The author(s) shown below used Federal funding provided by the U.S. Department of Justice to prepare the following resource:

Document Title: Detection and Characterization of Subpopulations and the Study of Algorithmic Bias in Forensic Identification of Source Problems

Author(s): Semhar Michael, Andrew Simpson, Dylan Borchert, Christopher Saunders, Larry Tang

Document Number: 308220

Date Received: December 2023

Award Number: N/A

This resource has not been published by the U.S. Department of Justice. This resource is being made publicly available through the Office of Justice Programs' National Criminal Justice Reference Service.

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Detection and Characterization of Subpopulations and the Study of Algorithmic Bias in Forensic Identification of Source Problems

Semhar Michael*, Andrew Simpson, Dylan Borchert, Christopher Saunders
Mathematics and Statistics Department
South Dakota State University
*Email: semhar.michael@sdstate.edu

Larry Tang
Department of Statistics and Data Science and National Center for Forensic Science
University of Central Florida, Orlando

A report on a presentation given at the International Conference on Forensic Inference and Statistics (ICFIS)
June 13, 2023
Lund, Sweden

This resource was prepared by the author(s) using Federal funds provided by the U.S.
Department of Justice. Opinions or points of view expressed are those of the author(s) and do not
necessarily reflect the official position or policies of the U.S. Department of Justice.

Contents

List of Figures

List of Tables

Detection and Characterization of Subpopulations and the Study of Algorithmic Bias in Forensic Identification of Source Problems

***Abstract*—The forensic source identification problem involves providing the summary of the forensic evidence to a decision-maker via the value of that evidence. This can be done via the forensic likelihood ratio which in turn requires modeling of a relevant background population. Some of the commonly used methods involve the assumption of normality. However, there might exist a latent variable representing an underlying subpopulation structure. In this work, we will focus on identifying and characterizing subpopulations in the relevant population when there are hierarchically structured data. This will be done through semi-supervised finite mixture models that are adjusted for the hierarchical sampling procedure. In addition, we will study systematic algorithmic biases that can occur as measured by rates of misleading evidence for each of the subpopulations when the subpopulation structure is not accounted for. We will illustrate this based on a simulation study using synthetic data and classical glass datasets.**

I. Introduction

In forensic science, the source identification problem involves providing the summary of the forensic evidence to a decision-maker about the origins of the given evidence. One way of providing a summary is through the value of that evidence. For glass trace evidence we can have the following questions

- Are the glass fragments found on the suspect from the same source as the specific crime scene window? *A specific source (window) problem.*
- Are the glass fragments found on these two different suspects from the same unknown crime scene window? *A common but unknown source (window) problem.*

For both problems, one common method for making inferences is through the likelihood ratio approaches (Bayes factor (BF) and likelihood ratio (LR)) [Aitken et al., 2007, Ommen and Saunders, 2021].

Generally, there are two competing propositions. For the common but unknown source problem, these propositions are H_p : E_{U_1} and E_{U_2} were generated from a common but unknown source from the relevant source population E_A and H_d : E_{U_1} and E_{U_2} were generated from two randomly selected sources from the relevant source population, E_A. These two hypotheses are commonly referred to as the prosecution's and the defense's proposition, respectively. Let M_d and M_p denote the models associated with the two competing propositions. Then the question is: "What is the likelihood of observing the evidence under the prosecution model? vs. What is the likelihood of observing the evidence under the defense model?" The LR can be written as

$$LR = \frac{f(e|M_p,\theta)}{f(e|M_d,\theta)},$$

where $e = \{e_{u_1}, e_{u_2}, e_a\}$ is the observed evidence.

Colin G. G. Aitken and David Lucy [2004] proposed using two types of LRs. The first is using the multivariate normal-based LR where we assume multivariate normality for both the between- and within-source distributions. The second is LR based on modeling the between-source distribution with a kernel density estimate applied to E_A. These approaches are implemented in the *comparison* R-package [Lucy et al., 2020]. Our goal is to find a more efficient model for the between-source distribution to be used in the calculation of the LR.

A Gaussian random effects model can be used to describe how the evidence arises. Let $\boldsymbol{X}_{ij}$ for $i = 1, ..., m$ and $j = 1, .., n_i$ be a $p \times 1$ vector containing the features of the j^{th} sample from the i^{th} source, where $\boldsymbol{X}^{mn}$ is the set of all samples in the background population. Assuming Gaussian within- and between-source distribution, a random effects model is given as $\boldsymbol{X}_{ij} = \boldsymbol{\mu} + \boldsymbol{a}_i + \boldsymbol{\epsilon}_{ij}$, where $\boldsymbol{\mu}$ is the overall mean of the sources, $\boldsymbol{a}_i \sim MVN(0, \boldsymbol{\Sigma}_a)$ describes the between source distribution, and $\boldsymbol{\epsilon}_{ij} \sim MVN(0, \boldsymbol{\Sigma}_\epsilon)$ describes the within-source distribution. We will explore the following sampling experiment for the jth sample from the ith source when a subpopulation structure is present in the relevant source population. Let $\boldsymbol{X}_{ij} = \boldsymbol{a}_i + \boldsymbol{\epsilon}_{ij}$, where $z_i \sim Multinoulli(\tau_1, ..., \tau_K)$, $\boldsymbol{a}_i|z_i = k \sim MVN(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, and $\boldsymbol{\epsilon}_{ij} \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma}_\epsilon)$, for $k = 1, \dots, K$ subpopulations with mixing proportions $\tau_1, \dots, \tau_K$, respectively.

