A Multivariate Extension to the Standard 4σ Criterion for Comparison of Forensic Glass Evidence

Oliver Lountain

October 11, 2021

Abstract

Contents

1 Introduction

1.1 Forensic Glass Evidence

When a person breaks a glass object – such as a window – minuscule fragments of the broken glass are transferred to the person's clothing. This process of fragmentation of the broken glass back

towards the object which broke it was observed by Nelson and Revell (1967) and is now referred to as backscatter fragmentation. Any person within "a few feet of the window" which is broken will likely have a number of minute fragments of glass transferred to their clothing. When a suspect is apprehended, these fragments can be found on their clothing and compared to the glass which was broken, potentially placing the suspect close to the object at the crime scene when it was broken. This analysis requires a careful treatment as one of course wishes to avoid wrongfully convicting an innocent person. As such, the development of measurement techniques with a high level of precision to discriminate between, say, two panes of glass manufactured at the same factory, on the same day, but installed as windows in two distinct buildings is of the utmost importance.

1.2   LA-ICPMS Data

Multiple methods have been used over time to measure glass evidence, including the colour, thickness and refractive index of the glass (Curran *et al.*, 2000), though more recently, the elemental composition has been the preferred approach. Elemental composition offers a multivariate measurement of the data, and as such can allow for a greater level of discrimination between samples. In Australia, LA-ICPMS (Houk, 1990, Curran *et al.*, 1997, Aeschliman *et al.*, 2003, Weis *et al.*, 2011, Trejos *et al.*, 2013, Dorn *et al.*, 2015) is conducted in some forensic laboratories and so data collected in this way will be the focus of this paper, though the techniques discussed can be applied to any elemental composition measurements. The motivation behind our research stems from the question of whether the currently employed interval-based criterion can be improved upon to take full advantage of the multivariate elemental data.

1.3   Data sets

We consider two data sets: a homogeneous set of observations taken from two glass factories in the USA and a much more diverse set of observations originating from South Australian casework from 2017 to 2020.

USA Ribbon Data

This first, homogeneous, data set was commissioned to serve as a database of chemical composition measurements by Park and Carriquiry (2019). The data is comprised only of measurements of float glass manufactured by two companies in the United States of America. These companies will be referred to as Company A and Company B. A sample of 31 panes of float glass was taken from Company A, labelled $AA, AB, \ldots, AAR$, and 17 manufactured by Company B labelled $BA, BB, \ldots, BR$. The panes sampled from Company A were manufactured between the third and 24th of January, 2017, and those produced by Company B were sampled a little earlier from the fifth to the 16th of December, 2016 (Park *et al.*, 2020).

Glass is manufactured in long continuous sheets known as ribbons, which are then cut into panes. At both of these factories, a large number of samples were taken from each ribbon of glass in order to develop an understanding of the level of variability within a source. To achieve this, on almost all days within the sampling periods, for both manufacturers, two glass panes were collected – one from each side of the ribbon. A sample of 24 fragments were then taken from each pane, and from 21 of these fragments, five replicate measurement were taken. For the other three, 20 replicate measurements were made, resulting in a total of 165 measurements per pane of glass.

In this study, the choice of elements to measure was made following Weis *et al.* (2011), who recommended that only 18 elements are used. Three major elements: calcium, sodium and magnesium; three minor elements: aluminium, potassium and iron; and 12 trace elements: lithium,

titanium, manganese, rubidium, strontium, zirconium, barium, lanthanum, cerium, neodymium, hafnium, and lead.

Australian Casework Data

The second data set comprises glass collected during casework in South Australia which has been measured and analysed by FSSA. The measurements in this data set span from 2016 to 2020 at the time of writing, and as the data is from real police casework, there is some variability in the number of measurements obtained. The data also come in two distinct forms: control samples and recovered samples. Control samples are taken from a known source at the scene of the crime, and typically at least nine individual fragments are measured, and between one and three replicate measurements are made of each fragment, depending on what is possible given the size of the fragment. Recovered samples are typically those taken from suspects' clothing as well as other sources such as the interior of a car. These samples are much less numerous, with between one and three fragments measured, and again between one and three replicate measurements for each fragment. Given that these observations are far less numerous and do not have a known source, only control fragments are considered for the analysis throughout in order to maintain a certain level of consistency in the observations used.

