SCI118: A multivariate approach to the grouping problem in forensic glass analysis

Summer Research 2022-2023 Student Report
Faculty of Science
School of Statistics

Toby Hayward
Bachelor of Science: Data Science

Supervised by:
Prof. James Curran

Career Comments

This project improved my abilities in a lot of areas in both Statistics and Computer Science. Most notably are the CS aspects that forced me out of my comfort zone in terms of programming languages, development procedures, and file concurrency for IT collaboration.
In particular, the project at its core is centred around the idea of building an **R Package** which I have never had experience doing before. In doing so, I learnt a lot about how packages are stored, tested and developed, and also about utilising more efficient languages like **C++** within the package for faster algorithms. C++ is a language I have never had experience with before, and I had a blast learning more about it, and getting to witness its computational benefits over programming in R. In order to work with James on the project, we utilised another service I have little experience in: **GitHub**. GitHub is a file concurrency manager which works with a localised program *Git* in order to allow programmers to work on coding projects together without constantly overwriting each others code.
Among other aspects; such as, a more robust understanding of statistical tests for between groups analysis and working with graphics and objects within R, this project gave me a larger skill set to draw from when going forward in my career in Data Science. I'm very grateful for both James's expert supervision and knowledge of the field, and for the privilege to exercise old and *new* skills in Computer Science and Statistics. I feel like I have had the opportunity to bring forward my own ideas to the table at every step, and I am very proud of the progress that has been made in project.

Project Summary

In forensics, it is common and effective practice to analyse glass fragments from the scene and suspects to gain evidence of placing a suspect at the crime scene. This kind of analysis involves comparing the physical and chemical attributes of glass fragments that exist on both the person and at the crime scene, and assessing the significance of any likeness that they share.
However, it is often the case that only a very small sample of glass fragments can be recovered; which limits both the strength and variety of analysis and hypothesis testing that can be conducted.
Professor James Curran and his associates conducted work into developing ideas to strengthen the approach towards comparing, and more specifically grouping fragments together based on particularly the glass's *Refractive Index* and *chemical compound concentration.* They theorised and tested their algorithms against conventional methods in forensic glass analysis and appeared successful in their approach. (Curran et al. 1997; Triggs et al. 1997)

This project aims to build on these ideas and develop a publicly available **R package** downloadable from the *Comprehensive R Archive Network (CRAN)* which incorporates their most successful algorithm for discriminating and grouping glass fragments based on their refractive index: The *Scott-Knott Modification 2 (SKM2).* We are also interested in extending the *Scott-Knott* idea to a more modern and rigorous *multivariate* context; which utilises multiple variable data from the glass fragments to discern and group glass fragments in the same context. This algorithm will also be made available for use in the R package labelled SK4FGA.

Abstract

From the Summer Research Scholarship - Science - Statistics page of projects:

When examining a sample of glass fragments recovered from a suspect in a forensic case, the question arises whether the fragments may come from several different sources. An existing method uses a single variable, the refractive index, to estimate the total number of sources and to group the fragments by their sources. The aim of this project is to extend the method to accommodate multivariate evidence data.

Conventional grouping methods, when applied to forensic glass analysis using the samples' *refractive index*; i.e. the method to determine whether two groups of glass samples are to be considered significantly alike enough to may as well have been from the same source, often violate the population size assumption which results in theoretical quantiles usually not being appropriate. Prof. Christopher Triggs and Prof. James Curran devised a more appropriate solution by re purposing the *Scott-Knott* algorithm for this particular problem and prove their modification to be more effective than other current methods (Triggs et al. 1997).
However with developments in glass manufacturing causing the variance in refractive indices to become narrower, has made this method more and more ineffective. It has been noted that the chemical composition of glass can provide further discrimination between samples (Curran et al. 1997), and thus arises a demand for algorithms which utilise this information for more robust forensic glass analysis.
This report outlines the success of developing an idea for this problem which incorporates the ideas outlined in James's thesis, and the development of an **R package** which makes the algorithms widely available and open source.

Aims

1. Explore the idea of extending the *Scott-Knott* algorithm to a more rigorous and data rich **multivariate** analysis.
2. Develop an open source and publicly available *R package* which encompasses useful functions and tools for performing forensic glass analysis in moth both the Univariate and Multivariate case.

```
library(SK4FGA)
```

Method

We begin by taking the existing **Scott-Knott** algorithm, and reforming it for a *forensic glass analysis* context as outlined in (Triggs et al. 1997).