II. Algorithmic bias

A. Rates of misleading evidence

For some cutoff c (usually $c = 1$), $LR > c$ suggests that the prosecution hypothesis is more likely, and $LR < c$ suggests that the defense hypothesis is more likely. Then the rate of misleading evidence in favor of the prosecution (RMEP) is defined as the proportion of different source comparisons with LR values greater than $\boldsymbol{c}$ and the rates of misleading evidence in favor of the defense (RMED) is the proportion of same source comparisons that resulted in LR values less than $\boldsymbol{c}$. The bias we are concerned about is the rate of misleading evidence in favor of the prosecution (RMEP). We will focus on the variation of this RMEP among subpopulations of differing sizes.

This resource was prepared by the author(s) using Federal funds provided by the U.S.
Department of Justice. Opinions or points of view expressed are those of the author(s) and do not
necessarily reflect the official position or policies of the U.S. Department of Justice.

B. *Study of algorithmic bias*

We introduce the following notation. Let $S_1 \subset \{1,\dots,n\}$ be the set of indices of objects that belong to the majority subpopulation. Similarly, let $S_2 \subset \{1,\dots,n\}$ be the set of indices of objects that belong to the minority subpopulation. Define $S_{k,k'} = \{(i,i') : i \in S_k, i' \in S_{k'}, i \neq i'\}$ for $k, k' \in \{1,2\}$ and the comparison $LR_{i,i'} = LR(e_{u_i}, e_{s_{i'}}|e_a)$. The algorithm used to study algorithmic bias represented by rates of misleading evidence in favor of the prosecution is given below. To study the effects of the size of subpopulations, we simulated a two-component mixture varying the mixing proportions representing the sizes of the subpoulations with $\tau \in \{.1, .15, .2, .25\}$. The following algorithm describes the steps.

Algorithm 1 Simulation

```
for $K = 2$, $\tau \in \{.1, .15, .2, .25\}$ do
    Generate $m = 200$ source means from the source level mixture model with $\pi$ as the mixing proportion for the minority population [Melnykov et al., 2012].
    Generate $n_i = 10, \forall i = 1,\dots,m$ observations from each within source distribution.
    Split the data into train and test sets, so that each set contains an equal number of sources preserving $\pi$. Let the train set be $e_A$.
    Using the test set, label the first 5 observations as the trace and the last 5 observations as the control for each source.
    Do all pairwise comparisons of traces and controls for the sources in the test set using a plug-in estimate of LR using $e_A$ as the relevant population.
    Calculate RMEP.
end for
Repeat $B = 100$ times
```

For the case of 2 subpopulations, there are four cases for the RMEP relative to subpopulations. Consider the scenario where the trace objects (e_u) come from the subpopulation of interest (All different source comparisons) and the control objects could come from both the majority and minority subpopulations. Then for the minority population, the RMEP is given as

$$RMEP_{1B} = \frac{1}{|S_{1,1} \cup S_{1,2}|} \sum_{(i,i') \in S_{1,1} \cup S_{1,2}} I(LR_{i,i'} > c) \quad (1)$$

Similarly, for the majority population, the RMEP is

$$RMEP_{2B} = \frac{1}{|S_{2,1} \cup S_{2,2}|} \sum_{(i,i') \in S_{2,1} \cup S_{2,2}} I(LR_{i,i'} > c). \quad (2)$$

Now consider the scenario where both the control and trace objects come from the subpopulation of interest. In the case of the minority subpopulation we can compute the RMEP as

$$RMEP_{1W} = \frac{1}{|S_{1,1}|} \sum_{(i,i') \in S_{1,1}} I(LR_{i,i'} > c). \quad (3)$$

Finally, in the case of the majority population we have

$$RMEP_{2W} = \frac{1}{|S_{2,2}|} \sum_{(i,i') \in S_{2,2}} I(LR_{i,i'} > c). \quad (4)$$

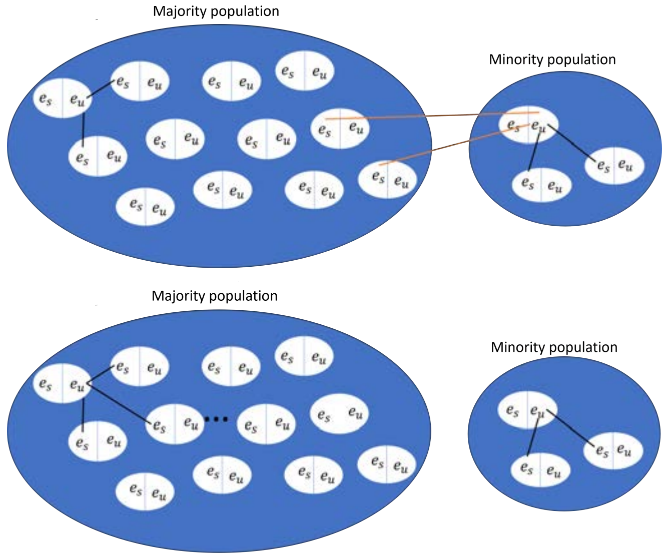

Fig. 1. Illustration of the pairwise source level comparisons. The top plot shows a case of the control being from both subpopulations and the bottom plot shows that the control source is taken from within the subpopulation only.

This difference is illustrated in Figure 1.