2 Match Criteria

The standard practice by forensic practitioners for comparing glass samples with LA-ICPMS elemental analysis is via some interval-based match criteria (Park and Carriquiry, 2019). Specifically, two international standards are used: ASTM-E2330-12 (2012) and ASTM-E2927-16 (2016). In this setting, we consider glass from a known source at a crime scene, referred to as the control source, and a glass sample obtained from a suspect's clothing, known as a recovered source. To simplify the notation, we will denote the known and questioned samples by $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ respectively. For each sample, the concentrations of p elements are measured, for n_l fragments of a given sample of broken glass (where $l = 1, 2$). In some cases there will also be replicate measurements taken of each fragment. In order to avoid over-complicated notation, we assume that replicate measurements are contained in the index spanning the number of fragments. That is, we take n_l to span the fragments and replicates for sample l. In full, the control and recovered measurements are denoted

$$\boldsymbol{y} = \{y_{ljk} \,|\, l = 1, 2, j = 1, \ldots, n_l, k = 1, \ldots, p\}.$$

From this, we have that each measurement is a vector of the form

$$\boldsymbol{y}_{lj} = (y_{lj1}, \ldots, y_{ljp}).$$

The mean vector for each element in sample l is then denoted

$$\bar{\boldsymbol{y}}_l = \frac{1}{n_l} \sum_{j=1}^{n_l} \boldsymbol{y}_{lj},$$

and similarly the vector of standard deviations is

$$\boldsymbol{\sigma}_l = \frac{1}{n_l - 1} \sum_{j=1}^{n_l} (\bar{\boldsymbol{y}} - \boldsymbol{y}_{lj})^2.$$

That is,

$$\bar{\boldsymbol{y}}_l = (\bar{y}_{l1}, \ldots, \bar{y}_{lp}),$$

and

$$\boldsymbol{\sigma}_l = (\sigma_{l1}, \ldots, \sigma_{lp}),$$

2.1 Standard 4σ Interval Criterion

Starting with the standard criterion (Koons and Buscaglia, 2002, ASTM-E2330-12, 2012, ASTM-E2927-16, 2016), we consider two glass samples $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, where $\boldsymbol{y}_1$ is from a known source and $\boldsymbol{y}_2$ is the sample in question. The guidelines suggest that at least nine measurements be taken from the known source via three replicates of three fragments and that "as many measurements as are practical" be taken of the sample in question (ASTM-E2927-16, 2016). Comparison intervals are then constructed for each element individually. The k-th comparison interval is computed as the mean of the control sample plus or minus four times its standard deviation. However, a lower bound is placed on the standard deviation of 3% of the mean. As such, the k-th comparison interval is defined as

$$\bar{y}_{1k} \pm 4 \times \max\left\{\sigma_{lk}, 0.03 \times \bar{y}_{1k}\right\}. \tag{1}$$

In other words, no matter the number of measurements obtained for the known fragment, the variability in this data can never be less than 3% of the control mean. The concentrations of each of the p elements in the questioned sample are then compared to the intervals calculated in Equation 1. If the concentrations lie within the interval for all elements, then $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are said to be chemically indistinguishable. In the event that one or more concentration lies outside its respective interval for any of the measurements, the samples are said to be chemically distinguishable. Equivalently, the distinction between two samples is quantified with a score determined by rearranging Equation (1). That is, we take the absolute difference between the concentrations of a the k-th element for the two samples and scale by the standard deviation of $\boldsymbol{y}_1$. We denote this by $S_{ASTM,k}$. The overall comparison score between the samples, denoted S_{ASTM}, is then taken to be the maximum across the p elements.

$$S_{ASTM,k} = \left| \frac{\bar{y}_{1k} - \bar{y}_{2k}}{\max\left\{\sigma_{1k}, 0.03 \times \bar{y}_{1k}\right\}} \right|. \tag{2}$$

$$S_{ASTM} = \max_{1 \leq k \leq p} S_{ASTM,k}.$$

This score exceeding four is equivalent to the questioned sample lying outside of the comparison interval and so the samples are declared distinguishable in this case (ASTM-E2330-12, 2012, ASTM-E2927-16, 2016).