The *original* Scott-Knott algorithm (SK)

For an array of values $y_1,...,y_k$ which are considered measurements of a single feature pertaining to a observation

1. Sort $y_1,...,y_k$ into ascending order $y_{(1)},...,y_{(k)}$.
2. Calculate $\hat{\sigma}_0^2$.
3. Find B_0 and record the position of which it was found (say j)
4. Calculate λ and ν_0
5. If $\lambda > \chi^2_{\nu_0}(\alpha)$ then split $y_{(1)},...,y_{(k)}$ into $\{y_{(1)},...,y_{(j)}\}$ and $\{y_{(j+1)},...,y_{(k)}\}$
6. If there was a split in step 5, repeat steps **2-5** for each new subgroup until there are no more splits.

With,

- $\hat{\sigma}_0^2$ = maximum likelihood estimate of σ^2 under the hypothesis that the means of the observations are equal.
- B_0 = the maximum *between groups sum of squares* when procedurally splitting the array into two subgroups $\{y_{(1)},...,y_{(j)}\}$ and $\{y_{(j+1)},...,y_{(k)}\}$. We determine the between groups sum of squares for all pairs of groups in the array for $1 \leq j \leq (k-1)$.
- $\lambda = \frac{\pi}{2(\pi-2)} \frac{B_0}{\hat{\sigma}_0^2}$; $\nu_0 = \frac{k}{\pi-2}$

(Triggs et al. 1997)

James's Alteration to the Scott-Knott algorithm (SKM2)

Since our analysis is specific to the usage of a glass fragment's *refractive index*, we can use an independently sampled sample variance for glass fragments $s_0 = 4 \times 10^{-5}$, which is chosen in accordance with both James's own analysis of New Zealand Case data, and that found in a another study conducted in the UK (Evett and Lambert 1982).
A similar issue arises when we try to approximate the distribution of λ with a Chi-Squared distribution. Due to insufficient data to assume such a distribution for λ when k is small (which is not unusual for

forensic glass analysis), we can test against an empirically estimated distribution for λ instead. Hence, James's modification to the Scott-Knott algorithm comes to:

1. Sort $y_1,\ldots,y_k$ into ascending order $y_{(1)},\ldots,y_{(k)}$.
2. Find B_0 and record the position of which it was found (say j)
3. Calculate λ
4. If $\lambda > \lambda(\alpha : k)$, then split $y_{(1)},\ldots,y_{(k)}$ into $\{y_{(1)},\ldots,y_{(j)}\}$ and $\{y_{(j+1)},\ldots,y_{(k)}\}$
5. If there was a split in step 4, repeat steps **2-4** for each new subgroup until there are no more splits.

With,

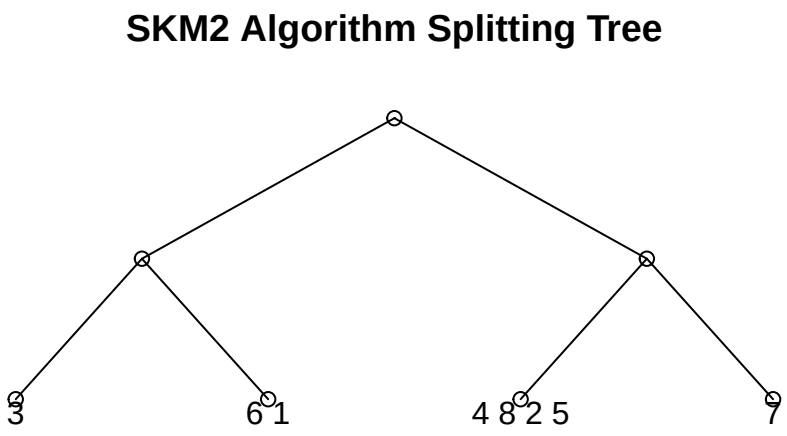
- $\lambda(\alpha : k)$ = the $100(1-\alpha)\%$ quantile of the empirically estimated distribution of λ for an array of size k.

