

基于多元玻璃物证的非平衡似然比比对检验模型建立及评估

周渝竣¹, 郭洪玲²

¹广东工业大学数学与统计学院, 广东 广州

²公安部物证鉴定中心, 北京

收稿日期: 2025年4月14日; 录用日期: 2025年5月7日; 发布日期: 2025年5月14日

摘要

对在犯罪现场和嫌疑人身上发现的证据特征进行测量和评估是刑事调查中的重要部分。Aitken等在2006年提出了样本平衡的多元正态分布多层次证据似然比模型。基于玻璃数据, 本文考虑了一种非平衡类型的多元正态多层次证据: 包含7种元素的玻璃碎片数据58组, 其中每组包含的玻璃样本数量不同(样本非平衡性); 共有396个玻璃样本, 每个样本都被测量4次。为降低数据维度, 利用图形模型(graphical models)找出具有高度相关的变量。假定数据服从多元正态分布, 建立样本非平衡的多元正态分布多层次证据似然比模型。在58对同一起来源对比中, $LR > 1$ 的有20对, 假阴性率约为34.48%; 在1653对不同来源对比中, $LR > 1$ 的有78对, 假阳性率约为4.7%。对于司法机关来说, 评价的假阳性率更为重要, 因为假阳性结果是对应于对无辜者的证据。本文建立的基于图形模型的非平衡似然比建模为司法科学中解决非平衡数据提供了一种新的方案。

关键词

法医科学, 多元数据, 似然比, 图形模型, 两类错误率

Establishment and Evaluation of Non-Equilibrium Likelihood Ratio Comparison Test Model Based on Multi-Element Glass Physical Evidence

Yujun Zhou¹, Hongling Guo²

¹School of Mathematics and Statistics, Guangdong University of Technology, Guangzhou Guangdong

²Ministry of Public Security, Institute of Forensic Science, Beijing

Received: Apr. 14th, 2025; accepted: May 7th, 2025; published: May 14th, 2025

文章引用: 周渝竣, 郭洪玲. 基于多元玻璃物证的非平衡似然比比对检验模型建立及评估[J]. 应用数学进展, 2025, 14(5): 100-109. DOI: 10.12677/aam.2025.145238

Abstract

In criminal investigation, the measurement and evaluation of the characteristics of the evidence found at the crime scene and on the suspect is an important part. Aitken *et al.* proposed a multivariate normal distribution three-level evidence likelihood ratio model with sample balance in 2006. Based on the glass data, we consider a non-equilibrium type of multivariate normal three-level evidence: 58 sets of glass fragment data containing 7 elements, each of which contains a different number of glass samples (sample non-equilibrium). There were 396 glass samples, each of which was measured four times. To reduce the dimensionality of the data, graphical models are used to find variables with high correlation. In the 58 pairs from the same source, there were 20 pairs $LR > 1$, and the false negative rate was about 34.48%. In the comparison of 1653 pairs from different sources, there were 78 pairs $LR > 1$, and the false positive rate was about 4.7%. For the prosecution, the false positive rate of the evaluation is more important because false positive results correspond to evidence against an innocent person. The non-equilibrium likelihood ratio modeling based on graphical model constructed in this paper provides a new scheme for solving non-equilibrium data in forensic science.

Keywords

Forensic Science, Multivariate Data, Likelihood Ratio, Graphical Models, Two Types of Error Rates

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

基于贝叶斯推理框架的似然比法[1]-[3] (likelihood ratio, LR)是目前世界公认的法庭科学领域证据价值评价方法之一,它基于先验知识和证据信息对某一事件发生的可能性进行逻辑推理和概率估计,自提出至今已被广泛应用在多个物证鉴定专业领域。例如玻璃碎片分析[4]、枪弹痕迹[5]、墨水笔迹[6]、毒品检验[7]等。

与国外相比,目前国内似然比的研究和应用仍处于起步阶段。虽然已有对某类微量物证进行似然比模型建立和计算的研究[8],但只是针对单变量的数据。对于更为复杂的、包含多个变量的微量物证,目前还没有相关的研究。

本文以玻璃物证为分析对象,基于多种元素玻璃检验数据,提出了一种评估多变量证据的似然比方法,该方法是对 Aitken 等[9]提出的样本平衡多层次正态分布证据似然比模型的扩展。为了避免“维度诅咒”,利用图形模型(Graphical model)找出具有高度相关的元素,使得高维数据集变为相互独立的几个低维数据集的乘积,从而尽可能保留原始数据集中的有用信息。

利用两类错误率来评估样品比对检验的结果是法庭科学领域常用的方法[10]。两类错误率是指假阴性率(False negative rate)和假阳性率(False positive rate) [11]来评价 LR 模型的判断效果。假阴性率是指所有来自同一样本的样本 LR 结果中,出现 $LR < 1$ 的样本个数占比;假阳性率则是指所有来自不同来源的样本 LR 结果中,出现 $LR > 1$ 的样本个数占比。通过两类错误率可以直观地对 LR 模型的好坏作一个评价。

2. 测量方法

玻璃在日常用品中被广泛使用,使得玻璃成为犯罪现场经常遇到的重要法医痕迹证据。在玻璃的生产过程中,许多不同的元素被纳入熔融混合物中,这些主要成分是硅、钠、钙、镁和钾等的氧化物。采用