2.2 Ellipsoid 4σ Criterion

The interval-based criteria are most intuitively thought of as p individual univariate tests. However, geometrically, this is equivalent to constructing a hyper-rectangle around the observations whose dimension is the number of elements measured. This hyper-rectangle would be located at the centroid of the data, and the length in each dimension would be 8σ. Then, a recovered sample is declared as matching if it lies within this hyper-rectangle, and non-matching if it lies outside. The construction of this hyper-rectangle assumes no relationship between any of the elements. Therefore, the most natural extension of the standard interval criterion to take into account correlations between the elements would be to consider a p-dimensional ellipsoid centred about the mean vector. We would then wish for this ellipsoid to be rotated such that its principal axes align with the directions of greatest variation, and that the surface is always four standard deviations from the centre. As far as the author is aware, there is no mention of this in the forensic science literature, though it can be constructed quite easily by employing the Mahalanobis distance. The Mahalanobis distance is an example of what is known as a statistical distance, and provides a well-defined notion of distance between observations. To define it, let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two random vectors from the same distribution with covariance matrix Σ. Then the Mahalanobis distance $d_M(\boldsymbol{X}, \boldsymbol{Y})$ between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is given by

$$d_M(\boldsymbol{X}, \boldsymbol{Y}) = \sqrt{(\boldsymbol{X} - \boldsymbol{Y})^T \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{Y})}.$$

The Mahalanobis distance is a generalisation of Euclidean distance, scaled by the standard deviation in the direction between the two points. As a result, the numerical value of the Mahalanobis distance is precisely the number of standard deviations between the two points. It is worth noting also that in one dimension, the Mahalanobis distance simplifies precisely to the standard 4σ score as expressed in Equation (2). Further, by considering a centroid $\boldsymbol{\mu}$ and an observation $\mathbf{x}$, both in $\mathbb{R}^p$, a p-dimensional ellipsoid centred at $\boldsymbol{\mu}$, whose distance from the edge to the centre is equal to r standard deviations is given by the equation

$$(d_M(\boldsymbol{\mu}, \mathbf{x}))^2 = (\boldsymbol{\mu} - \mathbf{x})^T \Sigma^{-1} (\boldsymbol{\mu} - \mathbf{x}) = r^2.$$

This is referred to as the standard deviational ellipsoid and any point which lies within this hyper-ellipsoid is within r standard deviations of the mean vector, $\boldsymbol{\mu}$. Thus, by classifying matches for comparisons where the Mahalanobis distance is less than four, and non-matches where it is greater than four, this provides a natural generalisation of the standard 4σ interval method. For control and recovered samples $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, the comparison score in this case is therefore given by

$$S_{\text{ellipsoid}} = d_M(\boldsymbol{y}_1, \boldsymbol{y}_2).$$

To help visualise the distinction between the standard and ellipsoid 4σ criteria consider Figure 1. The data used in this figure has been simulated to clearly demonstrate the potential distinction. To keep the visualisation simple, only two variables have been used. Variables 1 and 2 have been simulated from a bivariate normal distribution, with covariance matrix given by

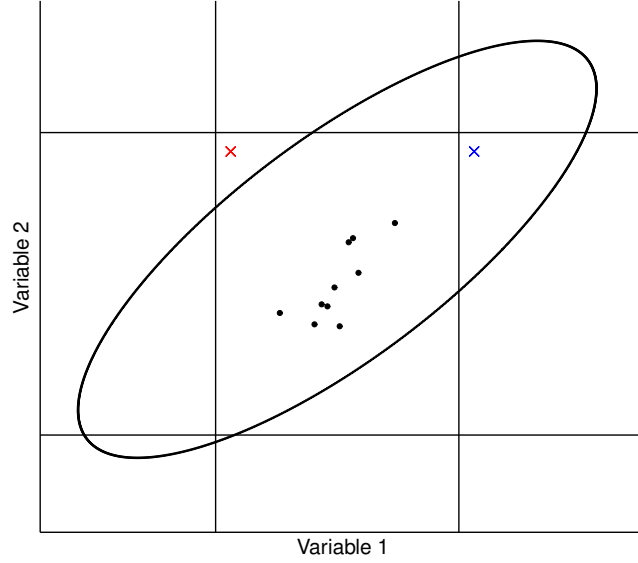$$\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}.$$

The black points show the simulated data, the straight black lines show the boundaries for the standard criterion in each variable, and the black ellipse is that for a Mahalanobis distance of four. The red cross shows a new observation which would be declared as matching by the standard criterion, but rejected by the ellipsoid method, while the opposite is true for the blue cross.

