Transformations for compositional data with zeros with an application to forensic evidence evaluation

Tereza Neocleous*[,a], Colin Aitken[b], Grzegorz Zadora[c]

[a]*School of Mathematics and Statistics, University of Glasgow, 15 University Gardens, Glasgow, G12 8QW, UK.*
[b]*School of Mathematics and The Joseph Bell Centre for Forensic Statistics and Legal Reasoning, The King's Buildings, The University of Edinburgh, Mayfield Road, Edinburgh, EH9 3JZ, UK.*
[c]*Institute of Forensic Research, Westerplatte 9, PL-31-033, Krakow, Poland.*

Abstract

In forensic science likelihood ratios provide a natural way of computing the value of evidence under competing propositions such as "the compared samples have originated from the same object" (prosecution) and "the compared samples have originated from different objects" (defence). We use a two-level multivariate likelihood ratio model for comparison of forensic glass evidence in the form of elemental compositions data under three data transformations: the logratio transformation, a complementary log-log type transformation and a hyperspherical transformation. The performances of the three transformations in the evaluation of evidence are assessed in simulation experiments through use of the proportions of false negatives and false positives.

Key words: likelihood ratio; compositional data; physicochemical data; glass fragments; forensic science.

1. Introduction

Statistical approaches to the evaluation of evidence of a forensic scientific nature have been developed over many years following a seminal paper by Lindley [16]. The underlying principle is that of the odds version of Bayes' Theorem. Two propositions are considered, thought of as the one put forward

*Corresponding author
Email address: tereza.neocleous@glasgow.ac.uk (Tereza Neocleous)

by the prosecution, denoted here as H_p, and the one put forward by the defence, denoted here as H_d. Denote the evidence to be evaluated by E. Then the odds form of Bayes' Theorem may be written as

$$\frac{\Pr(H_p \mid E)}{\Pr(H_d \mid E)} = \frac{\Pr(E \mid H_p)}{\Pr(E \mid H_d)} \frac{\Pr(H_p)}{\Pr(H_d)}.$$

In text, this may be written as the posterior odds in favour of the prosecution proposition equals the product of the likelihood ratio ($LR = \Pr(E \mid H_p)/\Pr(E \mid \Pr(H_d))$ and the prior odds in favour of the prosecution proposition. Values of LR above 1 support H_p and values of LR below 1 support H_d. A value of LR close to 1 provides little support for either proposition. Also the larger (the lower) the value of the LR, the stronger (the weaker) the support of E for H_p.

A particular type of forensic evidence for which this approach is very appropriate is the type known as trace evidence, which is simply evidence that is in the form of traces, such as traces (or fragments) of glass, traces (or stains) of blood or semen, or traces of gun-shot residue. Trace evidence that is of particular interest is that known as transfer evidence for the reason that it is transferred from one place to another. The particular example used here to illustrate the method described will be that of fragments of glass.

Evidence is evaluated by a comparison of trace evidence found at a crime scene with trace evidence, that corresponds in some sense to the crime scene evidence, found in association with a suspect. The rarity and similarity of the two sets of evidence are assessed with reference to a background population of the same type of evidence.

A sample of evidence whose origin is known is called a control sample. A sample of evidence whose origin is not known is called a recovered sample. For example, consider a window broken at a crime scene in the course of a burglary. Fragments of glass from the window will form a control sample. A suspect is identified and fragments of glass are found on his clothing. These form a recovered sample as their origin is not known. It may be the window at the crime scene but it may not. The assumed transfer is that of glass fragments from the crime scene to the criminal. Alternatively, a footmark may be found in soil beneath the window which could be thought to come from a shoe worn by the burglar. The footmark is a recovered sample as the shoe that made the mark is unknown. A shoe is found at the suspect's home. A mark made by that shoe would be a control mark as its origin is known.

Thus it is that control and recovered samples may or may not be associated with crime scenes or suspects.

Various types of materials such as glass fragments are routinely subjected to physico-chemical examination by forensic scientists. For example, glass fragments, identified as coming from a car headlamp, could be obtained from the debris on the road or from the clothes of the victim of a hit-and-run accident. This could comprise a mixture of pieces of sand, soil, dust, glass and other transfer evidence. The glass fragments are of interest for this paper and form the recovered sample. Their origin is not known; they may or may not have come from the car involved in the hit-and-run accident. A suspect car is identified for reasons other than those of the characteristics of its glass. Glass from its headlamps is examined. This glass is the control sample; its origin is known.

One of the purposes of analysing materials found in debris is to address the question whether two samples (*e.g.* a glass fragment found on the clothes of the victim of a hit-and-run accident, and a glass fragment collected from the suspected car) could have originated from the same object. The size of recovered fragments of glass is very small (of linear dimension 0.1-0.5 mm), and therefore this task requires information obtained during physico-chemical analysis; that is quantitative and semi-quantitative data such as the concentration of elements in a glass fragment [3, 2]. The question is addressed by considering the likelihood ratio $LR = \dfrac{\Pr(E \mid H_p)}{\Pr(E \mid H_d)}$ where E denotes the measurements on the control and recovered samples of glass, H_p denotes the proposition that the control and recovered samples came from the same source and H_d denotes the proposition that the control and recovered samples came from different sources.

The importance of glass as evidence has been recognised for many years (see, for instance, [5, 7]). The GRIM (Glass Refractive Index Measurement) method and Scanning Electron Microscopy coupled with an Energy Dispersive X-ray spectrometer (SEM-EDX) are routinely used in many forensic institutes for the investigation of glass and other forensic problems [3, 24]. Other methods of elemental analysis of glass fragments are μ-X-Ray Fluorescence [11] and Laser Ablation-Inductively Coupled Plasma-Mass Spectrometry [22]. However, these methods require relatively large fragments of glass; for example LA-ICP-MS gives reliable results with pieces of glass larger than 0.5 mm. SEM-EDX has the drawback that it can only provide information about major and minor elements, such as oxygen (O), sodium (Na), magne-

sium (Mg), aluminium (Al), silicon (Si), potassium (K), calcium (Ca) and iron (Fe), from any glass fragment. Trace elements exist in concentrations below the detection limits of this method. It is commonly believed that trace element concentrations are essential to enable the glass investigator to compare glass evidence effectively. However, some progress can be made on the basis of only the major and minor element concentrations [3, 24]. These data could be used to test the same-source hypothesis if recovered and control glass samples are available.