SEM-EDX 法[12][13]对玻璃样品的元素进行测定。先用树脂包埋法处理需要测量的玻璃, 然后利用专业的扫描电子显微镜(SEM-EDX 设备)分析玻璃样品的 SEM 图像和获得的玻璃样品的光谱, 最终测定了七种($p=7$)氧化物的含量, 分别是氧化钠(NaO)、氧化镁(MgO)、氧化铝(AlO)、氧化硅(SiO)、氧化钾(KO)、氧化钙(CaO)和氧化铁(FeO)。

2.1. 基础数据排版规范的完整性

考虑 $s=58$ 组玻璃样本, 每组样本包含不同的样本数量 $n_i (i=1, \dots, s)$ (样本非平衡), $M = \sum_{i=1}^s n_i = 396$, 每个样本独立重复测量 $m=4$ 次。这样分组-组内样本-测量次数就构成了三层次框架。为使数据更符合多元正态分布, 对数据进行对数变换($\log_{10}(L)$, L 为测定的元素含量)。以上所有已被测定的数据从《Statistical Analysis in Forensic Science》获得。

2.2. 样本数据和检材数据

假设在犯罪现场收集到的 G_1 个样本证据为样本数据, 这些数据的来源假定是已知的, 对它们分别进行 N_1 次独立重复测量; 假设在嫌疑人身上收集到的 G_2 个样本证据为检材数据, 这些数据的来源假定是未知的, 对它们分别进行了 N_2 次的独立重复测量。

3. 建立物证比对似然比模型

Aitken 等在 2006 年提出的多元正态分布三层次证据似然比模型, 考虑的是样本平衡的数据, 即每组都包含相同数量的样本。但在实际法医工作中, 更多地会出现样本非平衡的情况, 即收集到的每组数据, 包含不同数量的样本。如果依然采用样本平衡下的似然比模型, 需要对每组数据进行部分删减, 毫无疑问会丢失大量有用的信息, 以及对后续的似然比结果的准确性产生影响。考虑到这种情况, 本文将提出样本非平衡的多元正态分布三层次证据似然比模型。相较于 Aitken 的模型, 该模型对正态分布假设和参数估计部分进行了改进, 使其体现出样本非平衡这一特性, 从而更好地保留了数据中的信息。

3.1. 正态分布假设

现从基础数据中随机抽样, 记第 i 组的第 j 个样本第 k 次测量的证据向量为 $\mathbf{X}_{ijk} = (x_{ijk1}, \dots, x_{ijkp})^T$; $i=1, \dots, s; j=1, \dots, n_i; k=1, \dots, m$ 。假设三层次数据服从正态分布, 具体情形如下:

- 测量层次: 由 $\bar{\mathbf{X}}_{ij} = (1/m) \sum_{k=1}^m \mathbf{X}_{ijk}$ 来估计第 i 组的第 j 个样本的测量均值向量 $\boldsymbol{\theta}_{ij}$, 假设所有样本的重复测量的协方差矩阵为 $\mathbf{U}$, 则 $\mathbf{X}_{ijk}$ 的分布为 $(\mathbf{X}_{ijk} | \boldsymbol{\theta}_{ij}, \mathbf{U}) \sim N(\boldsymbol{\theta}_{ij} | \mathbf{U}); i=1, \dots, s; j=1, \dots, n_i; k=1, \dots, m$;

- 组内层次: 由 $\bar{\mathbf{X}}_{i\cdot} = (1/n_i) \sum_{j=1}^{n_i} \bar{\mathbf{X}}_{ij}$ 来估计第 i 组的组内均值向量 $\boldsymbol{\sigma}_i$, 假设组内协方差矩阵为 $\mathbf{V}$, 则测量均值向量 $\boldsymbol{\theta}_{ij}$ 的分布为 $(\boldsymbol{\theta}_{ij} | \boldsymbol{\sigma}_i, \mathbf{V}) \sim N(\boldsymbol{\sigma}_i | \mathbf{V}); i=1, \dots, s; j=1, \dots, n_i$;

- 组间层次: 由 $\bar{\mathbf{X}}_{\dots} = (1/M) \sum_{i=1}^s \bar{\mathbf{X}}_{i\cdot}$ 来估计总体均值向量 $\boldsymbol{\mu}$, 假设组间协方差矩阵为 $\mathbf{W}$, 则组内均值向量 $\boldsymbol{\sigma}_i$ 的分布为 $(\boldsymbol{\sigma}_i | \boldsymbol{\mu}, \mathbf{W}) \sim N(\boldsymbol{\mu} | \mathbf{W}); i=1, \dots, s$ 。

假设检材数据和样本数据分别记为 $\mathbf{y}_1, \mathbf{y}_2$ 。为方便以下表述, 引入指标 d : $d=1$ 代表检材数据, $d=2$ 代表样本数据。记指标 d 的第 j 个样本第 k 次测量的证据向量为 $\mathbf{y}_{djk} = (y_{djk1}, \dots, y_{djkp})^T$;