For each of the subpopulations and both cases between and within, we computed RMEP for $B = 100$ replicates and obtain mean and standard deviations. The results of the RMEPs computed in the four cases are shown in Table I. The column labeled "All" represents the RMEPs computed using Eq 1 and Eq 2 for each mixture. The column labeled "Subpopulation" indicates the RMEPs computed using Eq 3 and Eq 4. The overall error rates are presented in the last column. First, comparing the RMEP for the majority and minority population with all the pairwise comparisons, we see that the error rates of the minority population are almost 3 to 7 times more than the majority population. This becomes even more stark when we look at error rates within the subpopulations where the minority RMEP is 18 to 30 times more than the majority RMEP. This means that traces from different sources within the minority population will be incorrectly identified as being from the same source 18 to 30 times more than different sources from the majority population. In addition, moving from "All" to within subpopulation, we see that the majority RMEP increases slightly by about 2 percent; however, the minority RMEP increases from 4 to 10 times. The results of normal-based vs. kernel-based LRs can also be seen in Table II. The RMEPS based on the kernel-based LRs are very similar to the normal-based LRs and the error rate gaps between the minority and majority groups did not improve. The RMEPs improved slightly for the minority population, but the differences in error rates within the minority population were still 16 to 24 times more than the within majority RMEP.

Glass data-based simulation:

We computed RMEP for the 3-dimensional glass data with three subpopulations [Colin G. G. Aitken and David Lucy, 2004]. Similar results are obtained. The overall RMEP is about

This resource was prepared by the author(s) using Federal funds provided by the U.S.
Department of Justice. Opinions or points of view expressed are those of the author(s) and do not
necessarily reflect the official position or policies of the U.S. Department of Justice.

TABLE I. The mean (SD) of the rates of misleading evidence in favor of the prosecution (RMEP) from 100 replicates in simulated data using normal-based LRs.

			RMEP		
Mixture	k	π_k	All	Subpopulation	Overall
1	1	.25	0.023 (0.006)	0.094 (0.025)	0.008 (0.002)
	2	.75	0.003 (0.001)	0.005 (0.001)	
2	1	.2	0.020 (0.007)	0.103 (0.034)	0.007 (0.002)
	2	.8	0.004 (0.001)	0.005 (0.001)	
3	1	.15	0.014 (0.006)	0.102 (0.038)	0.005 (0.001)
	2	.85	0.004 (0.001)	0.005 (0.001)	
4	1	.1	0.011 (0.006)	0.125 (0.065)	0.005 (0.001)
	2	.9	0.004 (0.001)	0.004 (0.001)	

TABLE II. Rates of misleading evidence in favor of the prosecution from simulated data both normal-based and Kernel-based LRs.

		RMEP			
		all		subpopulation	
π_k	k	LR_1	LR_2	LR_1	LR_2
.1	1	0.011 (0.006)	0.009 (0.005)	0.125 (0.065)	0.096 (0.055)
.9	2	0.004 (0.001)	0.004 (0.001)	0.004 (0.001)	0.004 (0.001)
.15	1	0.014 (0.006)	0.012 (0.005)	0.102 (0.038)	0.084 (0.032)
.85	2	0.004 (0.001)	0.004 (0.001)	0.005 (0.001)	0.005 (0.001)
.2	1	0.020 (0.007)	0.017 (0.006)	0.103 (0.034)	0.087 (0.030)
.8	2	0.004 (0.001)	0.004 (0.001)	0.005 (0.001)	0.005 (0.001)
.25	1	0.023 (0.006)	0.019 (0.006)	0.094 (0.025)	0.080 (0.022)
.75	2	0.003 (0.001)	0.004 (0.001)	0.005 (0.001)	0.005 (0.001)

5% and subpopulation-specific RMEP ranged from 3 to 5% when all between source comparisons. However, when looking at the within-source comparisons the RMEP increased by 8% for the largest subpopulation and to 14% and 21% for the other two subpopulations of equal sizes.

Therefore, in all the simulations when comparing a trace to controls in the entire population the error rate in the majority subpopulation is not too different from the overall error rate. But when we decompose the overall error rate by subpopulation, we see that the minority subpopulation has higher error rates than the majority. This is even more apparent when traces and control comparisons are from within a subpopulation the RMEP is up to ten-fold in minority subpopulations than in the majority population. This in practice can lead to incorrect decisions being made at a higher rate in minority subpopulations. This is problematic since there can be undetected bias when we do not know there is a subpopulation structure in the relevant source population.

TABLE III. Rates of misleading evidence from 3-dimensional glass data with three labeled subpopulations.

		RMEP			
		all		subpopulation	
$\hat{\pi}_k$	k	LR_1	LR_2	LR_1	LR_2
—	1-3	0.0409	0.0398	—	—
.258	1	0.0333	0.0333	0.1429	0.1429
.258	2	0.0500	0.0458	0.2143	0.1964
.484	3	0.0400	0.0400	0.0857	0.0857

III. Relative efficiency

Currently, there are multiple versions of the likelihood ratio. For example the normal-based likelihood ratio, kernel density-based likelihood ratio Colin G. G. Aitken and David Lucy [2004], and mixture-based likelihood ratio Franco-Pedroso et al. [2016]. Since there are multiple versions of the likelihood ratio, how should one compare likelihood ratios on a given type of evidence to determine which performs better? We propose using a relative efficiency measure, which is the ratio of the variances of the estimated LRs to the true LR. Thus, if the relative efficiency is less than 1, the LR version that estimated the top of the ratio is more efficient, which means on average it is closer to the true LR value.