The evaluation of evidence in this context is based on analytical data obtained during the physico-chemical analysis. Comparison of the control and recovered materials requires that careful attention be paid to the following considerations.

1. The possible sources of uncertainty which will include, at least:
   (a) the variation of measurements of characteristics within the recovered and control items,
   (b) the variation of measurements of characteristics between various objects in the relevant population (*e.g.* glass object population);
2. Information about the rarity of the determined physico-chemical characteristics (*e.g.* elemental composition of compared samples) for recovered and control samples in the relevant population;
3. The level of association between different characteristics when more than one characteristic has been measured; and
4. Information about the similarity of the recovered material to the control sample.

Consider a case where the fact finder such as a prosecutor or judge asks a forensic scientist to evaluate evidence in the form of a recovered material, of unknown origin, and a control material, whose origin is known. The result of such a comparison will be referred to as E. The relevant propositions for the fact finder arise from the circumstances of the case. Often, because of the adversarial nature of legal systems, they are:

- H_p: the control and recovered samples come from the same source (*prosecution proposition*),
- H_d: the control and recovered samples come from different sources both belonging to a relevant population (*defence proposition*).

The rest of the paper considers an approach to obtaining the likelihood ratio for the strength of evidence under propositions H_p and H_d, when the evidence is in the form of compositional data arising from a forensic glass database. The dataset and the data transformations considered are described in Section 2.

Section 3 describes in detail the statistical approach used for obtaining the likelihood ratios. Section 4 describes the simulation experiments performed to assess each of the transformations and finally Section 5 discusses the results of the method comparisons.

2. Physicochemical glass data

Three replicate measurements were made of the elemental concentrations of each of four glass fragments, with surfaces as smooth and flat as possible, collected from each of 320 glass objects (105 building windows, 94 car windows, 26 bulbs, 16 headlamps and 79 containers. The elemental concentrations measured were those of oxygen (O), sodium (Na), magnesium (Mg), aluminium (Al), silicon (Si), potassium (K), calcium (Ca) and iron (Fe).

The mean of the three replicate measurements from each fragment was used for the analyses. The variance in the replications was very much smaller than the variance between the four fragments so it has been ignored. The data consist of eight variables which represent the % wt. of each of the eight elements whose concentrations were measured. The data are compositional as they add up to 100%, and they often include zero concentrations of certain elements. Percentages of zeros for each variable are given in Table 1. Physico-chemical data frequently contain zero values. In glass the presence or absence of a particular component is related to the nature of the object analysed; for instance, iron is an additive used in order to obtain a green or brown colour. It is also the cheapest additive that adds colour to a glass object as it is present in sand. However, unless added at the manufacturing stage, iron appears in concentrations that are usually below the detection limits of the SEM-EDX method and as a result most of the iron values recorded are zero. Hence it can be either argued that zero iron concentrations are structural zeros, or that they are simply below-detection-limit values. Similarly magnesium, aluminium and potassium could appear in concentrations below the detection limit of SEM-EDX, while at the same time oxygen, sodium, silicon and calcium concentrations are non-zero for soda-lima-silica glass.

In analysing the data it is necessary to take into account their compositional nature and the presence of zero values. Compositional data provide information about relative values of components, and therefore ratios can be used to model them. In particular, the logratio transformation [1] of a composition $\mathbf{z} = (z_1, \ldots, z_P)$ with $z_P \neq 0$ and $\sum_{i=1}^{P} z_i = 1$ is given by

$$u_1 = \log_{10}\left(\frac{z_1}{z_P}\right), \ldots, u_p = \log_{10}\left(\frac{z_{P-1}}{z_P}\right). \tag{1}$$

This reduces the data vector to $\mathbf{u} = (u_1, \ldots, u_{P-1})$, of dimension $p = P - 1$, which removes the problem of the constrained sample space and transforms the data closer to normality. Another nice characteristic of the logratio transformation (sometimes also referred to as additive logratio or alr transformation) is its invariance to permutations.

When some (in certain cases many) of the $\{z_i\}$ are zero, they can be replaced by a very small number to enable computation of the logratio. This implicitly assumes that zero values are simply values below the detection limit of the measuring equipment. Methods for choosing a suitable small number have been proposed in [10], [17] and [18] among others, the latter two detailing a parametric and nonparametric approach respectively. In practice the simpler approach of replacing zeros by a small constant (0.0001 for this application) appears to work reasonably well. Alternatively the presence or absence of certain components can itself be modelled if the zeros are assumed to be structural; see [26] for a recent example of such modelling.

For the glass data the ratios are taken with respect to oxygen, with zero concentrations substituted by 0.0001 before taking the logarithm of the ratio. The resulting data vector $\mathbf{u}$ contains $p = 7$ variables:

$$\mathbf{u} = \left(\log_{10}\frac{Na}{O}, \log_{10}\frac{Mg}{O}, \log_{10}\frac{Al}{O}, \log_{10}\frac{Si}{O}, \log_{10}\frac{K}{O}, \log_{10}\frac{Ca}{O}, \log_{10}\frac{Fe}{O}\right).$$

A further improvement in the normality of data can be achieved by a complementary log-log type transformation, which involves taking the logarithm of the negative of the logratio-transformed data. This is possible if all logratios are negative, that is, if the concentration of oxygen is always larger than those of the other elements. For glass this is usually the case, but in our dataset there were two exceptions: two fragments of glass for which the silicon concentration was slightly higher than the oxygen concentration. For

this reason, a small constant was added to the logratios before taking the logarithm. The result is a data vector $\mathbf{v}$, to be referred to as *complementary log-log*, with components

$$v_1 = \log_{10}(-u_1 + 0.01), \ldots, v_p = \log_{10}(-u_p + 0.01). \tag{2}$$

An alternative approach to the logratio and the complementary log-log is a spherical transformation [23]. The compositional vectors $\mathbf{z} = (z_1, \ldots, z_P)$ are transformed by first taking the square root, $s_i = \sqrt{z_i},\ i = 1, \ldots, P$, and then applying the following recursive relationship.