$j=1, \dots, G_d; k=1, \dots, N_d$ 。此时, 记指标 d 的第 j 个样本的重复测量的样本均值向量为 $\bar{\mathbf{y}}_{dj} = (1/N_d) \sum_{k=1}^{N_d} \mathbf{y}_{djk}$ 。

记指标 d 的重复测量的样本均值向量为 $\bar{\mathbf{y}}_d = (1/G_d) \sum_{j=1}^{G_d} \bar{\mathbf{y}}_{dj}$ 。现定义加权均值 $\bar{\mathbf{y}}$, 则 $\bar{\mathbf{y}} = (G_1 N_1 \bar{\mathbf{y}}_1 + G_2 N_2 \bar{\mathbf{y}}_2) / t$, 其中 $t = G_1 N_1 + G_2 N_2$, 加权均值 $\bar{\mathbf{y}}$ 的分布为: $(\bar{\mathbf{y}} | \boldsymbol{\mu}, \mathbf{U}, \mathbf{V}, \mathbf{W}) \sim N(\boldsymbol{\mu}, (G_d N_d)^{-1} \mathbf{U} + N_d^{-1} \mathbf{V} + \mathbf{W})$ 。

3.2. 参数估计

按照经典的极大似然理论[14], 总体均值向量 $\boldsymbol{\mu}$ 由 $\bar{\mathbf{X}}_{\dots}$ 来估计。

利用 $\{\mathbf{X}_{ijk}\}$ 估计测量层次的协方差矩阵 $\mathbf{U}$, 其表达式如下:

$$\hat{\mathbf{U}} = \frac{\mathbf{S}_{\mathbf{U}}}{M(m-1)} \quad (1)$$

其中, $\mathbf{S}_{\mathbf{U}} = \sum_{i=1}^s \sum_{j=1}^{n_i} \sum_{k=1}^m (\mathbf{X}_{ijk} - \bar{\mathbf{X}}_{ij.})(\mathbf{X}_{ijk} - \bar{\mathbf{X}}_{ij.})^T$ 。

利用 $\{\mathbf{X}_{ijk}\}$ 估计组内层次的协方差矩阵 $\mathbf{V}$, 其表达式如下:

$$\hat{\mathbf{V}} = \frac{\mathbf{S}_{\mathbf{V}}}{M-s} - \frac{\hat{\mathbf{U}}}{m} \quad (2)$$

其中, $\mathbf{S}_{\mathbf{V}} = \sum_{i=1}^s \sum_{j=1}^{n_i} (\bar{\mathbf{X}}_{ij.} - \bar{\mathbf{X}}_{i..})(\bar{\mathbf{X}}_{ij.} - \bar{\mathbf{X}}_{i..})^T$ 。

利用 $\{\mathbf{X}_{ijk}\}$ 估计组间层次的协方差矩阵 $\mathbf{W}$, 其表达式如下:

$$\hat{\mathbf{W}} = \frac{\mathbf{S}_{\mathbf{W}}}{s-1} - \frac{\hat{\mathbf{U}}}{m} - \frac{\hat{\mathbf{V}}}{m} \quad (3)$$

其中, $\mathbf{S}_{\mathbf{W}} = \sum_{i=1}^s n_i (\bar{\mathbf{X}}_{i..} - \bar{\mathbf{X}}_{\dots})(\bar{\mathbf{X}}_{i..} - \bar{\mathbf{X}}_{\dots})^T$ 。

3.3. 非平衡多元正态性假设的似然比模型

从犯罪现场采集的证据样本为 p_1 , 从嫌疑人身上采集的证据样本为 p_2 , E 为由 p_1 和 p_2 组成的证据样本总体。记控方假设(Prosecution hypothesis)和辩方假设(Defense hypothesis)分别为 H_p 、 H_d [15], 具体意义如下:

H_p : $P_1 = P_2$, 二者是相同来源, 认为嫌疑人和犯罪现场存在相关。

H_d : $P_1 \neq P_2$, 二者不是相同来源, 认为嫌疑人和犯罪现场不存在相关。

对这两种竞争假设条件下获得证据的概率比进行计算, 即似然比为 $\text{LR} = P(E | H_p) / P(E | H_d)$ 。

基于控方假设 H_p 和辩方假设 H_d , 考虑连续性三层次非平衡的正态分布模型样品, 参考文献[9], 可得 LR 值的计算公式如下:

$$\text{LR} = \frac{f(E | H_p)}{f(E | H_d)} = \frac{f(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\mu}, \mathbf{U}, \mathbf{V}, \mathbf{W})}{f(\mathbf{y}_1 | \boldsymbol{\mu}, \mathbf{U}, \mathbf{V}, \mathbf{W}) f(\mathbf{y}_2 | \boldsymbol{\mu}, \mathbf{U}, \mathbf{V}, \mathbf{W})} \quad (4)$$