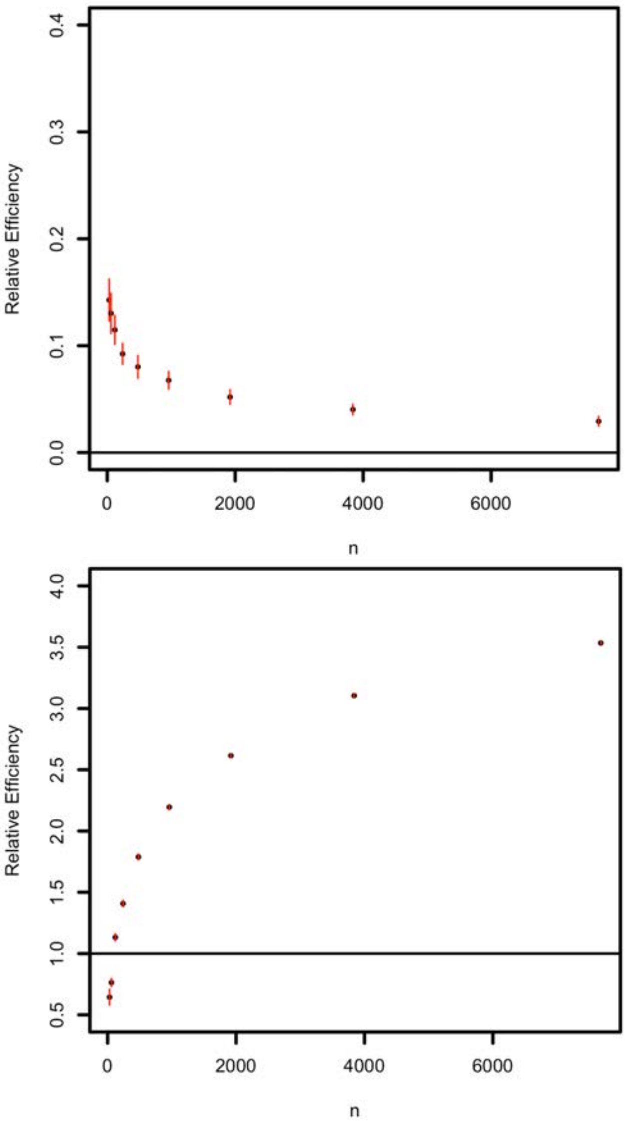

Fig. 2. Relative efficiency of normal-based likelihood ratio compared to KDE-based likelihood ratio. The top plot is for when E_A has a homogeneous population and the bottom plot is when there is a subpopulation structure.

Suppose we have two likelihood ratio methods for the common but unknown source identification problem denoted as $L_1(e^{(n_s)})$ and $L_2(e^{(n_s)})$ respectively, where $e^{(n_s)} = \{e_{u1}, e_{u2}, e_a^{(n_s)}\}$ and $e_a^{(n_s)}$ is the s^{th} randomly sampled background population consisting of n sources. Finally, let $L(e_{u1}, e_{u2})$ be the true likelihood ratio value. The relative

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

efficiency of the two likelihood ratio methods is defined as

$$\gamma_n(\hat{L}_1, \hat{L}_2) = \frac{\sum_{s=1}^{S} \left(L(e_{u1}, e_{u2}) - \hat{L}_1(e^{(n_s)}) \right)^2}{\sum_{s=1}^{S} \left(L(e_{u1}, e_{u2}) - \hat{L}_2(e^{(n_s)}) \right)^2}$$

$$= \sum_{s=1}^{S} \frac{\left(L(e_{u1}, e_{u2}) - \hat{L}_1(e^{(n_s)}) \right)^2}{\sum_{s'=1}^{S} \left(L(e_{u1}, e_{u2}) - \hat{L}_2(e^{(n_{s'})}) \right)^2}.$$

The asymptotic relative efficiency is,

$$\gamma(\hat{L}_1, \hat{L}_2) = \lim_{n \to \infty} \gamma_n(\hat{L}_1, \hat{L}_2).$$

Let $\hat{L}_1$ be the normal-based plug-in LR and $\hat{L}_2$ be the kernel-based plug-in LR (based on Colin G. G. Aitken and David Lucy [2004]). For the common but unknown source likelihood ratio, there are two sets of evidence with unknown sources, e_{u_1} and e_{u_2}. For this simulation, the prosecution model will be true which means e_{u_1} and e_{u_2} will be randomly sampled from the same source which in this case will be the mean of the first subpopulation. Both e_{u1} and e_{u2} consists of 5 samples.The relative efficiency $\gamma_n(L_1, L_2)$ is then calculated through simulation for $n = 30, 60, 120, 240, 480, 960, 1920, 3840, 7680$ where $S = 250$. Relative efficiency is calculated when only the first subpopulation, in which e_{u1} and e_{u2} reside, is in the background population. Hence, there is no subpopulation structure in the background population. The relative efficiency is then calculated when the second subpopulation is added to the background population along with the first subpopulation. Hence, there is a subpopulation structure in the background population. The results can be seen in Figure 2. The top plot shows the relative efficiency values as n increases for the case when there is a single subpopulation. It can be seen from this plot that the normal-based likelihood ratio is around 20 times more efficient than the kernel-based likelihood ratio in this case. This means that the normal-based likelihood ratio, on average, is 20 times closer to the true value of the likelihood ratio in terms of mean squared error compared to the kernel-based likelihood ratio. The bottom plot in Figure 2 shows the relative efficiency when there are two subpopulations. It can be seen that as n increases the kernel-based likelihood ratio is about 3.5 times for efficient than the normal-based likelihood ratio. This is because the kernel-based likelihood ratio is flexible and better able to model the two subpopulations. Interestingly, for small sample sizes (*i.e.*, $n = 30, 60$) the normal-based likelihood ratio is more efficient even though the assumptions of normally are known to be broken. This means that the kernel-based estimation method is failing to adequately model the population given low sampling sizes and is being outperformed by the normal-based method in terms of LR mean squared error even though the assumptions do not hold.