3 Results

3.1 USA Ribbon Data

We begin by applying the two methods to the USA ribbon data. We note in Table 1 that the standard 4σ criterion was only approximately 60% accurate, and that its low value of kappa suggests little difference between it and a uniform random classifier. It received a perfect score for sensitivity, meaning that it never incorrectly classified matching pairs as non-matching, but received quite a low score for specificity, meaning a large number of false positive predictions. The ellipsoid 4σ criterion appears to have significantly improved upon the standard method, achieving a raw accuracy of 0.950. It also scored above 0.9 for both sensitivity and specificity, with specificity in particular scoring above 0.95, meaning that false positives were minimised.

To provide another interpretation of sensitivity and specificity, we see that the standard 4σ criterion predicted perfectly on same source pairs, but only 60% of the time on different source pairs. The ellipsoid criterion, by contrast, predicted correctly 91.7% of the time on same source pairs, and 95% of the time on different source pairs. In other words, the ellipsoid criterion sacrifices approximately 8% accuracy on same source pairs, in order to increase different source accuracy by 35 percentage points.

Figure 1: Comparison of standard and ellipsoid 4σ match criteria in two variables. Variables 1 and 2 have been simulated from a bivariate normal distribution, each with variance 1, and the covariance is 0.5. The black points show the simulated data, the straight black lines show the boundaries for the standard criterion in each variable, and the black ellipse is that for a Mahalanobis distance of four. The red cross shows a new observation which would be declared as matching by the standard criterion, but rejected by the ellipsoid method, while the opposite is true for the blue cross.

3.2 Australian Casework Data

Moving on to the Australian casework data, Table 2 displays the performance metrics of these two criteria on this data set. We see that in terms of raw accuracy, both methods perform quite well, with the ellipsoid criterion performing slightly better with a near-perfect score of 0.992, an increase of 0.031 over the standard criterion. In Cohen's kappa coefficient, the difference is more pronounced, with a significant increase of 0.34 from the standard to ellipsoid criterion, suggesting that accounting for correlations has a significant impact on classification as compared to a random classifier. Looking at sensitivity and specificity, we note that the standard criterion achieves a perfect score for sensitivity, suggesting that it predicted no false negatives, that is, no pairs of fragments were classified as non-matching, when they were in fact matching. The standard criterion traded this high sensitivity for a slightly lower specificity, suggesting that it made a small number of false positive predictions. The ellipsoid criterion, achieved a balance with close to perfect scores for both sensitivity and specificity.

The score of 1.000 for sensitivity means that the standard 4σ criterion performed perfectly on same source pairs, and the specificity shows that it correctly predicted 96% of the time on different source pairs. By contrast, the ellipsoid 4σ criterion made correct predictions 99% of the time on same source pairs, and 99.2% of the time on different source pairs.

4 Discussion

Overall, we note that both the standard and ellipsoid criteria perform quite well on the diverse Australian casework data set, with the ellipsoid method offering substantial improvement on different source classifications. The standard criterion performed quite poorly on the homogeneous USA ribbon data, being able to classify same sources pairs as matching, but struggling to correctly

method	accuracy	kappa	sensitivity	specificity
Standard 4σ	0.616	0.109	1.000	0.600
Ellipsoid 4σ	0.950	0.575	0.917	0.951

Table 1: Performance metrics for standard and ellipsoid 4σ criterion applied to USA ribbon data. The standard criterion is only approximately 60% accurate while the ellipsoid criterion is 95% accurate. The standard criterion scored 0.109 for Cohen's Kappa, suggesting only a slight improvement over a uniform classifier, while the ellipsoid criterion received a score of 0.575 The standard criterion received a perfect score for sensitivity, meaning that it never incorrectly classified matching pairs as non-matching, but received quite a low score for specificity, meaning a large number of false positive predictions. The ellipsoid criterion has high scores for both sensitivity and specificity, with both above 90%.