其中, f 表示多元正态分布的概率密度函数。分子计算如下:

$$\begin{aligned} & f(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\mu}, \mathbf{U}, \mathbf{V}, \mathbf{W}) \\ &= |2\pi(\mathbf{V} + \mathbf{W})|^{-1/2} |2\pi\mathbf{U}|^{-T/2} \times \left| 2\pi \left\{ t\mathbf{U}^{-1} + (\mathbf{V} + \mathbf{W})^{-1} \right\}^{-1} \right|^{1/2} \times \exp \left\{ -\frac{1}{2} (H_1 + H_2) \right\} \\ & \times \exp \left\{ -\frac{1}{2} \sum_{j=1}^{G_1} \text{tr}(S_{1j} \mathbf{U}^{-1}) - \frac{1}{2} \sum_{j=1}^{G_2} \text{tr}(S_{2j} \mathbf{U}^{-1}) - \frac{1}{2} N_1 \text{tr}(S_1 \mathbf{U}^{-1}) - \frac{1}{2} N_2 \text{tr}(S_2 \mathbf{U}^{-1}) \right\} \end{aligned}$$

其中:

$$\begin{aligned}\mathbf{D}_1 &= N_1 G_1^{-1} \mathbf{U}; \mathbf{D}_2 = N_2 G_2^{-1} \mathbf{U}; H_1 = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)(\mathbf{D}_1 + \mathbf{D}_2)^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)^T \\ H_2 &= (\bar{\mathbf{y}} - \boldsymbol{\mu})(T^{-1} \mathbf{U} + (\mathbf{V} + \mathbf{W}))^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu})^T; S_d = \sum_{j=1}^{N_d} (\bar{\mathbf{y}}_{dj} - \bar{\mathbf{y}}_d)(\bar{\mathbf{y}}_{dj} - \bar{\mathbf{y}}_d)^T; \\ S_{dj} &= \sum_{k=1}^{N_d} (\mathbf{y}_{djk} - \bar{\mathbf{y}}_{dj})(\mathbf{y}_{djk} - \bar{\mathbf{y}}_{dj})^T; d = 1, 2\end{aligned}$$

分母计算如下:

$$\begin{aligned}f(\mathbf{y}_1 | \boldsymbol{\mu}, \mathbf{U}, \mathbf{V}, \mathbf{W}) &= |2\pi \mathbf{U}|^{-(N_1 G_1)/2} |2\pi(\mathbf{V} + \mathbf{W})|^{-1/2} \left[2\pi [\mathbf{D}_1^{-1} + (\mathbf{V} + \mathbf{W})^{-1}] \right]^{-1/2} \\ &\times \exp \left\{ -\frac{1}{2} \sum_{j=1}^{G_1} \text{tr}(S_{1j} \mathbf{U}^{-1}) - \frac{1}{2} N_1 \text{tr}(S_1 \mathbf{U}^{-1}) \right\} \exp \left\{ -\frac{1}{2} (\bar{\mathbf{y}}_1 - \boldsymbol{\mu})(\mathbf{D}_1 + \mathbf{V} + \mathbf{W})^{-1}(\bar{\mathbf{y}}_1 - \boldsymbol{\mu})^T \right\}\end{aligned}$$

同理, 可得 $f(\mathbf{y}_2 | \boldsymbol{\mu}, \mathbf{U}, \mathbf{V}, \mathbf{W})$ 的表达式:

$$\begin{aligned}f(\mathbf{y}_2 | \boldsymbol{\mu}, \mathbf{U}, \mathbf{V}, \mathbf{W}) &= |2\pi \mathbf{U}|^{-(N_2 G_2)/2} |2\pi(\mathbf{V} + \mathbf{W})|^{-1/2} \left[2\pi [\mathbf{D}_2^{-1} + (\mathbf{V} + \mathbf{W})^{-1}] \right]^{-1/2} \\ &\times \exp \left\{ -\frac{1}{2} \sum_{j=1}^{G_2} \text{tr}(S_{2j} \mathbf{U}^{-1}) - \frac{1}{2} N_2 \text{tr}(S_2 \mathbf{U}^{-1}) \right\} \exp \left\{ -\frac{1}{2} (\bar{\mathbf{y}}_2 - \boldsymbol{\mu})(\mathbf{D}_2 + \mathbf{V} + \mathbf{W})^{-1}(\bar{\mathbf{y}}_2 - \boldsymbol{\mu})^T \right\}\end{aligned}$$

将两者相乘, 得到(4)式的分母。

将以上得到的结果相应地代入(4)式的分子分母, 经过简化运算, 可以得到 LR 值的最简表达式如下:

$$\text{LR} = \frac{|\mathbf{V} + \mathbf{W}|^{1/2} |\mathbf{U}^{-1} + (\mathbf{V} + \mathbf{W})^{-1}|^{-1/2} \exp \left\{ -\frac{1}{2} (H_1 + H_2) \right\}}{|\mathbf{D}_1^{-1} + (\mathbf{V} + \mathbf{W})^{-1}|^{-1/2} |\mathbf{D}_2^{-1} + (\mathbf{V} + \mathbf{W})^{-1}|^{-1/2} \exp \left\{ -\frac{1}{2} (H_3 + H_4) \right\}} \quad (5)$$

4. 图形模型

图形模型是一种基于偏相关矩阵和图论的降维技术。该模型具有很好的变量可解释性, 同时减少了需要估计的参数数量。在图中, 每个变量代表一个节点, 具有高度相关的变量由一条线(边)连接。几个具有相关性的变量联结在一起称为团。而不同变量间的相关性由偏相关矩阵来确定。在本文中, 偏相关矩阵由组间协方差矩阵和组内协方差矩阵之和 $(\mathbf{V} + \mathbf{W})$ 确定。通过先对 $(\mathbf{V} + \mathbf{W})$ 求逆, 再进行缩放, 可得到偏相关矩阵[16]。一般地, 偏相关矩阵是正定的; 如果仅是半正定, 可通过重新参数化变成正定[17]。

