. the evidence against the suspect strongly disputes the hypothesis that the suspect was at the crime scene.

Conclusions

Hotelling's T^2 test for the difference in two mean vectors provides a valid statistical method for the discrimination between two samples of glass based on elemental data. However, it is subject to problems and does not answer the real question adequately. The Bayesian approach along with the continuous extension is the only method that fulfils the requirements of the forensic scientist, the statistician, and the court. All analyses of elemental data should use the continuous Bayesian approach.

The authors realise that while calculation of the lr_{cont} statistic is mathematically simple, it is computationally intensive. For that purpose a small software package that calculates lr_{cont}, T^2, and also returns the relevant probability from the F-distribution has been provided. Versions are available for MS-DOS, MS-Windows, and UNIX and may be requested via post or e-mail from the corresponding authors. In order to run this software a database of elemental compositions for glass fragments is needed.

Appendix A

If n_c control fragments are to be compared with n_r recovered fragments on p elements, then let

$$\begin{bmatrix} x_{11} & \cdots & x_{n_c1} \\ x_{12} & \ddots & x_{n_c2} \\ \vdots & \ddots & \vdots \\ x_{1p} & \cdots & x_{n_cp} \end{bmatrix} \text{and} \begin{bmatrix} y_{11} & \cdots & y_{n_r1} \\ y_{12} & \ddots & y_{n_r2} \\ \vdots & \ddots & \vdots \\ y_{1p} & \cdots & y_{n_rp} \end{bmatrix}$$

represent the measurements.

Let $x_j = [x_{j1}, x_{j2}, \ldots, x_{jp}]^T$ and $y_j = [y_{j1}, y_{j2}, \ldots, y_{jp}]^T$, then the summary statistics are defined by

$$\bar{x} = \frac{1}{n_c}\sum_{j=1}^{n_c} x_j \quad \text{and} \quad \bar{y} = \frac{1}{n_r}\sum_{j=1}^{n_r} y_j$$

$$S_X = \frac{1}{n_c - 1}\sum_{j=1}^{n_c}(x_j - \bar{x})(x_j - \bar{x})^T \text{ and } S_Y = \frac{1}{n_r - 1}\sum_{j=1}^{n_r}(y_j - \bar{y})(y_j - \bar{y})^T$$

respectively. An estimate of the common covariance, Σ, is given by

$$S_{pooled} = \frac{(n_c - 1)S_X + (n_r - 1)S_Y}{n_c + n_r - 2}$$

Hotelling's T^2 is then defined by

$$T^2 = (\bar{x} - \bar{y})^T \left[\left(\frac{1}{n_c} + \frac{1}{n_r}\right) S_{pooled}\right]^{-1} (\bar{x} - \bar{y})$$

It can be shown that the vector of coefficients, l, that defines the maximum population difference (this vector defines the first principal component, an axis that explains the largest proportion of the variation present in the data) is proportional to $[S_{pooled}]^{-1}(\bar{x} - \bar{y})$.

If $\hat{l} = [S_{pooled}]^{-1}(\bar{x} - \bar{y})$, then $t_{\hat{l}}^2$ is defined as

$$t_{\hat{l}}^2 = \frac{(\bar{x} - \bar{y})^2}{\left(\frac{1}{n_c} + \frac{1}{n_r}\right)\hat{l}^T S_{pooled}\, \hat{l}}$$

In fact $T^2 = \sup_l t_l^2$. The denominator of $t_{\hat{l}}^2$ is the scaling. Therefore

$$lr_{cont} = \frac{f\left(\frac{t_{\hat{l}}^2}{k}\right)}{\hat{g}\left((\hat{l}^T \bar{y})^2\right)} \times \frac{1}{D}$$

where $D = \left(\frac{1}{n_c} + \frac{1}{n_r}\right)\hat{l}^T S_{pooled}\, \hat{l}$

and $\hat{g}\left((\hat{l}^T \bar{y})^2\right)$ is a kernel estimate of the density at the point $(\hat{l}^T \bar{y})^2$.

$\hat{g}(y)$ is a traditional kernel density estimate. It works by smoothing the data by replacing each data point by a density function (typically a Gaussian), or kernel, centred at the data point. The resulting density estimate at any point, y, is the mean of the values of the individual densities for each datum at that point. The smoothness of the resulting density estimate, $\hat{g}(y)$, is controlled by a tuning parameter, or 'window', h.

$\hat{g}(y)$ is defined as

$$\hat{g}(y) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y - x_i}{h}\right)$$

$h = 1.06\, \hat{\sigma}_y n^{-1/5}$ [12] is the 'window' width, where $\hat{\sigma}_y$ is standard deviation of the data. The smooth kernel function,

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

is the probability density function for a standard Gaussian variable.

Acknowledgement

This research was made possible by a PhD scholarship from ESR: Forensic.

References

1. Curran JM, Triggs CM, Almirall JR, Buckleton JS and Walsh KAJ. The interpretation of elemental composition measurements from forensic glass evidence: I. Science & Justice 1997; 37: 241–244.
2. Robertson BW and Vignaux GA. DNA evidence: wrong answers or wrong questions. Genetica 1995; 96: 145–152.
3. Lindley DV. A problem in forensic science. Biometrika 1977; 64: 207–213.
4. Walsh KAJ, Triggs CM and Buckleton JS. A practical example of glass interpretation. Journal of the Forensic Science Society 1996; 36: 213–218.
5. McQuillan J and Edgar KA. Survey of the distribution of glass on clothing. Journal of the Forensic Science Society 1992; 32: 333–348.
6. Lambert JA, Sattterthwaite MJ and Harrison PH. A survey of glass fragments recovered from clothing of persons suspected of involvement in crime. Science & Justice 1995; 35: 273–281.
7. Pearson EF, May RW and Dabbs MDG. Glass and paint fragments found in men's outer clothing. Journal of Forensic Sciences 1971; 16: 283–300.
8. Lindley DV. A statistical paradox. Biometrika 1957; 44: 187–192.
9. Evett IW and Buckleton JS. The interpretation of glass evidence: a practical approach. Journal of the Forensic Science Society 1990; 30: 215–223.
10. Seber GAF. Multivariate Observations. New York: John Wiley & Sons, 1984.
11. Johnson RA and Wichern DA. Applied multivariate analysis. New Jersey: Prentice-Hall, Inc., 1982.
12. Scott DW. Multivariate density estimation. New York: John Wiley & Sons, 1992.