IV. Finite mixture model-based solutions

We will consider a glass data example [Aitken et al., 2007] where subpopulation structures are evident. Glass data are from the Institute of Forensic Research in Krakow, Poland [Aitken et al., 2007]. The data have seven elemental compositions and 2400 observations. There are 200 windows with four fragments measured three times. The rest of the variables represent the elemental compositions of the glass fragments: $log(NaO)$, $log(MgO)$, $log(AlO)$, $log(SiO)$, $log(KO)$, $log(CaO)$ $log(FeO)$ where $log(NaO) = log_{10}(\frac{Na}{O})$. See the pairwise scatter-plot of the data in Figure 3.

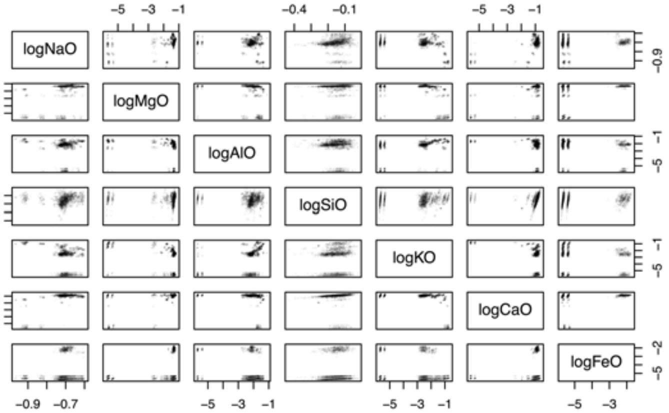

Fig. 3. Pairwise scatterplot of the observations from the glass dataset (Zadora's)

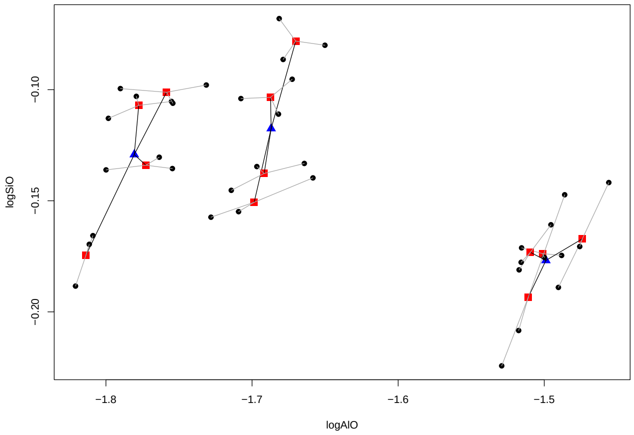

Fig. 4. Three windows from the glass dataset (Zadora's) where the blue triangles are the window (source) means, the red triangles are the fragment (trace) means are the black circles are the measurements (technical replicates)

In addition to subpopulations, forensic evidence often arises from a hierarchical sampling process. *i.e.*

1) The source of the fragments is first sampled from the alternative source population with a between-source distribution.
2) Then from that source, the fragments are sampled from that specific within-source distribution.
3) Finally, the measurement (elemental composition) is

taken from the fragment.

4) Further, there could be technical replicates from each fragment.

This sampling process can be seen in Figure 4 where we see three window means with lines to means of fragment replicates indicated by lines.

The research problem is to develop a model that can account for both heterogeneity and hierarchical structures present in data. Previous approaches used in the literature assume Gaussian mixture models for modeling heterogeneity but do not account for hierarchical structure [Dettman et al., 2014, Franco-Pedroso et al., 2016]. One of these approaches is the Gaussian finite mixture model (FMM) approach which fits an FMM using source means. In this paper, this was fitted using the $mclust$ package in R [R Core Team, 2022, Scrucca et al., 2016]. Further, this model can be improved by adding the estimated within-source covariance to the covariance estimates from the FMM which we will refer to as FMM+C.

A. Proposed method

Recall the random effects model with K between source subpopulations. Let $Z_i \sim Multinoulli(\tau_1, \tau_2, ..., \tau_K)$ be the subpopulation membership of the i^{th} source where τ_k is the probability that a source is in the k^{th} subpopulation. In this case, the random effects model can be rewritten as

$$\boldsymbol{X}_{ij} = \boldsymbol{a}_i + \boldsymbol{\epsilon}_{ij}, \tag{5}$$

where $\boldsymbol{a}_i | Z_i = k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the source sampled from the k^{th} subpopulation. In this case, it is still assumed that the within-source covariance $\boldsymbol{\Sigma}_\epsilon$ remains the same over the K subpopulations. Since $\boldsymbol{a}_i | Z_i = k$ and $\boldsymbol{\epsilon}_{ij}$ have Gaussian distributions, $\boldsymbol{X}_{ij} | Z_i = k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k^*)$ where $\boldsymbol{\Sigma}_k^* = \boldsymbol{\Sigma}_k + \boldsymbol{\Sigma}_\epsilon$.

B. SSFMM and Expectation-maximization algorithm

The mixture model based on the random effects model can be written as

$$f(\boldsymbol{x}_{ij} | \boldsymbol{\Psi}) = \sum_{k=1}^{K} \tau_k \phi(\boldsymbol{x}_{ij} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k^*), \tag{6}$$

where $\boldsymbol{\Psi} = \{\tau_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k^*\}_{k=1,...,K}$ needs to be estimated with the constraint that fragments coming from the same source need to come from the same subpopulation.