(Triggs et al. 1997)

```
set.seed(20)
ri = generate_indices(8, 4); names(ri) = 1:8; ri
```

```
       1        2        3        4        5        6        7        8
1.518186 1.517906 1.518286 1.517787 1.517929 1.518091 1.517538 1.517861
```

```
plot(partition(ri))
```



[1] An example array of refractive indices and an example output of the *SKM2* algorithm.

A repurposed, multivariate analog to the Scott-Knott (SKmulti)

The real interest of the study was extending this idea to the multivariate case by including the glass chemical composition data. However this breeds a couple challenges to overcome:

- How to compare two *n*-dimensional glass fragment observations.
- How to compare *multiple n*-dimensional glass fragment observations in a *SK* type fashion.

At least for the first point, James has already identified a sensible hypothesis test: *Hotellings* T^2; the multivariate analog to the *Students t-test.* (Campbell and Curran 2009)

Used in place of λ, we can test the T^2 statistic against the T^2 distribution to determine significance between groups and develop a *partition* tree in the same fashion.
Although a consideration must be made to a particular property of the Scott-Knott; the observations must be arranged in some order such that those which would be of statistically significant proximity to each other will be side by side in the array of data points.
Ordering points that effectively exist in *n*-dimensional space isn't as simple as ordering scalar values since the same transitive properties of size are not maintained. Thus a simple solution is to order the points by *Euclidean distance* to the *mean vector.* The effectiveness of this approach are to be tested.
Therefore, our multivariate analog of the Scott-Knott algorithm becomes:

1. Sort $Y_1,\ldots,Y_k$ into ascending order $Y_{(1)},\ldots,Y_{(k)}$ according to Euclidean distance to the array's mean vector.
2. Find T_0 and record the position of which it was found (say j)
3. If $T_0 > T^2_{j*,(k-j)*,n}(\alpha)$, then split $Y_{(1)},\ldots,Y_{(k)}$ into $\{Y_{(1)},\ldots,Y_{(j)}\}$ and $\{Y_{(j+1)},\ldots,Y_{(k)}\}$
4. If there was a split in step 4, repeat steps **2-4** for each new subgroup until there are no more splits.

With,

- T_0 = the maximum T^2 statistic when procedurally splitting the array into two subgroups $\{y_{(1)},\ldots,y_{(j)}\}$ and $\{y_{(j+1)},\ldots,y_{(k)}\}$ and binding the data into a single matrix. We determine T^2 for all pairs of groups in the array for $1 \leq j \leq (k-1)$.
- $T^2_{j*,(k-j)*,n}(\alpha)$ = The T^2 distribution quantile estimated at the α tail end.
- $j* = \sum_{i=1}^{j} rows_i$; n = number of variables for each observation.

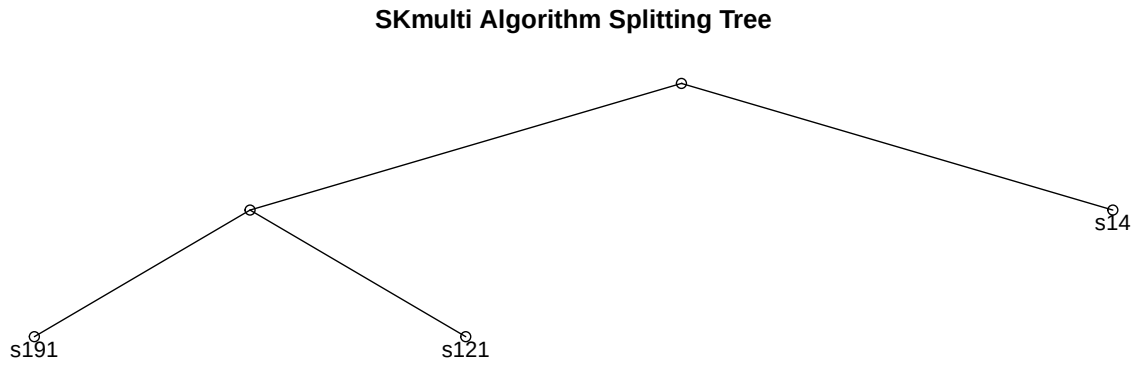
An all encompassing *R package* to the benefit of all