method	accuracy	kappa	sensitivity	specificity
Standard 4σ	0.961	0.497	1.000	0.960
Ellipsoid 4σ	0.992	0.837	0.990	0.992

Table 2: Performance metrics for standard and ellipsoid 4σ criterion applied to Australian data. In terms of raw accuracy, both methods perform quite well, with the ellipsoid criterion performing slightly better with a near-perfect score of 0.992, an increase of 0.031 over the standard criterion. In Cohen's kappa coefficient, the difference is more pronounced, with a significant increase of 0.34 from the standard to ellipsoid criterion, suggesting that accounting for correlations has a significant impact on classification as compared to a random classifier. We note also that the standard criterion achieves a perfect score for sensitivity, suggesting that it predicted no false negatives.

classify different source pairs. The ellipsoid method, taking into account the correlation structure between the variables, closed this gap and offered substantial improvement on different source classification, while maintaining good performance on same source comparisons.

References

David B. Aeschliman, Stanley J. Bajic, David P. Baldwin, and R. S. Houk. Spatially-resolved analysis of solids by laser ablation-inductively coupled plasma-mass spectrometry: trace elemental quantification without matrix-matched solid standards. *Journal of Analytical Atomic Spectrometry*, 18(8):872, 2003.

ASTM-E2330-12. Test method for determination of concentrations of elements in glass samples using inductively coupled plasma mass spectrometry (ICP-MS) for forensic comparisons, 2012.

ASTM-E2927-16. Test method for determination of trace elements in soda-lime glass samples using laser ablation inductively coupled plasma mass spectrometry for forensic comparisons, 2016.

J.M. Curran, C.M. Triggs, J.R. Almirall, J.S. Buckleton, and K.A.J. Walsh. The interpretation of elemental composition measurements from forensic glass evidence: II. *Science & Justice*, 37(4):245–249, oct 1997.

James Michael Curran, Tacha Natalie Hicks Champod, and John S Buckleton. *Forensic interpretation of glass evidence.* CRC Press, 2000.

Hendrik Dorn, David E. Ruddell, Alex Heydon, and Brenda D. Burton. Discrimination of float glass by LA-ICP-MS: assessment of exclusion criteria using casework samples. *Canadian Society of Forensic Science Journal*, 48(2):85–96, mar 2015.

R.S. Houk. Chapter 91 elemental analysis by atomic emission and mass spectrometry with inductively coupled plasmas. In *Handbook on the Physics and Chemistry of Rare Earths*, pages 385–421. Elsevier, 1990.

Robert Koons and JoAnn Buscaglia. Interpretation of glass composition measurements: The effects of match criteria on discrimination capability. *Journal of forensic sciences*, 47:505–12, 06 2002.

D.F. Nelson and B.C. Revell. Backward fragmentation from breaking glass. *Journal of the Forensic Science Society*, 7(2):58–61, apr 1967.

Soyoung Park and Alicia Carriquiry. Learning algorithms to evaluate forensic glass evidence. *The Annals of Applied Statistics*, 13(2):1068–1102, 2019.

Soyoung Park, Alicia Carriquiry, L. Kenneth Horkley, and David W. Peate. A database of elemental compositions of architectural float glass samples measured by LA-ICP-MS. *Data in Brief*, 30:105449, jun 2020.

Tatiana Trejos, Robert Koons, Stefan Becker, Ted Berman, JoAnn Buscaglia, Marc Duecking, Tiffany Eckert-Lumsdon, Troy Ernst, Christopher Hanlon, Alex Heydon, Kim Mooney, Randall Nelson, Kristine Olsson, Christopher Palenik, Edward Chip Pollock, David Rudell, Scott Ryland, Anamary Tarifa, Melissa Valadez, Peter Weis, and Jose Almirall. Cross-validation and evaluation of the performance of methods for the elemental analysis of forensic glass by μ-XRF, ICP-MS, and LA-ICP-MS. *Analytical and Bioanalytical Chemistry*, 405(16):5393–5409, may 2013.

Peter Weis, Marc Dücking, Peter Watzke, Sonja Menges, and Stefan Becker. Establishing a match criterion in forensic comparison analysis of float glass using laser ablation inductively coupled plasma mass spectrometry. *Journal of Analytical Atomic Spectrometry*, 26(6):1273, 2011.