Define the i^{th} source indexing set as $S_i = \{i1, i2, \ldots, im\}$. Therefore $\bigcup_{i=1}^{n} S_i = \{11, 12, ..., nm\}$ and $S_i \bigcap S_{i'} = \emptyset \;\; \forall i, i' \in \{1, , , n\}$ which is needed for the semi-supervised algorithm in Melnykov et al. [2015]. Let $i(j)$ be defined such that $S_{i(j)} = S_i$. Let $\boldsymbol{Z}^+ = \{S_1, S_2, ..., S_n\}$ be the set of positive constraints in which $j, j' \in S_i$ implies $Z_{ij} = Z_{ij'}$ where Z_{ij} is the subpopulation membership of the j^{th} trace from the i^{th} source. The *E-step* becomes

$$\pi_{ijk}^{(t+1)} = \frac{\tau^{(t)}{}_{k}^{|S_{i(j)}|} \prod_{q \in S_{i(j)}} \phi(\boldsymbol{x}_{iq} | \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})}{\sum_{k'=1}^{K} \dot{\tau}_{k'}^{|S_{i(j)}|} \prod_{q \in S_{i(j)}} \phi(\boldsymbol{x}_{iq} | \boldsymbol{\mu}_{k'}^{(t)}, \boldsymbol{\Sigma}^*{}_{k'}^{(t)}))}. \tag{7}$$

where $|S_{i(j)}|$ denotes the cardinality of the set $S_{i(j)}$. Note that with this we have $\forall j, j' \in S_i$, the posterior probabilities $\ddot{\pi}_{ijk}$ and $\ddot{\pi}_{ij'k}$ are equal satisfying the constraint. The M-step is

$$\tau_k^{(t+1)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ijk}^{(t+1)}}{mn}, \boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \boldsymbol{x}_{ij} \pi_{ijk}^{(t+1)}}{\sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ijk}^{(t+1)}},$$

$$\boldsymbol{\Sigma}_k^{*(t+1)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} (\boldsymbol{x}_{ij} - \boldsymbol{\mu}_k^{(t+1)})(\boldsymbol{x}_{ij} - \boldsymbol{\mu}_k^{(t+1)})' \pi_{ijk}^{(t+1)}}{\sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ijk}^{(t+1)}}. \tag{8}$$

Note that the M-step is similar to the usual FMM model and the constraint only affects the E-step of the algorithm. The EM algorithm then iterates between the E-step and M-step until a prespecified convergence is reached. The relative change in likelihood values is used in our work with $\epsilon = 10e-6$. This criterion is given as

$$\frac{|L(\boldsymbol{\Psi}^{(t+1)} | \boldsymbol{x}^{mn}) - L(\boldsymbol{\Psi}^{(t)} | \boldsymbol{x}^{mn})|}{L(\boldsymbol{\Psi}^{(t+1)} | \boldsymbol{x}^{mn})} < \epsilon$$

C. Comparison metric

As the BIC can not be used for model selection in this case (likelihood is guaranteed to decrease in the constrained model), we opt to use a version of cross-validation approach for comparing the different models.

The Train-test split approach:

1) One sample(fragment) is randomly removed from each source in the dataset and is placed into the test set.
2) The rest are used in the training set to fit each of the competing models.
3) Let $Z_j^{(l)}$ be the membership assigned to the j^{th} source by the l^{th} model.
4) Then let $\hat{Z}_{jr}^{(l)}$ be the predicted subpopulation membership by the l^{th} model of the removed sample from the j^{th} source.
5) If $\hat{Z}_{jr}^{(l)} = Z_j^{(l)}$ then we say the subpopulation membership of that sample was correctly identified by the l^{th} model.
6) The accuracy of the model: $acc^{(l)} = \frac{1}{m} \sum_{j=1}^{m} I(\hat{Z}_{jr}^{(l)} = Z_j^{(l)})$.

Using these steps we can study the out-of-sample performance of the competing approaches.

D. Simulation setup

A simulation study was conducted by randomly generating mixture models by varying four settings using the MixSim [Melnykov et al., 2012] R package; where $K \in \{2, 5\}$, the number of subpopulations, $p \in \{2, 5, 10\}$, the number of variables, $\omega \in \{0.01, 0.1\}$, the average overlap between components, which is described in Maitra and Melnykov [2010], and $\alpha \in \{0.1, 0.2\}$ the within scale, is used to generate $\boldsymbol{\Sigma}_\epsilon$. To obtain the within covariance $\boldsymbol{\Sigma}_\epsilon$ the pooled covariance between all K between covariance structures is computed and then multiplied by α. Between- and within-source samples are

This resource was prepared by the author(s) using Federal funds provided by the U.S.
Department of Justice. Opinions or points of view expressed are those of the author(s) and do not
necessarily reflect the official position or policies of the U.S. Department of Justice.

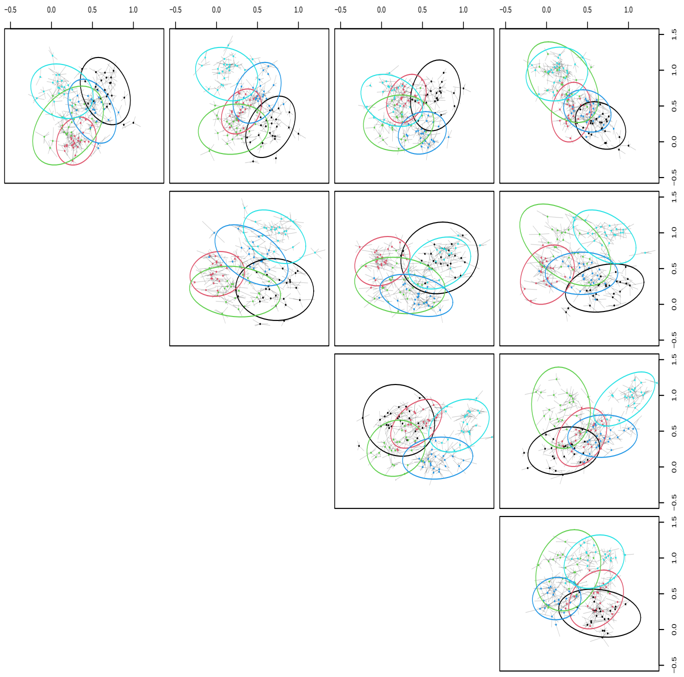

Fig. 5. A pairwise scatterplot of example data generated from a randomly generated 5-dimensional mixture model with 5 subpopulations, dots representing sources means, and gray lines representing trace object measurements from a given source. The 95% probability contours for mixtures are given.