在实际数据分析时, 需要设定一个相关系数的阈值 R_0 , $0 < R_0 < 1$ 。对于 $\boldsymbol{\Omega}$ 中的相关系数 $R_{ab}(a, b = 1, \dots, p, a \neq b)$, 若 $|R_{ab}| \geq R_0$, 则保留该相关系数; 否则令 $R_{ab} = 0$ 。 R_0 取值不同, 得到的偏相关矩阵(记为 $\boldsymbol{\Omega}_{R_0}$)也不同。根据不同的偏相关矩阵会产生不同的图形模型。

利用图形模型, 在计算似然比时, 可以把高维的数据集, 变为几个低维的数据集的乘积。计算公式如下:

$$\text{LR}(C_i | S_i) = \frac{\prod \text{LR}(C_i)}{\prod \text{LR}(S_i)} \quad (6)$$

其中, C_i 为图形模型中的第 i 个团, 即所有节点相互连接的变量子集。第 i 个团的所有分隔符的集合用 S_i 表示, 即两个团相交处的一个节点或一组节点。

5. 实验与分析

5.1. 实验流程

在第 2.1 节, 给出了多种化合物玻璃数据的具体信息; 假设其中被测量的 7 个变量在经过对数变换后分布为 7 维的多元正态分布, 7 维向量的分量为 $\log_{10}(\text{NaO})$ 、 $\log_{10}(\text{MgO})$ 、 $\log_{10}(\text{AlO})$ 、 $\log_{10}(\text{SiO})$ 、 $\log_{10}(\text{KO})$ 、 $\log_{10}(\text{CaO})$ 、 $\log_{10}(\text{FeO})$ 。

5.2. 参数结果

基于基础数据, 利用第 3.1 节的公式依次计算, 得到参数 μ 的估计:

$$\hat{\mu} = (-0.7095, -2.0796, -2.2549, -0.1408, -2.8842, -1.2408, -4.6718)^T$$

参数 U 的估计:

$$\hat{U} = \begin{pmatrix} 2.029 \times 10^{-4} & 3.051 \times 10^{-4} & -2.177 \times 10^{-4} & 1.813 \times 10^{-5} & -1.923 \times 10^{-4} & \dots 2.799 \times 10^{-4} & 2.980 \times 10^{-4} \\ 3.051 \times 10^{-4} & 1.711 \times 10^{-1} & -1.387 \times 10^{-3} & 3.387 \times 10^{-4} & -1.633 \times 10^{-3} & -3.671 \times 10^{-3} & -3.897 \times 10^{-3} \\ -2.177 \times 10^{-4} & 1.387 \times 10^{-3} & -1.528 \times 10^{-1} & 2.984 \times 10^{-4} & 1.810 \times 10^{-3} & -6.095 \times 10^{-6} & 1.848 \times 10^{-3} \\ 1.813 \times 10^{-5} & 3.387 \times 10^{-4} & -2.984 \times 10^{-4} & 8.009 \times 10^{-4} & 2.275 \times 10^{-3} & 1.123 \times 10^{-3} & 6.104 \times 10^{-4} \\ -1.923 \times 10^{-4} & -1.633 \times 10^{-3} & -1.810 \times 10^{-3} & 2.275 \times 10^{-3} & 2.531 \times 10^{-1} & 1.203 \times 10^{-3} & -2.318 \times 10^{-2} \\ 2.799 \times 10^{-4} & -3.671 \times 10^{-3} & -6.059 \times 10^{-6} & 1.123 \times 10^{-3} & 1.203 \times 10^{-3} & -3.352 \times 10^{-2} & -4.165 \times 10^{-4} \\ 2.980 \times 10^{-4} & -3.897 \times 10^{-3} & 1.848 \times 10^{-6} & 6.104 \times 10^{-4} & -2.318 \times 10^{-2} & 4.165 \times 10^{-3} & 3.104 \times 10^{-4} \end{pmatrix}$$

参数 V 的估计:

$$\hat{V} = \begin{pmatrix} 3.140 \times 10^{-3} & 1.966 \times 10^{-2} & -8.823 \times 10^{-4} & 7.461 \times 10^{-6} & -1.109 \times 10^{-2} & 1.989 \times 10^{-2} & 2.980 \times 10^{-3} \\ 1.966 \times 10^{-2} & 6.562 \times 10^{-1} & -1.030 \times 10^{-1} & -7.660 \times 10^{-4} & -8.508 \times 10^{-2} & 2.584 \times 10^{-1} & -3.897 \times 10^{-2} \\ 8.823 \times 10^{-4} & -1.030 \times 10^{-1} & 3.317 \times 10^{-1} & -1.609 \times 10^{-3} & 1.016 \times 10^{-1} & -1.965 \times 10^{-2} & 1.848 \times 10^{-1} \\ 7.461 \times 10^{-6} & -7.660 \times 10^{-4} & -1.609 \times 10^{-3} & 1.323 \times 10^{-3} & 6.330 \times 10^{-3} & 1.656 \times 10^{-3} & 6.104 \times 10^{-3} \\ -1.109 \times 10^{-2} & -8.508 \times 10^{-2} & 1.016 \times 10^{-1} & 6.330 \times 10^{-3} & 8.868 \times 10^{-1} & -7.847 \times 10^{-2} & -2.318 \times 10^{-2} \\ 1.989 \times 10^{-2} & 2.584 \times 10^{-1} & -1.965 \times 10^{-2} & 1.656 \times 10^{-3} & -7.847 \times 10^{-2} & 3.148 \times 10^{-1} & 4.165 \times 10^{-2} \\ 5.381 \times 10^{-3} & 3.062 \times 10^{-2} & -1.221 \times 10^{-1} & 3.363 \times 10^{-3} & 7.784 \times 10^{-2} & 2.919 \times 10^{-2} & 1.1110 \end{pmatrix}$$

参数 W 的估计:

$$\hat{W} = \begin{pmatrix} 1.172 & 3.548 & 3.234 & 0.229 & 3.971 & 2.559 & 7.767 \\ 3.548 & 20.26 & 7.913 & 0.665 & 5.661 & 11.06 & 26.29 \\ 3.234 & 7.913 & 13.89 & 0.640 & 16.77 & 5.132 & 19.42 \\ 0.229 & 0.665 & 0.640 & 0.049 & 0.812 & 0.444 & 1.516 \\ 3.971 & 5.661 & 16.77 & 0.812 & 27.46 & 4.056 & 24.69 \\ 2.559 & 11.06 & 5.132 & 0.444 & 4.056 & 11.37 & 17.03 \\ 7.767 & 26.29 & 19.42 & 1.516 & 24.69 & 17.03 & 58.4 \end{pmatrix}$$

5.3. 建立图形模型

根据图形模型方法, 先求协方差矩阵估计 $(V+W)$ 的逆, 再缩放得到逆缩放矩阵 Ω , 如表 1 所示。

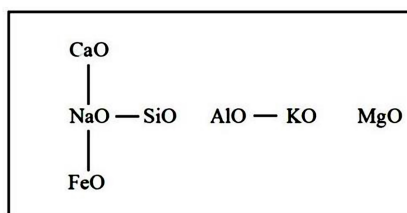
Table 1. Scaled inverse matrix Ω **表 1.** 缩放逆矩阵 Ω

	logNaO	logMgO	logAlO	logSiO	logKO	logCaO	logFeO
logNaO	1.000	0.0019	-0.4119	-0.6394	-0.0413	-0.5071	-0.5820
logMgO		1.000	-0.2569	-0.0347	0.4348	-0.3062	-0.3720
logAlO			1.000	0.0956	-0.6360	0.1500	0.3807
logSiO				1.000	-0.0740	0.2653	0.0172
logKO					1.000	0.1529	-0.2288
logCaO						1.000	0.1882
logFeO							1.000

取阈值 $R_0 = 0.5$, 对于 Ω 来说, $|R| < 0.5$ 的值全部设定为 0, 可以得到第一个偏相关矩阵 $\Omega_{0.5}$:

$$\Omega_{0.5} = \begin{pmatrix} 1.0000 & * & * & -0.6394 & * & -0.5071 & -0.5820 \\ & 1.0000 & * & * & * & * & * \\ & & 1.0000 & * & -0.6360 & * & * \\ & & & 1.0000 & * & * & * \\ & & & & 1.0000 & * & * \\ & & & & & 1.0000 & * \\ & & & & & & 1.0000 \end{pmatrix}$$

将偏相关矩阵 $\Omega_{0.5}$ 中的相关系数绝对值从大到小排序, 可以确定以下的图形模型(见图 1)。

**Figure 1.** Graphical model of partial correlation matrix $\Omega_{0.5}$ **图 1.** 偏相关矩阵 $\Omega_{0.5}$ 的图形模型

由图 1 得到, CaO、NaO、SiO、FeO 具有相关性, AlO、KO 具有相关性, 化合物 MgO 与其他元素没有相关性。进一步, 由图 1 可以确定基于偏相关矩阵 $\Omega_{0.5}$ 的 LR 值计算公式:

$$LR = \frac{LR(Ca, Si)LR(Si, Na)LR(Fe, Na)LR(Al, K)LR(Mg)}{LR(Na)LR(Na)} \quad (7)$$

若取阈值 $R_0 = 0.4$, 对于 Ω 来说, $|R| < 0.4$ 的值全部设定为 0, 可以得到第二个偏相关矩阵 $\Omega_{0.4}$:

$$\Omega_{0.4} = \begin{pmatrix} 1.0000 & * & -0.4119 & -0.6394 & * & -0.5071 & -0.5820 \\ & 1.0000 & * & * & 0.4348 & * & * \\ & & 1.0000 & * & -0.6360 & * & * \\ & & & 1.0000 & * & * & * \\ & & & & 1.0000 & * & * \\ & & & & & 1.0000 & * \\ & & & & & & 1.0000 \end{pmatrix}$$