The development and implementation of these algorithms was written in a combination of **R** and **C++**. An *R package* makes this process simple, reproducible and well documented, but also just as importantly it allows communication and collaboration between the two programming languages. C++ is useful for its computational efficiency, and hence all of the computationally complex code will be written in that language. R proves to be sufficiently fast in most other aspects of the algorithm and therefore the majority of the package is developed in R.

```
(glass.data.view = glass[glass$item %in% sample(unique(glass$item), 3) &
                         glass$fragment %in% c('f1', 'f2'),])
```

	item	fragment	logNaO	logMgO	logAlO	logSiO	logKO	logCaO	logFeO
157	s14	f1	-0.6009	-1.3485	-1.6785	-0.1019	-2.6974	-1.0305	-5.6517
158	s14	f1	-0.6009	-1.3494	-1.6707	-0.0853	-2.5299	-1.0024	-5.3428
159	s14	f1	-0.6034	-1.3478	-1.6580	-0.0591	-2.4270	-0.9508	-5.6311
160	s14	f2	-0.6023	-1.3501	-1.6799	-0.1143	-2.7034	-1.0534	-5.6576
161	s14	f2	-0.6096	-1.3697	-1.6957	-0.1277	-2.7611	-1.0698	-5.3631
162	s14	f2	-0.6061	-1.3688	-1.6764	-0.1455	-2.7689	-1.1145	-5.6720
1441	s121	f1	-0.6584	-1.1721	-2.7403	-0.0814	-2.4972	-0.8524	-5.6434
1442	s121	f1	-0.6571	-1.1687	-5.6430	-0.0783	-2.3642	-0.8492	-5.3419
1443	s121	f1	-0.6640	-1.1669	-5.6440	-0.0808	-2.3888	-0.8524	-5.6440
1444	s121	f2	-0.6664	-1.1656	-2.6113	-0.0976	-2.5388	-0.8782	-5.6527
1445	s121	f2	-0.6664	-1.1656	-2.6113	-0.0976	-2.5388	-0.8782	-5.3517
1446	s121	f2	-0.6670	-1.1785	-5.6542	-0.0993	-2.4781	-0.8819	-5.6542
2281	s191	f1	-0.7151	-1.3393	-2.2001	-0.1704	-5.3904	-0.9402	-5.6914
2282	s191	f1	-0.7167	-1.3508	-2.3354	-0.1834	-5.6971	-0.9505	-5.3961
2283	s191	f1	-0.7164	-1.3694	-2.3714	-0.1739	-5.6936	-0.9370	-5.6936
2284	s191	f2	-0.7232	-1.3591	-2.2464	-0.1740	-5.3925	-0.9324	-5.6936
2285	s191	f2	-0.7232	-1.3725	-2.2875	-0.2357	-5.7188	-1.0340	-5.4178
2286	s191	f2	-0.7234	-1.3616	-2.3100	-0.2079	-5.7079	-0.9911	-5.7079

```
plot(partition.multi(prepare_data(glass.data.view, 1)))
```