$$\begin{aligned}
\omega_1 &= \arccos(s_1) \\
\omega_2 &= \arccos\left(\frac{s_2}{\sin\omega_1}\right) \\
\omega_3 &= \arccos\left(\frac{s_3}{\sin\omega_2 \sin\omega_1}\right) \\
&\vdots \\
\omega_{P-1} &= \arccos\left(\frac{s_{P-1}}{\sin\omega_1 \sin\omega_2 \ldots \sin\omega_{P-2}}\right)
\end{aligned} \tag{3}$$

The compositions lie on the unit hypersphere and this transformation essentially maps the Cartesian coordinates to polar coordinates. Note that the dimension of the resulting $\boldsymbol{\omega}$-vector is $p = P-1$, and the zeros simply map to $\arccos(0) = \pi/2$. Ordering the variables based on concentration from highest to lowest, and taking oxygen to be the Pth variable, the resulting data vector is

$$\boldsymbol{\omega} = (\omega_{Si}, \omega_{Na}, \omega_{Ca}, \omega_{Mg}, \omega_{Al}, \omega_K, \omega_{Fe}).$$

To summarise, the following three data transformations were considered:

1. logratio (with zeros substituted by 0.0001) given by expression (1);
2. complementary log-log given by expression (2);
3. spherical given by expression (3).

3. Statistical methods

3.1. Two-level random effects model

The glass database consists of $m = 320$ objects with $n = 4$ measurements each (corresponding to four fragment means from each object) of $P = 8$

variables in the form of compositions $\{z_{ijk}\}, i = 1, \ldots, m, j = 1, \ldots, n, k = 1, \ldots, P$ with $z_{ijP} \neq 0$ and $z_{ij1} + \ldots + z_{ijP} = 1$. Given the sum constraint on the compositions, $p = P - 1$ variables suffice for describing such data and thus, for all the analyses presented here, data transformations were applied that result in $p = P - 1$ variables.

Denote the database of m objects with p variables each of which is measured n times within each object, by

$$\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^\top; i = 1, \ldots, m, j = 1, \ldots, n,$$

giving a total of $N = nm$ sets of p measurements. Suppose that two sets exist (control and recovered), one of n_1 and one of n_2 measurements and a comparison between the two sets is required. Let $\bar{\mathbf{y}}_1$ be a vector of means of the n_1 measurements $\mathbf{y}_{1j}$, $j = 1, 2, \ldots, n_1$ and $\bar{\mathbf{y}}_2$ be a vector of means of the n_2 measurements $\mathbf{y}_{2j}$, $j = 1, 2, \ldots, n_2$ from the second object.

Two sources of variation are considered, that between replicates within the same object (within-object variability) and that between objects (between-object variability). Following [25], it is assumed that the within-object distribution is normal with constant variance. The between-object distribution can be estimated either assuming multivariate normality (Model 1), or, more realistically, using density estimation with Gaussian kernels (Model 2).

Denote the mean vector within the ith object by $\boldsymbol{\theta}_i$ and the within-object covariance matrix by $\mathbf{U}$. Then, given $\boldsymbol{\theta}_i$ and $\mathbf{U}$,

$$(\mathbf{X}_{ij} \mid \boldsymbol{\theta}_i, \mathbf{U}) \sim N_p(\boldsymbol{\theta}_i, \mathbf{U});\ i = 1, \ldots, m, j = 1, \ldots, n.$$

Under Model 1 it is assumed that

$$(\boldsymbol{\theta}_i \mid \boldsymbol{\mu}, \mathbf{C}) \sim N_p(\boldsymbol{\mu}, \mathbf{C});\ i = 1, \ldots, m,$$

while under Model 2 the between-source distribution is estimated from the group means, $\bar{\mathbf{x}}_i = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{ij}$ using a multivariate normal kernel density function with mean $\bar{\mathbf{x}}_i$ and covariance matrix $\mathbf{H}$, and denoted by $K(\boldsymbol{\theta} \mid \bar{\mathbf{x}}_i, \mathbf{H})$ where

$$K(\boldsymbol{\theta} \mid \bar{\mathbf{x}}_i, \mathbf{H}) = (2\pi)^{-p/2}|\mathbf{H}|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \bar{\mathbf{x}}_i)^\top \mathbf{H}^{-1}(\boldsymbol{\theta} - \bar{\mathbf{x}}_i)\right\} \quad (4)$$

is a multivariate Gaussian kernel function.

The estimate $f(\boldsymbol{\theta} \mid \bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_m, \mathbf{H})$ of the between-object probability distribution function under Model 2 is then

$$f(\boldsymbol{\theta} \mid \bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_m, \mathbf{H}) = \frac{1}{m} \sum_{i=1}^{m} K(\boldsymbol{\theta} \mid \bar{\mathbf{x}}_i, \mathbf{H})$$

which is a function of the object means, $\bar{\mathbf{x}}_i$, and the kernel bandwidth matrix $\mathbf{H}$.

The between- and within-group covariance matrices $\mathbf{U}$ and $\mathbf{C}$ can be estimated from the background database of m objects by

$$\hat{\mathbf{U}} = \frac{\mathbf{S}_w}{m(n-1)}$$

where

$$\mathbf{S}_w = \sum_{i=1}^{m} \sum_{j=1}^{n} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^\top,$$

and by

$$\hat{\mathbf{C}} = \frac{\mathbf{S}^*}{m-1} - \frac{\mathbf{S}_w}{nm(n-1)}$$

where

$$\mathbf{S}^* = \sum_{i=1}^{m} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top,$$

respectively as discussed in [4].