selected such that $n_b^* \in \{5, 15\}$ and $n_w \in \{5, 15\}$ where $n_b = n_b^* * p * K$. From each combination $(K, p, \omega, \alpha, n_b, n_w)$, 100 datasets are simulated. An example from one of these randomly generated mixture models can be seen in Figure 5.

E. *Results of simulation*

We can see from Figure 6 that as the complexity of the problem increases by increasing the number of subpopulations, the dimensionality, the overlap between subpopulations, and the within-source covariance, then the accuracy of all methods decreases. We present four cases in Figure 6. Comparing our proposed method SSFMM to FMM and FMM+c note the following:

- SSFMM has overall higher accuracy than FMM and FMM+C in almost all the cases.
- SSFMM utilizes all the data, hence even with small within samples when the other two methods degrade, SSFMM performs well.
- The difference in accuracy between the new SSFMM method and the other two methods increases as the complexity of the problem increases.

F. *Application to forensic glass data (Zadora's)*

In this section, we fitted FMM, FMM+C, and SSFMM to the glass data presented in Section IV. The models are fitted on the training set for various values of K and $acc^{(l)}$ is computed based on the test set. This was repeated 25 times. See Figure 7 for the out-of-sample performance of the three approaches. Similar to the simulation study, in this glass data, we can see that for varying numbers of components, the SSFMM method outperforms the FMM+C.

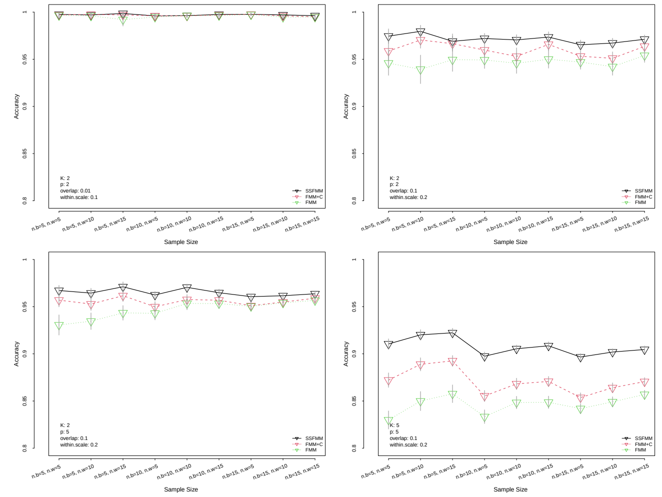

Fig. 6. Four examples of the (K, p, ω, α) combinations for the simulation study results. The accuracy of each method vs the within and between- sample sizes for SSFMM (black-solid), FMM+C (red-dashed), FMM (green-dotted) lines are given.

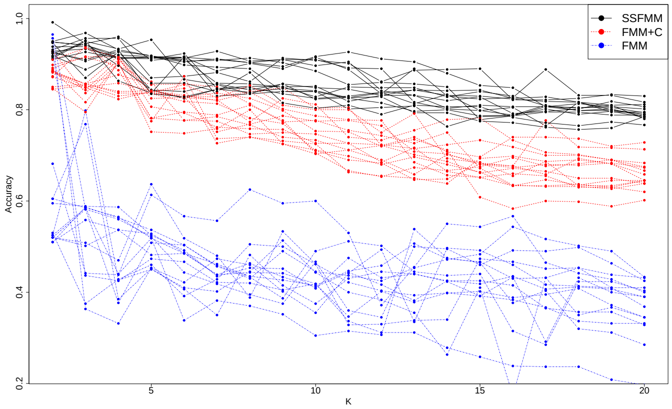

Fig. 7. Performance of the three models considered: FMM, FMM+C, and SSFMM for ranging the number of components.

In addition, recall that we have three technical replicates within each fragment. We can assess if the proposed method consistently assigns technical replicates to the same subpopulation. Let $\mathbf{Z}_i^l = \{\hat{Z}_{i1}^{(l)}, \hat{Z}_{i2}^{(l)}, \hat{Z}_{i3}^{(l)}\}$ be the set of estimated subpopulation memberships of the technical replicate of the removed fragments belonging to the i^{th} window in the test set for the l^{th} model. Let C_i be the number of unique subpopulation memberships in $\mathbf{Z}_i^{(l)}$. Then the comparison statistic is defined as $\bar{C}^{(l)} = \frac{1}{m}\sum_{i=1}^{m} C_i$. The results are shown in Figure 8. As the number of subpopulations increases we see overall all three approaches tend to assign replicates to more than one subpopulation. However, the SSFMM method produces the lowest number of subpopulations on average with smaller

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

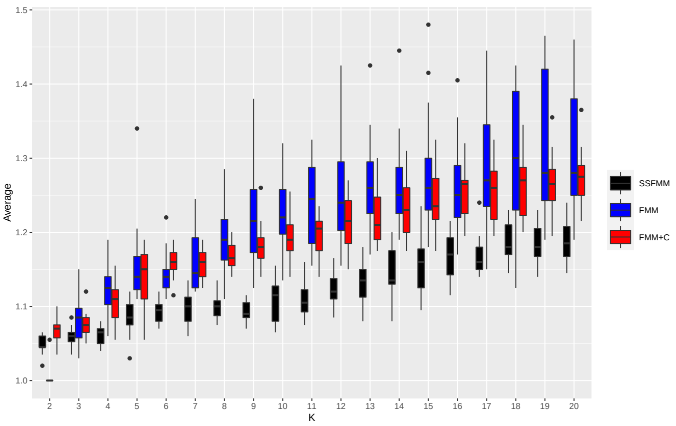

Fig. 8. Boxplot of $\bar{C}^{(l)}$ measure of consistency on membership assignment of technical replicates over varying values of K.

variability.