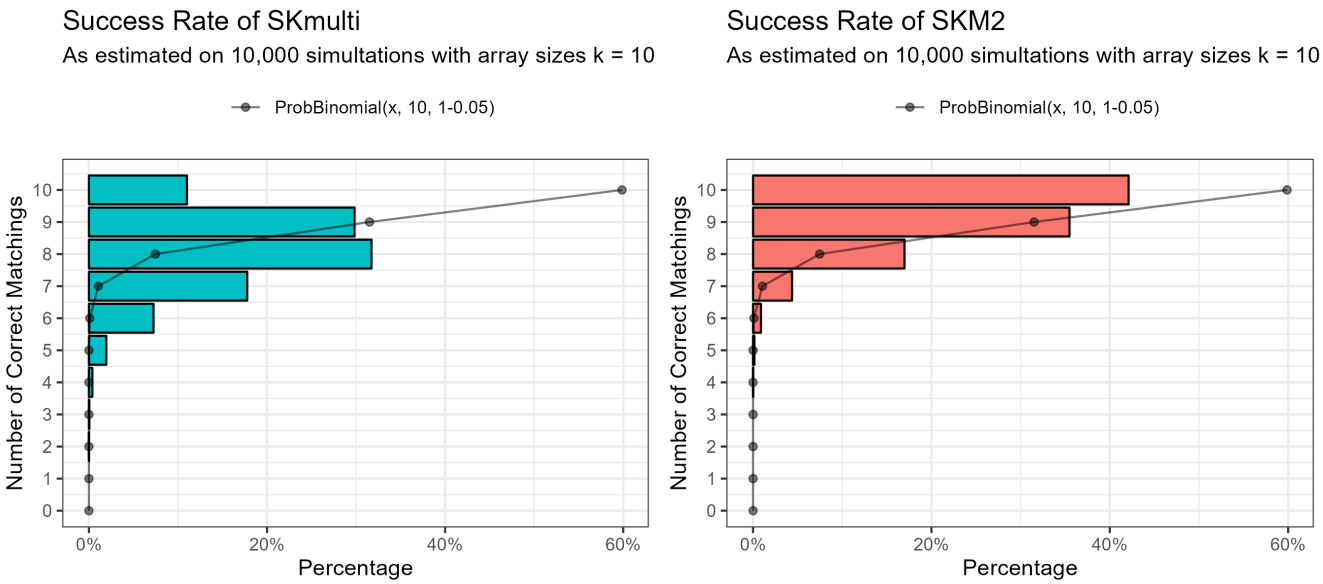


[2] An example SEM-EDX measured dataset of glass fragments recovered from various items; measured and supplied by Grzegorz (Greg) Zadora at the Institute of Forensic Research in Krakow, Poland. And the output from the SKmulti algorithm.

Results

Addressing the first aim of the project to do with the success of the multivariate algorithm, a glaring issue and oversight with the idea arose with *how to order n-dimensional objects* such that alike objects will be next to each other in the array. A natural choice was to order the array by each objects *Euclidean distance to the mean vector*, and was the decision of ours to proceed with this approach.
A simple test of efficacy was to take a sample of k objects, sampled from the `glass` dataset (provided by Grzegorz Zadora) and randomly generate k more corresponding objects based on the covariance matrices and mean vectors of each object from the random selection of objects. The idea here is that each randomly generated object should be grouped alongside its corresponding "inspiration", and by counting the number of successful pairings we can get an idea of how well the algorithm can pair a large array of objects with corresponding "matches" when convoluted with a bunch of random other objects.
Under the null hypothesis and with a perfectly working algorithm, we should expect to see the distribution of correct matchings approach a *binomial distribution* as the number of simulations approaches ∞:

$$CorrectMatchings \sim Binomial(n = k, p = (1 - \alpha))$$



[3] Performance of algorithms **SKmulti** and **SKM2** on arrays of size 10 at a threshold of $\alpha = 0.05$. We hope to witness a percentage of perfect matchings to be around $(1 - 0.05)^{10} \times 100 \approx 60\%$.

Shown above, the *SKM2* algorithm appears to represent roughly what we were expecting, but the *SKmulti* algorithm seems to fail this assumption. We can only assume that the reason for this is due to the algorithm failing to order the objects correctly in the array. However, we can imagine that this method is producing better results than letting the array remain unordered, and hence it remains that this is at least a somewhat working method at producing correct matchings.

The R package

As of now, the R Package is available for download from **CRAN** and can be loaded and installed using the following command.

```
install.packages("SK4FGA")
```

The package includes a range of useful functions for forensic glass analysis. The most important and main ones include:

SK4FGA::	Description
partition	Completes the *SKM2* algorithm on a vector of glass Refractive Indices.
partition.multi	Completes the *SKmulti* algorithm on a list of same format data.frames containing glass chemical composition data.
plot.sk_partition_tree	For plotting the outputs from `partition` and `partition.multi`.
generate_indices	Function for quickly generating an array of Refractive indices.
prepare_data	Processes a data.frame of items and their respective chemical composition data into a list object that `partition.multi` can understand.

Amongst other functions that are mostly just useful from within the package (i.e. they're used by the functions above), there is also three data files that come included in the package also:

SK4FGA::	Description
glass	200 Glass samples with 7 corresponding chemical concentration oxygen ratios.
glass2	A limited dataset containing glass work on 15 variables.
vehicle.glass	A comprehensive vehicle glass dataset with 761 samples on 15 variables.