Under Model 2 the numerator of the likelihood ratio, for which H_p is assumed true, can be shown to be given by:

$$\begin{aligned} f(\bar{\mathbf{y}}_\mathbf{1}, \bar{\mathbf{y}}_\mathbf{2} | H_p) &= f(\bar{\mathbf{y}}_\mathbf{1} - \bar{\mathbf{y}}_\mathbf{2}, \bar{\mathbf{y}}^* | \mathbf{U}, \mathbf{H}) = \\ &= (2\pi)^{-p/2} \left| \frac{\mathbf{U}}{n_1} + \frac{\mathbf{U}}{n_2} \right|^{-1/2} \exp\left\{ -\frac{1}{2} (\bar{\mathbf{y}}_\mathbf{1} - \bar{\mathbf{y}}_\mathbf{2})^\top \left(\frac{\mathbf{U}}{n_1} + \frac{\mathbf{U}}{n_2} \right)^{-1} (\bar{\mathbf{y}}_\mathbf{1} - \bar{\mathbf{y}}_\mathbf{2}) \right\} \\ &\quad \times \frac{1}{m} \sum_{i=1}^{m} \left\{ (2\pi)^{-p/2} \left| \frac{\mathbf{U}}{n_1 + n_2} + \mathbf{H} \right|^{-1/2} \right. \\ &\quad \left. \times \exp\left[-\frac{1}{2} (\bar{\mathbf{y}}^* - \bar{\mathbf{x}}_\mathbf{i})^\top \left(\frac{\mathbf{U}}{n_1 + n_2} + \mathbf{H} \right)^{-1} (\bar{\mathbf{y}}^* - \bar{\mathbf{x}}_\mathbf{i}) \right] \right\} \end{aligned} \tag{5}$$

where

$$\bar{\mathbf{y}}^* = \frac{n_1\bar{\mathbf{y}}_\mathbf{1} + n_2\bar{\mathbf{y}}_\mathbf{2}}{n_1 + n_2}. \tag{6}$$

Note that expression (5) is a simplified version of the corresponding equation in [4], which uses a bandwidth matrix of the form $\mathbf{H} = h^2\mathbf{C}$. Similarly, under Model 2 the denominator of the likelihood ratio, for which H_d is assumed true, can be shown to be given by:

$$f(\bar{\mathbf{y}}_\mathbf{1}, \bar{\mathbf{y}}_\mathbf{2}|H_d) = f(\bar{\mathbf{y}}_\mathbf{1}|\mathbf{U}, \mathbf{H})f(\bar{\mathbf{y}}_\mathbf{2}|\mathbf{U}, \mathbf{H})$$

with

$$f(\bar{\mathbf{y}}_\mathbf{l}|\mathbf{U}, \mathbf{H}) = \frac{(2\pi)^{-p/2}}{m}\left|\frac{\mathbf{U}}{n_l} + \mathbf{H}\right|^{-1/2} \sum_{i=1}^{m} \exp\left\{-\frac{1}{2}(\bar{\mathbf{y}}_\mathbf{l} - \bar{\mathbf{x}}_\mathbf{i})^\top \left(\frac{\mathbf{U}}{n_l} + \mathbf{H}\right)^{-1} (\bar{\mathbf{y}}_\mathbf{l} - \bar{\mathbf{x}}_\mathbf{i})\right\} \tag{7}$$

for $l = 1, 2$ and $i = 1, \ldots, m$.

Under Model 1, which assumes multivariate normality for the between-object distribution, expression (5) for the numerator of the likelihood ratio is replaced by

$$\begin{aligned}(2\pi)^{-p}&\left|\frac{\mathbf{U}}{n_1} + \frac{\mathbf{U}}{n_2}\right|^{-1/2} \exp\left\{-\frac{1}{2}(\bar{\mathbf{y}}_\mathbf{1} - \bar{\mathbf{y}}_\mathbf{2})^\top \left(\frac{\mathbf{U}}{n_1} + \frac{\mathbf{U}}{n_2}\right)^{-1} (\bar{\mathbf{y}}_\mathbf{1} - \bar{\mathbf{y}}_\mathbf{2})\right\} \\ &\times \left|\frac{\mathbf{U}}{n_1 + n_2} + \mathbf{C}\right|^{-1/2} \exp\left\{-\frac{1}{2}(\bar{\mathbf{y}}^* - \boldsymbol{\mu})^\top \left(\frac{\mathbf{U}}{n_1 + n_2} + \mathbf{C}\right)^{-1} (\bar{\mathbf{y}}^* - \boldsymbol{\mu})\right\}\end{aligned} \tag{8}$$

and expression (7) for the denominator by

$$f(\bar{\mathbf{y}}_\mathbf{l}|\boldsymbol{\mu}, \mathbf{U}, \mathbf{C}) = (2\pi)^{-p/2}\left|\frac{\mathbf{U}}{n_l} + \mathbf{C}\right|^{-1/2} \exp\left\{-\frac{1}{2}(\bar{\mathbf{y}}_\mathbf{l} - \boldsymbol{\mu})^\top \left(\frac{\mathbf{U}}{n_l} + \mathbf{C}\right)^{-1} (\bar{\mathbf{y}}_\mathbf{l} - \boldsymbol{\mu})\right\} \tag{9}$$

for $l = 1, 2$ and $i = 1, \ldots, m$ with $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i$ [4].

Parameter estimates for $\boldsymbol{\mu}, \mathbf{C}$ and $\mathbf{U}$ obtained from the entire database of $m = 320$ objects, for each of the three data transformations, are shown in Tables 2-5.

3.2. Dimension reduction using graphical models

High-dimensional estimation and in particular multivariate kernel density estimation in seven dimensions can be computationally challenging. Graphical models [15], a probabilistic tool for studying and visualising conditional independence relationships between random variables, is used as a way to reduce the seven-dimensional problem into several lower-dimensional problems without disregarding potentially informative interdependencies in the variables measured. Values of partial correlation can be used to construct a decomposable graphical model of the full density into cliques representing the product of several density functions in lower dimensions. Various data-driven methods have been proposed for obtaining a graphical model from the partial correlation coefficients. In this work, the method used was the PC algorithm (named after its authors, Peter and Clark [21]). This algorithm starts from a complete graph and recursively deletes edges based on conditional independence. In [12] asymptotic consistency of the PC algorithm is shown for Gaussian data, and in [13] a robust version of the algorithm is proposed. The PC algorithm was implemented in R [19] using the `pcalg` package [14]. The robust version was also applied in this work, but did not give good results because it is designed to deal with outliers, not major deviations from normality. Thus, the robust PC algorithm-generated graphical models and results obtained using these models are not presented here.