V. Conclusions

The work in this report identified that there could be an alarming amount of algorithmic bias towards a minority population as measured by rates of misleading evidence. The likelihood ratio methods available and widely used such as the Aitken et al. [2007] LR's are susceptible to this algorithmic bias. The study of the efficiency of these LR methods showed that both can be inefficient depending on the existence of subpopulation structures and the number of sources in the background population. This suggests the need for finding a robust LR method that is efficient and will mitigate the algorithmic bias observed. We proposed a mixture-based solution to model subpopulations in hierarchically structured data. The semi-supervised model was more accurate over random train test split validation. The semi-supervised approach also performs better at assigning the same membership to technical replicates of the same fragments. The smaller variability supports that the semi-supervised approach gives a more reliable model. More work is needed to implement the developed semi-supervised mixture models into an LR. This line of work aligns with the current NIJ initiatives. Particularly, in a recent interview [National Institute of Justice, May 15, 2023] with the NIJ director Dr. La Vigne states: "Foster(ing) rigorous research to promote safer communities and more equitable justice system" ... where "... researchers should be intentional in examining potential structural inequalities that may generate disparate outcomes based on one's gender, race, ethnicity, religion, sexual identity or citizenship status, regardless of research topic".

Acknowledgements

The research detailed in this presentation was partly supported by Award No. 2019-DU-BX-0011 (Dr. Larry Tang as the PI) awarded by the National Institute of Justice, Justice Programs, US Department of Justice.

The presenter of this work was funded by the National Justice Institute in the form of a Travel Grant to attend this conference. Some of the research detailed in this presentation was supported in part by NSF NRT: Cyber-Physical-Social System for Understanding and Thwarting the Illicit Economy (DGE 1828492, 1828302, 1828462, 1828288).

The materials presented here solely represent the views of the authors and do not represent the view of the US Department of Justice, the National Science Foundation, or the United States government.

References

Colin G. G. Aitken, Grzegorz Zadora, and David Lucy. A Two-Level Model for Evidence Evaluation. *Journal of Forensic Sciences*, 52(2):412–419, March 2007. ISSN 0022-1198, 1556-4029. doi: 10.1111/j.1556-4029.2006.00358.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1556-4029.2006.00358.x.

Colin G. G. Aitken and David Lucy. Evaluation of Trace Evidence in the Form of Multivariate Data. 53(1): 109–122, 2004. URL JournaloftheRoyalStatisticalSociety.SeriesC(AppliedStatistics).

Joshua R Dettman, Alyssa A Cassabaum, Christopher P Saunders, Deanna L Snyder, and JoAnn Buscaglia. Forensic discrimination of copper wire using trace element concentrations. *Analytical chemistry*, 86(16):8176–8182, 2014.

Javier Franco-Pedroso, Daniel Ramos, and Joaquin Gonzalez-Rodriguez. Gaussian Mixture Models of Between-Source Variation for Likelihood Ratio Computation from Multivariate Data. *PLOS ONE*, 11(2):e0149958, 2016. doi: 10.1371/journal.pone.0149958.

David Lucy, James Curran, and Agnieszka Martyna. *comparison: Multivariate Likelihood Ratio Calculation and Evaluation*, 2020. URL https://CRAN.R-project.org/package=comparison. R package version 1.0-5.

Ranjan Maitra and Volodymyr Melnykov. Simulating data to study performance of finite mixture modeling and clustering algorithms. *Journal of Computational and Graphical Statistics*, 19(2):354–376, 2010.

Volodymyr Melnykov, Wei-Chen Chen, and Ranjan Maitra. MixSim: An R package for simulating data to study performance of clustering algorithms. *Journal of Statistical Software*, 51(12):1–25, 2012. URL https://www.jstatsoft.org/v51/i12/.

Volodymyr Melnykov, Igor Melnykov, and Semhar Michael. Semi-supervised model-based clustering with positive and negative constraints. *Advances in Data Analysis and Classification*, 10(3):327–349, 2015. doi: 10.1007/s11634-015-0200-3.

National Institute of Justice. The united nations crime prevention and criminal justice programme network of institutes interviews nij director la vigne, May 15, 2023. nij.ojp.gov: https://nij.ojp.gov/topics/articles/united-nations-crime-prevention-and-criminal-justice-programme-network-institutes, Last accessed on June 2023.

Danica M. Ommen and Christopher P. Saunders. A Problem in Forensic Science Highlighting the Differences between the Bayes Factor and Likelihood Ratio. *Statistical Science*, 36

This resource was prepared by the author(s) using Federal funds provided by the U.S.
Department of Justice. Opinions or points of view expressed are those of the author(s) and do not
necessarily reflect the official position or policies of the U.S. Department of Justice.

(3):344 – 359, 2021. doi: 10.1214/20-STS805. URL https://doi.org/10.1214/20-STS805.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.

Luca Scrucca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289–317, 2016. URL https://doi.org/10.32614/RJ-2016-021.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.