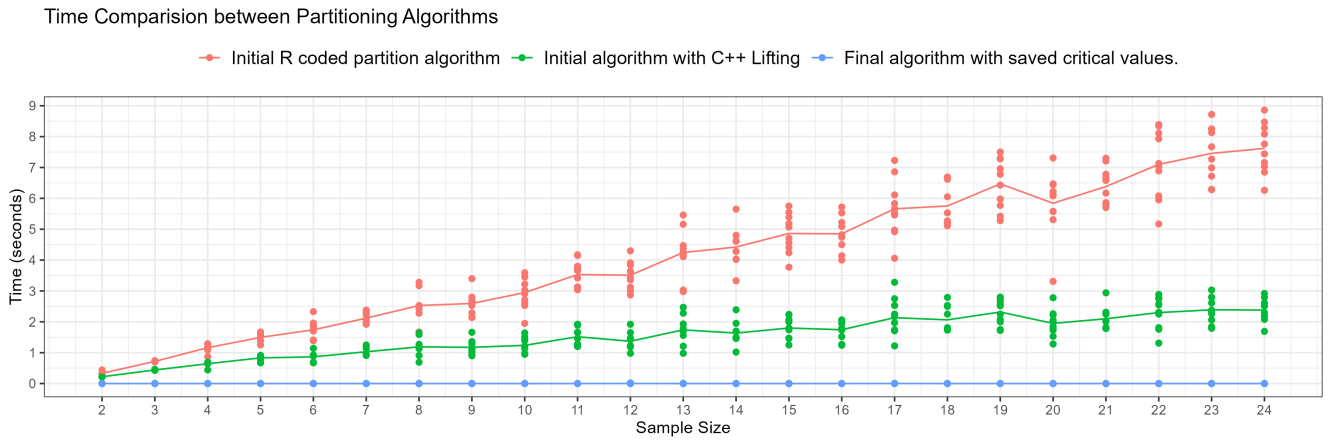
Discussion

SKmulti

As mentioned, the multivariate analog (SKmulti) is problematic; and it remains an open question about how to properly order the array of objects' data. For now I am satisfied with its performance and don't intend on improving the matter myself.
Another issue with the algorithm to tackle is with assuming a T^2 distribution. Much like the issue with the Scott-Knott, there is usually not enough data to safely warrant an assumption that the T^2 statistic came from exactly the T^2 distribution under the null hypothesis. The distribution is likely slightly different; skewed, or chaotic in shape that perhaps a more sensible solution would be to empirically estimate the distribution in the same vein as *SKM2*.

Computational Efficiency

Especially when coming to computational efficiency, came a calling to utilising the **C++** computational benefits in the package. `partition` was drafted and coded multiple times with heavy lifting code blocks being swapped for the C++ equivalents. Over the course of the project, the speed of the algorithm became faster and faster until it became negligible; especially with the decision to maintain an object within the package that saved critical values for the distribution of λ to be called upon and use the chi-squared distribution for all array sizes above 20 (just like in the original Scott-Knott).



[4] Speed performance of the same algorithm rewritten over the course of the project.

References

Campbell, Gareth P., and James M. Curran. 2009. "The Interpretation of Elemental Composition Measurements from Forensic Glass Evidence III." *Science & Justice* 49 (1): 2–7. https://doi.org/https://doi.org/10.1016/j.scijus.2008.09.001.

Curran, J. M., C. M. Triggs, J. R. Almirall, J. S. Buckleton, and K. A. J. Walsh. 1997. "The Interpretation of Elemental Composition Measurements from Forensic Glass Evidence: i." *Science & Justice* 37 (4): 241–44. https://doi.org/https://doi.org/10.1016/S1355-0306(97)72197-X.

Evett, I. W., and J. A. Lambert. 1982. "The Interpretation of Refractive Index Measurements. III." *Forensic Science International* 20 (3): 237–45. https://doi.org/https://doi.org/10.1016/0379-0738(82)90123-2.

Triggs, Christopher M., James M. Curran, John S. Buckleton, and Kevan A. J. Walsh. 1997. "The Grouping Problem in Forensic Glass Analysis: A Divisive approach1This Research Was Made Possible by a Ph.d. Scholarship from ESR: Forensic.1." *Forensic Science International* 85 (1): 1–14. https://doi.org/https://doi.org/10.1016/S0379-0738(96)02037-3.

Thanks

Lewis Kendall-Jones For supporting the development of the package by overseeing the initial C++ code.