In addition to the PC algorithm-generated graphical models shown in Figs 1-3 for each of the three data transformations, a simplified graphical model which is based on inspection of the partial correlations matrix was also considered. For an example of how a graphical model can be constructed from the partial correlations matrix, and how the corresponding density factorisation is obtained, see [25]. For the logratio-transformed data and the complementary log-log transformed data, the same model was obtained (shown in Fig. 4), which captures the main chemical relationships between the various glass components. This model gave the density factorisation

$$f(Na', Mg', Al', Si', K', Ca', Fe') = f(Na', Si', Ca')f(Al', K')f(Mg')f(Fe'). \tag{10}$$

where the prime denotes the transformed version of the variable corresponding to each element.

For the spherically transformed data, a slightly different graphical model was obtained (shown in Fig. 5), which resulted in the density factorisation

$$
\begin{aligned}
&f(Na', Mg', Al', Si', K', Ca', Fe') \\
&= \frac{f(Si', Ca')f(Na', Al', K')f(Ca', K')f(Mg', Al')f(Fe')}{f(Ca')f(K')f(Al')}.
\end{aligned} \tag{11}
$$

3.3. Bandwidth selection for kernel density estimation

Under Model 2, which utilises kernel density estimation for the between-object distribution, the numerator (5) and the denominator (7) of the likelihood ratio are estimated using multivariate Gaussian kernels with bandwidth matrix $\mathbf{H}$. This matrix was estimated in two different ways:

KDE1 – Following [4], assume that $\mathbf{H}$ is of the form $h^2\mathbf{C}$ and use a rule-of-thumb formula based on [20] for estimating h:

$$
\hat{h} = \left(\frac{4}{2p+1}\right)^{1/(p+4)} m^{-1/(p+4)}. \tag{12}
$$

KDE2 – Allow $\mathbf{H}$ to be an unconstrained matrix obtained using least squares cross-validation or smoothed cross-validation as described in [9]. Estimation of the unconstrained bandwidth matrix $\mathbf{H}$ was implemented using the `ks` package [8] in R. Due to computational difficulties with the smoothed cross-validation method for high-dimensional data, this method was replaced by least squares cross-validation when working with the full seven-dimensional density.

Note that both of the above kernel density estimation procedures were applied to the logratio, complementary log-log and spherically transformed data. For an alternative approach to kernel density estimation for compositional data using a plug-in method see [6].

4. Simulation experiments

The performance of each method and data transformation was assessed in terms of the percentage of false negative and positive answers. A false negative answer (type I error) is an answer where compared glass samples originating from the same glass sample are evaluated as having originated from different glass samples ($LR < 1$). A false positive answer (type II error) is an answer where compared glass samples originating from different

glass objects are evaluated as having originated from the same glass object ($LR > 1$). Control of the level of false positive answers is especially important from the forensic point of view as the statement that two samples of glass could have the same origin, which does not correspond with the true facts, could have serious legal consequences for the suspect.

Four-fold cross-validation was used in the simulation experiments, for which the data were divided into four parts at random. There were 320 items altogether, so each part consisted of 80 items. One part was kept as the test data, and the rest of the data were considered as the training set, from which parameters were estimated and graphical models obtained. This was repeated four times, yielding four sets of false positive (FP) and false negative (FN) rates from each test set. Each cell in Table 6 shows the average of those four rates.

The following experiments were performed in order to study the level of false positive and false negative answers:

1. Experiment 1 (estimation of the percentage of false negative answers). The measurements of the first two glass fragments of a total of four analysed from a particular glass object were selected for the simulated measurements of sample A (recovered). The measurements of the other two glass fragments were assigned to sample B (control). Each simulated sample A was compared with a simulated sample B. Four such sets of 80 objects were created and a total of 320 comparisons were made. The desirable answer was $LR > 1$ and each answer with $LR < 1$ was a false negative answer.
2. Experiment 2 (estimation of percentage of false positive answers). All four measurements from each of two different glass objects were selected to form a pair of samples to compare, *i.e.* samples A and B. Four sets of 80 glass samples were available in the database, and thus $4 \times \binom{80}{2} = 12,640$ such pairs were formed. The desirable answer was $LR < 1$ and each answer with $LR > 1$ was a false positive answer.

5. Results

The results of the simulation experiment described in Section 4 are shown in Table 6 in the form of false positive and false negative rates. Three models were considered: normal, which assumes that the between-object distribution

is (multivariate) normal, and KDE1 and KDE2 which estimate the between-object distribution using Gaussian kernels. The difference between KDE1 and KDE2 is that the former uses a kernel bandwidth matrix of the form $\mathbf{H} = h^2\mathbf{C}$, where h is estimated by expression (12), while the latter estimates an unconstrained kernel bandwidth matrix $\mathbf{H}$ using cross-validation as described in Section 3.3 above. These models were applied to the three transformed datasets described in Section 2: logratio, complementary log-log and spherically transformed data. Firstly the full seven-dimensional density (denoted as Full in Table 6) was estimated under propositions H_p and H_d and the likelihood ratio was obtained. In addition, the likelihood ratio was obtained under the two density factorisations described in Section 3.2 based on decomposable graphical models which are denoted by GM1 and GM2. GM1 is the model obtained using the PC algorithm and it is shown for the three data transformations in Figs 1-3 respectively. GM2 is the graphical model obtained based on the main relationships between the chemical elements that form glass and is shown in Fig. 4 for the logratio and complementary log-log transformed data. Its resulting factorisation given by expression (10) has been previously used for logratio-transformed data in [26]. This model corresponds well with the partial correlations matrices for the logratio and complementary log-log transformed data. For spherically transformed data a slightly different graphical model was obtained based on inspection of the partial correlations matrix with corresponding density factorisation given by expression (11). In general there were similarities between the graphical models obtained using the PC algorithm and inspection of the partial correlations matrices for each data transformation. Also worth noting is that the graphical models obtained for each of the four subsets used in the four-fold cross-validation procedure for the simulation experiments, are almost identical to each other as can be seen in Figs 1 - 3.

The false positive and negative rates obtained from all methods and data transformations range from 1.9% to 5.9%, rates which indicate good performance in general. Low false positive rates are of particular interest as the error to which they apply is that of convicting an innocent person. This is generally thought to be a much worse error than failing to convict a guilty person. With this criterion, the spherical transformation is preferable to the logratio and complementary log-log transformation. The spherical transformation has the additional advantage of enabling the analysis of data including zeros; the other models require the addition of a small amount to zero

or a more sophisticated model that allows for zeros. Kernel density estimation with an unconstrained bandwidth matrix (KDE2) and a simplified graphical-model based factorisation of the density with a spherical transformation yields the lowest false positive rates (1.9%).