将偏相关矩阵 $\Omega_{0.4}$ 中的相关系数绝对值从大到小排序, 可以确定以下的图形模型(见图 2)。

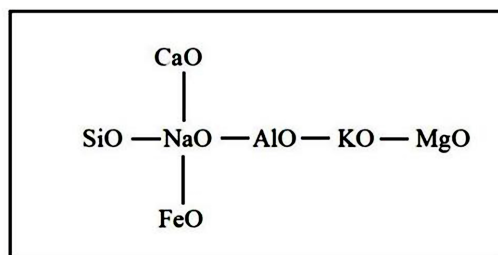


Figure 2. Graphical model of partial correlation matrix $\Omega_{0.4}$

图 2. 偏相关矩阵 $\Omega_{0.4}$ 的图形模型

由图 2 得到, 7 种元素依次具有相关性。再由图 2, 可以确定基于偏相关矩阵 $\Omega_{0.4}$ 的 LR 值计算公式:

$$LR = \frac{LR(Ca, Na)LR(Si, Na)LR(Fe, Na)LR(Na, Al)LR(Al, K)LR(K, Mg)}{LR(Na)LR(Na)LR(Na)LR(Al)LR(K)} \quad (8)$$

5.4. 同来源的玻璃数据比对

为进行比对实验, 现在从基础数据中抽取样本数据和检材数据, 将剩余的数据用于计算参数。具体操作如下, 从 $s = 58$ 组中随机抽取一组, 选取该组的任意一个样本作为检材数据 ($N_2 = 1$), 该组余下的样本作为样本数据。本文做实验时, 仅设定选取第一个样本作为检材数据; 剔除已被选定的一组, 将余下的 57 组数据用作参数估计。先利用第 5.3 节给出的公式(7)或公式(8)确定 LR 值计算公式, 再进行比对检验。最终每个图形模型都输出了 58 个比对检验结果, 根据不同数量级, 输出 LR 值计数统计结果见表 2。

Table 2. Comparison of two graphical models when recovered data and control data from the same source

表 2. 样本数据和检材数据同源时两种图形模型比较

似然比	$ R < 0.5$	$ R < 0.4$
<1	20	20
1~10	1	1
10~10 ²	1	1
10 ² ~10 ³	6	6
10 ³ ~10 ⁴	15	15
10 ⁴ ~10 ⁵	15	15
>10 ⁵	0	0
所有样本	58	58
真阳性(%)	65.52	65.52
假阴性(%)	34.48	34.48

5.5. 不同来源的玻璃数据比对

具体操作如下, 从 58 组数据中随机抽取两组, 从抽取的第一组中, 随机地抽取一个样本作为检材数据 ($N_2 = 1$), 将第二组全体作为样本数据。剔除已被抽取的两组, 将余下的 56 组数据用作参数估计。随机抽取两组的组合数为 $C_{58}^2 = 1653$, 最终输出了 1653 个比对检验结果, 计数统计结果见表 3。

Table 3. Comparison of two graphical models when recovered data and control data from the different source
表 3. 样本数据和检材数据不同来源时两种图形模型的比较

似然比	$ R < 0.5$	$ R < 0.4$
>1	78	76
$1 \sim 10^{-1}$	17	16
$10^{-1} \sim 10^{-2}$	18	16
$10^{-2} \sim 10^{-3}$	16	20
$10^{-3} \sim 10^{-4}$	12	7
$10^{-4} \sim 10^{-5}$	17	20
$<10^5$	1495	1498
所有样本	1653	1653
真阴性(%)	95.28	95.40
假阳性(%)	4.718	4.597

6. 结论

对于同一来源的玻璃碎片, 假阴性的比例比较大, 有接近 35% 的样品似然比小于 1, 并被判别为来自不同来源。对于不同来源的玻璃碎片, 假阳性占比非常低, 只有 4.7% 左右的碎片似然比大于 1, 并被判别为来自同一来源。虽然相较于 Aitken 和 Lucy 的方法, 两类错误率下降较为明显(在 Aitken 的方法中, 其假阴性率大于 40%, 假阳性率在 7% 左右), 但可以发现, 假阴性的比例依然较高。

在文献[9]中, 对两类错误率的高低情况作出了具体的解释。司法机关会更看重假阳性结果, 因为假阳性结果对应于无辜者的判别结果。而假阴性对于后续的司法工作影响较小。

在进行数据降维时, 分别使用了由 $|R| < 0.5$ 或 $|R| < 0.4$ 导出的两组图形模型。由两类错误率来看, 两组模型基本没有太大的区别。但是考虑到 $|R| < 0.4$ 时导出的图形模型, 所有元素都存在相关性。因此, 考虑使用 $|R| < 0.4$ 时的图形模型。

对于司法机关来说, 似然比强调的是物证检验结果本身的概率之比, 其数值大小反映的是物证支持两个假设的程度, 而不是指来源可能性之比。因此, 在司法实践中, 要想回答物证是否同源、同源的可能性有多大等关键问题, 法官还需要采用似然比与其他证据相互印证、排除有效怀疑的方式等诸多方法[18]进行进一步的评估。

本文描述了一种以非平衡三层次模型的形式评价证据的方法。该方法可适用于高维数据, 通过考虑由缩放逆协方差矩阵得到的图形模型降低维度进行似然比计算, 最终提供了一个低假阳性率的结果, 这在法医科学中是可以接受的。

基金项目

本课题来自公安部物证鉴定中心《基于理化检验数据的似然比证据价值评价方法研究》, 是公安部物证鉴定中心基本科研业务费专项资金自主选题项目, 项目编号为 2022JC12。

参考文献

- [1] 王桂强. 物证鉴定范式发展(二): 贝叶斯似然比框架[J]. 刑事技术, 2024, 49(5): 441-455.
- [2] Bozza, S., Franco, T. and Alex, B. (2022) Bayes Factors for Forensic Decision Analyses with R. Springer Nature, 2-5.
- [3] Menzyk, A., Martyna, A., Damin, A., Vincenti, M. and Zadora, G. (2023) Breaking with Trends in Forensic Dating:

- A Likelihood Ratio-Based Comparison Approach. *Forensic Science International*, **349**, Article ID: 111763. <https://doi.org/10.1016/j.forsciint.2023.111763>
- [4] Falcone, R., Sommariva, G. and Verità, M. (2006) WDXRF, EPMA and SEM/EDX Quantitative Chemical Analyses of Small Glass Samples. *Microchimica Acta*, **155**, 137-140. <https://doi.org/10.1007/s00604-006-0531-z>
 - [5] Guo, H., Hu, C., Wang, P., Mei, H., Li, Y. and Zhu, J. (2023) Application of Lead Isotope Signature and Likelihood Ratio Evaluation in a Shooting Incident Investigation. *Forensic Science International*, **351**, Article ID: 111812. <https://doi.org/10.1016/j.forsciint.2023.111812>
 - [6] Martyna, A., Lucy, D., Zadora, G., Trzcinska, B.M., Ramos, D. and Parczewski, A. (2013) The Evidential Value of Microspectrophotometry Measurements Made for Pen Inks. *Analytical Methods*, **5**, 6788-6795. <https://doi.org/10.1039/c3ay41622d>
 - [7] Klemenc, S. (2001) In Common Batch Searching of Illicit Heroin Samples—Evaluation of Data by Chemometrics Methods. *Forensic Science International*, **115**, 43-52. [https://doi.org/10.1016/s0379-0738\(00\)00306-6](https://doi.org/10.1016/s0379-0738(00)00306-6)
 - [8] 郭洪玲, 王萍, 胡灿, 等. 基于折射率数据的玻璃物证比对检验似然比模型建立及评估[J]. 刑事技术, 2023, 48(4): 355-363.
 - [9] Aitken, C.G.G., Lucy, D., Zadora, G. and Curran, J.M. (2006) Evaluation of Transfer Evidence for Three-Level Multivariate Data with the Use of Graphical Models. *Computational Statistics & Data Analysis*, **50**, 2571-2588. <https://doi.org/10.1016/j.csda.2005.04.005>
 - [10] Taroni, F., Bozza, S., Biedermann, A., Garbolino, P. and Aitken, C. (2010) Data Analysis in Forensic Science: A Bayesian Decision Perspective. Wiley, 323-340. <https://doi.org/10.1002/9780470665084>
 - [11] Zadora, G., Martyna, A., Ramos, D. and Aitken, C. (2013) Statistical Analysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data. Wiley, 187-188. <https://doi.org/10.1002/9781118763155>
 - [12] Curran, J.M., Triggs, C.M., Almirall, J.R., Buckleton, J.S. and Walsh, K.A.J. (1997) The Interpretation of Elemental Composition Measurements from Forensic Glass Evidence: II. *Science & Justice*, **37**, 245-249. [https://doi.org/10.1016/s1355-0306\(97\)72198-1](https://doi.org/10.1016/s1355-0306(97)72198-1)
 - [13] Campbell, G.P. and Curran, J.M. (2009) The Interpretation of Elemental Composition Measurements from Forensic Glass Evidence III. *Science & Justice*, **49**, 2-7. <https://doi.org/10.1016/j.scijus.2008.09.001>
 - [14] 赵军圣, 庄光明, 王增桂. 极大似然估计方法介绍[J]. 长春理工大学学报, 2010, 5(6): 53-54.
 - [15] Zadora, G., Neocleous, T. and Aitken, C. (2010) A Two-Level Model for Evidence Evaluation in the Presence of Zeros. *Journal of Forensic Sciences*, **55**, 371-384. <https://doi.org/10.1111/j.1556-4029.2009.01316.x>
 - [16] Zadora, G., Martyna, A., Ramos, D. and Aitken, C. (2013) Statistical Analysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data. Wiley, 99-105. <https://doi.org/10.1002/9781118763155>
 - [17] Hayes, J.F. and Hill, W.G. (1980) A Reparameterization of a Genetic Selection Index to Locate Its Sampling Properties. *Biometrics*, **36**, 237-248. <https://doi.org/10.2307/2529975>
 - [18] Aitken, C., Roberts, P. and Jackson, G. (2010) Communicating and Interpreting Statistical Evidence in the Administration of Criminal Justice. Practitioner Guide No. 1: Fundamentals of Probability and Statistical Evidence in Criminal Proceedings. Royal Statistical Society.