All error rates are less than 6% so all could be used in practice. The full model has lower false positive rates than GM1 or GM2 for the Normal model. Results from models that use a graphical model-based density factorisation have lower false positive rates than for the full model for kernel density estimation with an unconstrained bandwidth matrix (KDE2). The false negative rates are higher for the logratio and log(-logratio) transformations with GM1 and GM2 than for the full model (again for KDE2).

In general GM1 and GM2 are to be preferred to the Full model as they require the estimation of only one-, two- or three-dimensional density functions. The Full model uses seven variables. Density estimation for such a large number of variables has a larger error associated with the quality of the fit. For example, the sample size required (accurate to about three significant figures) to ensure that the relative mean square error at zero is less than 0.1, when estimating a standard multivariate normal density using a normal kernel and a window width that minimizes the mean square error at zero is 4 for one dimension, 19 for two, 67 for three and 10,700 for seven dimensions [20]. The simulation results here are based on samples of size 80 so satisfy these criteria up to three dimensions.

There are three factors to consider when evaluating evidence in the form of compositional data, as shown in Table 6. For the reasons given above, it is recommended that the evaluation of evidence in the form of compositional data be made with

- a kernel density estimation procedure with an unconstrained bandwidth matrix;
- a spherical transformation of the data;
- a simplified graphical model based on the partial correlation matrix.

This method has the lowest false positive rate and requires only estimation of univariate, bivariate and three-dimensional densities as seen in expression (11).

Acknowledgments

The authors are grateful to Josep Martín-Fernández for helpful comments and suggestions on an early version of this article.

References

[1] J. Aitchison. *The statistical analysis of compositional data.* Chapman & Hall, London, 1986.

[2] C. G. G. Aitken, D. Lucy, G. Zadora, and J. M. Curran. Evaluation of trace evidence for three-level multivariate data with the use of graphical models. *Computational Statistics and Data Analysis*, 50:2571–2588, 2006.

[3] C. G. G. Aitken, G. Zadora, and D. Lucy. A two level model for evidence evaluation. *Journal of Forensic Sciences*, 52:412–419, 2007.

[4] C.G.G. Aitken and D. Lucy. Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 53:109–122, 2004.

[5] B. Caddy. *Forensic examination of glass and paint.* CRC Press, Boca Raton, FL, 2001.

[6] J. E. Chacón, G. Mateu-Figueras, and J. A. Martín-Fernández. Gaussian kernels for density estimation with compositional data. To appear in *Computers & Geosciences.* DOI: 10.1016/j.cageo.2009.12.011.

[7] J. M. Curran, T. N. Hicks, and J. S. Buckleton. *Forensic interpretation of glass evidence.* CRC Press, Boca Raton, FL, 2000.

[8] T. Duong. ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *Journal of Statistical Software*, 21(7), 2007.

[9] T. Duong and M. L. Hazelton. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, 32:485–506, 2005.

[10] J. Fry, T. Fry, and K. McLaren. Compositional data analysis and zeros in micro data. *Applied Economics*, 32:953–959, 2000.

[11] T. C. Hicks, F. Monard Sermier, T. Goldmann, A. Brunelle, C. Champod, and P. Margot. The classification and discrimination of glass fragments using non destructive energy dispersive x-ray fluorescence. *Forensic Science International*, 137:107–118, 2003.

[12] M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8:613–636, 2007.

[13] M. Kalisch and P. Bühlmann. Robustification of the PC-algorithm for directed acyclic graphs. *Journal of Computational and Graphical Statistics*, 17:773–789, 2008.

[14] M. Kalisch, N. Maechler, and D. Colombo. *pcalg: Estimation of CPDAG/PAG and causal inference using the IDA algorithm*, 2010. R package version 1.1-2.

[15] S. L. Lauritzen. *Graphical models*. The Clarendon Press Oxford University Press, New York, 1996. Oxford Science Publications.

[16] D. V. Lindley. A problem in forensic science. *Biometrika*, 64:207–213, 1977.

[17] J. A. Martín-Fernández, C. Barceló-Vidal, and V. Pawlowsky-Glahn. Dealing with zeros and missing values in compositional data sets. *Mathematical Geology*, 35:253–278, 2003.

[18] J. Palarea-Albaladejo and J. A. Martín-Fernández. A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. *Computers & Geosciences*, 34:902–917, 2008.

[19] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[20] B. W. Silverman. *Density estimation for statistics and data analysis*. Chapman and Hall, London, 1986.

[21] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction and search*. MIT Press, Cambridge, MA, 2nd edition, 2000.

[22] T. Trejos and J. R. Almirall. Sampling strategies for the analysis of glass fragments LA-ICP-MS Part 1: micro-homogeneity study of glass and its application to the interpretation of forensic evidence. *Talanta*, 67:388–395, 2005.

[23] H. Wang, Q. Liu, H. Mok, L. Fu, and W. Man Tse. A hyperspherical transformation forecasting model for compositional data. *European Journal of Operational Research*, 179:459–468, 2007.

[24] G. Zadora. Glass analysis for forensic purposes - a comparison of classification methods. *Journal of Chemometrics*, 21:174–186, 2007.

[25] G. Zadora and T. Neocleous. Likelihood ratio model for classification of forensic evidence. *Analytica Chimica Acta*, 642:266–278, 2009.

[26] G. Zadora, T. Neocleous, and C. G. G. Aitken. A two-level model for evidence evaluation in the presence of zeros. *Journal of Forensic Sciences*, 55:371–384, 2010.

6. Tables

Table 1: Number (out of 320) and percentage of objects with zeros for each variable in the glass data.

Variable	O	Si	Na	Ca	Mg	Al	K	Fe
Number	0	0	0	9	22	17	97	253
Percentage	0.0	0.0	0.0	2.8	6.9	5.3	30.3	79.1

Table 2: Sample mean vector, $\hat{\boldsymbol{\mu}}$, for each transformation of the glass data.

Transformation	Si	Na	Ca	Mg	Al	K	Fe
logratio	-0.170	-0.718	-1.070	-1.792	-2.179	-3.270	-4.955
log (- log ratio)	-0.768	-0.140	-0.007	0.212	0.321	0.464	0.669
spherical	0.957	1.186	1.256	1.400	1.471	1.508	1.553

Table 3: Sample within-object variance-covariance matrix, $\hat{\mathbf{U}}$, and between-object variance-covariance matrix, $\hat{\mathbf{C}}$, for the logratio transformation of the glass data. The values shown are the variances and covariances multiplied by 1000.

$\hat{\mathbf{U}}$	Si	Na	Ca	Mg	Al	K	Fe
Si	1.066	-0.002	1.911	0.227	0.456	1.292	0.706
Na		0.202	0.108	0.201	-0.125	-0.18	0.022
Ca			3.831	0.558	0.596	2.147	1.305
Mg				0.857	-0.53	0.029	0.172
Al					8.166	0.682	0.269
K						18.74	0.829
Fe							6.737
$\hat{\mathbf{C}}$	Si	Na	Ca	Mg	Al	K	Fe
Si	1.864	1.800	20.428	15.501	-6.363	-10.140	7.409
Na		5.118	49.045	41.268	-5.631	-37.855	11.131
Ca			655.656	535.343	-60.497	-337.730	131.239
Mg				1232.555	-187.652	-424.815	286.963
Al					732.257	577.464	-326.544
K						2664.552	-292.354
Fe							2045.787

Table 4: Sample within-object variance-covariance matrix, $\hat{\mathbf{U}}$, and between-object variance-covariance matrix, $\hat{\mathbf{C}}$, for the log(- log ratio) transformation of the glass data. The values shown are the variances and covariances multiplied by 1000.

$\hat{\mathbf{U}}$	Si	Na	Ca	Mg	Al	K	Fe
Si	14.342	-0.077	3.156	0.142	0.372	0.809	0.299
Na		0.063	0.011	0.031	-0.017	-0.034	0.003
Ca			0.873	0.062	0.072	0.194	0.088
Mg				0.049	-0.032	-0.01	0.007
Al					0.228	0.044	0.009
K						0.393	0.021
Fe							0.106
$\hat{\mathbf{C}}$	Si	Na	Ca	Mg	Al	K	Fe
Si	10.841	2.074	8.437	4.429	-1.911	-4.091	1.516
Na		1.360	4.126	2.993	-0.458	-3.545	0.628
Ca			19.394	11.683	-2.247	-12.000	2.979
Mg				27.533	-5.983	-12.865	5.598
Al					13.236	11.402	-5.244
K						44.348	-5.215
Fe							29.046

Table 5: Sample within-object variance-covariance matrix, $\hat{\mathbf{U}}$, and between-object variance-covariance matrix, $\hat{\mathbf{C}}$, for the spherical transformation of the glass data. The values shown are the variances and covariances multiplied by 1000.

$\hat{\mathbf{U}}$	Si	Na	Ca	Mg	Al	K	Fe
Si	0.211	-0.043	0.287	0.016	0.025	0.054	0.016
Na		0.039	-0.054	0.011	-0.014	-0.029	-0.002
Ca			0.457	0.030	0.026	0.057	0.023
Mg				0.020	-0.011	-0.009	0.002
Al					0.061	0.016	0.002
K						0.071	0.003
Fe							0.012
$\hat{\mathbf{C}}$	Si	Na	Ca	Mg	Al	K	Fe
Si	0.312	0.154	0.663	0.269	-0.093	-0.295	0.063
Na		0.615	1.128	0.664	-0.095	-1.157	0.070
Ca			4.125	1.350	-0.476	-2.836	0.284
Mg				4.210	-1.231	-2.287	0.495
Al					1.352	1.233	-0.312
K						4.339	-0.374
Fe							1.213

Table 6: Means of simulation results for 320 comparisons within groups for the estimation of false negatives (FN) and for 12,640 comparisons between groups for the estimation of false positives (FP). FP and FN rates for each model (normal, KDE1, KDE2) and data transformation considered. KDE1: kernel density estimation with bandwidth matrix of the form $\mathbf{H} = h^2\mathbf{C}$ and h obtained using expression (12), KDE2: kernel density estimation with unconstrained bandwidth matrix $\mathbf{H}$ estimated using cross-validation. Full: seven-dimensional densities, GM1: factorisation based on graphical model using PC algorithm, GM2: factorisation based on simplified graphical model.

Model	Data Transformation	Error Type	Factorisation Full	GM1	GM2
Normal	1. logratio	FP	4.9%	5.8%	5.2%
		FN	3.4%	2.8%	3.1%
	2. log(-log ratio)	FP	3.9%	4.4%	4.5%
		FN	3.8%	3.8%	3.8%
	3. spherical	FP	2.6%	3.0%	3.0%
		FN	4.7%	4.7%	4.4%
KDE1	1. logratio	FP	4.1%	4.6%	4.2%
		FN	3.4%	3.1%	3.8%
	2. log(-log ratio)	FP	3.3%	3.6%	3.6%
		FN	4.4%	4.7%	5.0%
	3. spherical	FP	2.2%	2.4%	2.4%
		FN	5.0%	5.3%	5.3%
KDE2	1. logratio	FP	3.9%	3.6%	3.2%
		FN	3.4%	4.1%	4.7%
	2. log(-log ratio)	FP	3.0%	2.8%	2.6%
		FN	4.1%	5.3%	5.6%
	3. spherical	FP	2.0%	2.0%	1.9%
		FN	5.9%	5.6%	5.3%

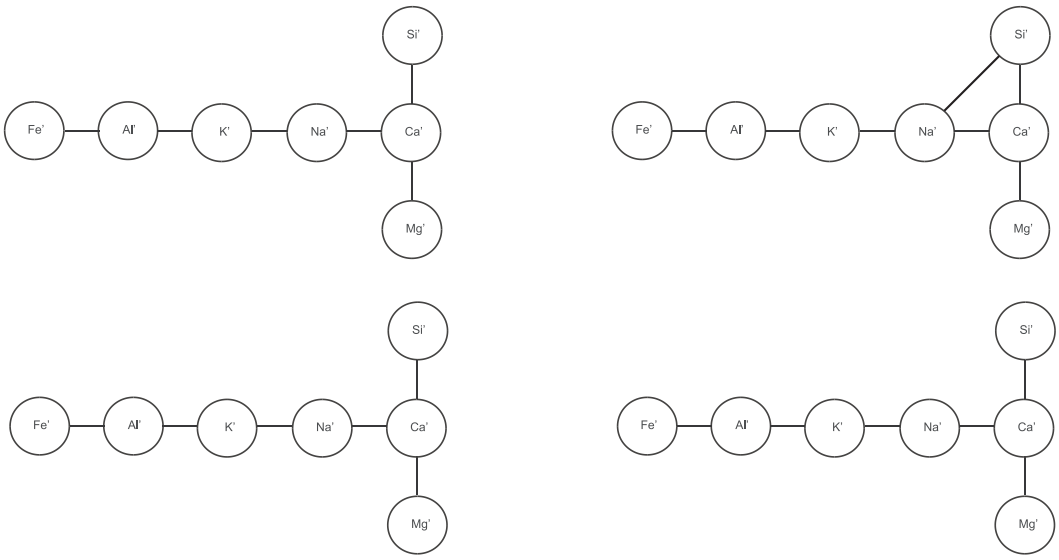

Figure 1: PC algorithm-generated graphical models shown clockwise for each of the training sets in the four-fold cross-validation procedure applied to the logratio-transformed data.

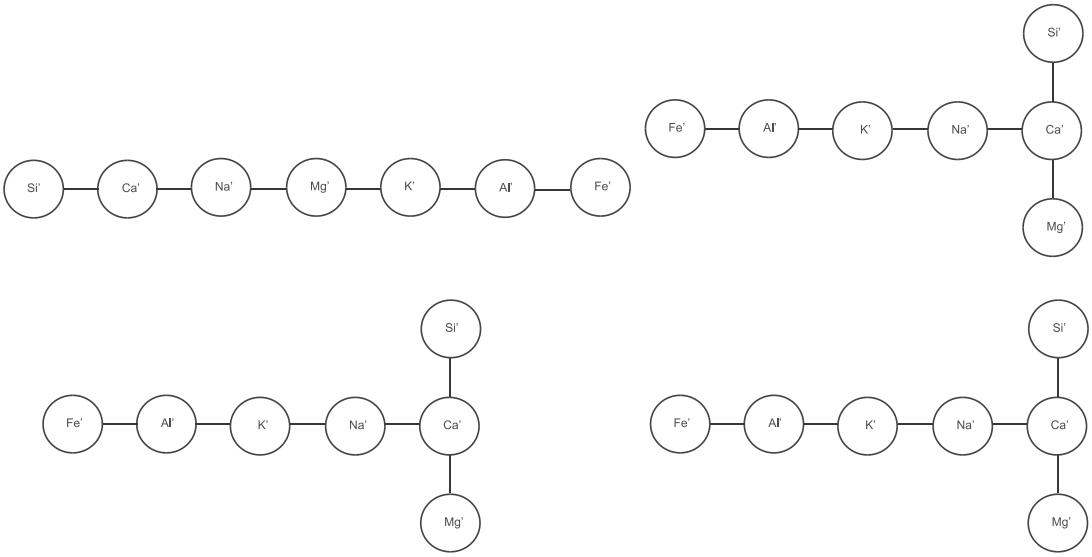

Figure 2: PC algorithm-generated graphical models shown clockwise for each of the training sets in the four-fold cross-validation procedure applied to the complementary log-log transformed data.

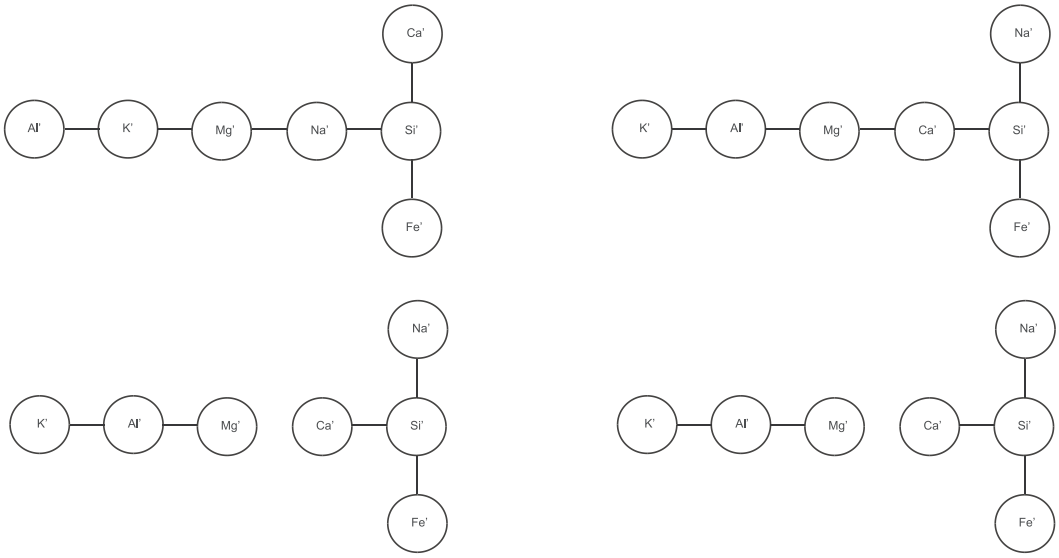

Figure 3: PC algorithm-generated graphical models shown clockwisefor each of the training sets in the four-fold cross-validation procedure applied to the spherically transformed data.

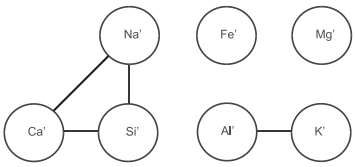

Figure 4: Simplified graphical model selected based on the partial correlation matrix for the logratio and complementary log-log transformed data.

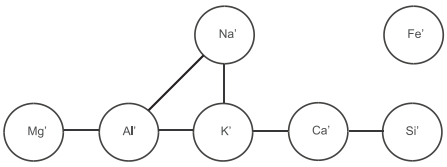

Figure 5: Simplified graphical model selected based on the partial correlation matrix for the spherically transformed data.