Estimating error rates for bullet comparisons in forensic science

Yawei Ge[1], Heike Hofmann[1]

[a] *Department, Street, City, State, Zip*
[b] *Department, Street, City, State, Zip*

Abstract

This is the abstract.

It consists of two paragraphs.

Text based on elsarticle sample manuscript, see http://www.elsevier.com/author-schemas/latex-instructions#elsarticle

Link for the outline document

https://github.com/yaweige/mixture-beta-model-fit/blob/master/other-informal-writeup/outline1.pdf

Introduction and literature review

Generally, the structure is fine, but you need to make sure that you introduce everything that you use. e.g. RF needs to be first introduced as Randomforest. The question then becomes whether it is necessary to restrict yourself to RF scores, or whether ccf scores have the same/similar structure. In case you go with RF scores, you need to make sure to sketch out the pipeline.

Include proper references and figures. I can't guess what you mean unless you actually include the things exxplicitly.

The firearm examiners are focusing on the same source and different source problems of bullets and cartridge cases which serve as important forensic evidence in the court/to the jurors. XX this is quite vague - could you try to find a number in the literature on how much evidence in trials is related to firearms or ballistic evidence? The subjectivity and lack of quantified error rates in the traditional forensic process is called to be reduced or complemented by more objective methods (**???**; National Research Council, 2009). Some XX 'some' is too vague - instead cite literature - i.e. something like: in response xxx, yyy, zzz, ... have introduced automatic matching algorithms ... automatic matching algorithms are developed which usually return a similarity score to quantify the similarity or the

Email addresses: `yaweige@iastate.edu` (Yawei Ge), `hofmann@iastate.edu` (Heike Hofmann)

probability to be an actual match for a certain comparison. However, this raises questions about how to interpret the reported scoresand how these scores are distributed. XXX more details before this thus - why is it XXX it? write out your reference not clear to draw inference? Thus, it XXX it? write out your reference is not all clear how to conduct inference based on these similarity scores. Song et al. (2018) proposed binomial and beta-binomial for the number of matched cells of the CMC method for cartridge case comparisons. Therefore, Song et al. (2018) provides a way to quantify the theoretical error rate of the algorithm. However, for the bullets LEA comparisons, quantitative measurements for the theoretical error rate have not been established. In this paper, we will evaluate the possible models/distributions for the LEA comparisons scores generated by the random forest proposed by Hare et al. (2016). And then, we will also evaluate the error rates based on the estimated distribution for the automatic matching algorithm.

In section 2, we will discuss distributional forms of random forest scores produced by the automatic LEA matching algorithm proposed by Hare et al. (2016). In section 3, we will introduce the LAPD data set and the estimated distributions. In section 4, the theoretical error rates based on the distributions are discussed. In section 5, we evaluate the performance of the estimation, stability of the distribution within a changing sample size context. In section 6, we will conclude the discussion.

The distributional forms of the similarity scores

The quantitative methods used to objectively measure the similarity between LEAs reports various quantities, such as counts, correlations, distances, probabilities and more generally similarity scores [references]. To understand how those quantities reflect the strength of evidence and to study the underlying the error rates in making decisions based on those quantities, distributional forms are usually set up [references]. Particularly, we are focusing on the similarity scores which range from 0 to 1 and the probabilities reported as the likelihood of an actual match. The similarity scores reported in the forensic researches are usually categorized into two categories. One is that the compared LEAs or other forensic evidences are actual matches, the other is that the compared LEAs or other forensic evidences are not matches. We name the former as known matches (KM) and the latter as known non-matches (KNM), and the corresponding distributions are named as KM distributions and KNM distributions as in Figure ... for example. However, this kind of classifications can be investigated only in the lab environment where the ground truth is known. When we make any decisions based on any quantitative measurement, we are actually making a distinguish between those two potential distributions. The strength of any identification process is also measured by the disparities of those distributions. However, in practice, we can hardly discriminate those two distributions entirely, thus, we are never 100% sure which distribution the observed score comes from. This is where the identification error raises.

For the purpose of illustration, we have a look at the example (could use Hamby and other sets, not necessarily LAPD) in Figure .... Those are RF scores generated from the automatic matching algorithm [references]. The scores from the RF can be explained as probabilities that calculated through the algorithm based on the trained model to quantify the likelihood that a pair of LEAs are actual a match. Or we can think of the RF scores as general similarity scores that quantify the similarities. As the name indicated, the higher RF scores imply higher chances to be a match. For different combination of ammunition and firearms, the scores are distributed differently. It is expected that systematic differences exist there for different cases. So, it is necessary to study the scores under controlled conditions, thus, a threshold for the scores to do classifications varies based on the changes of the underlying distributions.

We can see from the Figure ..., which is a typical one we usually have for the scores, that the distributions are apart for the majorities. In the bullet LEA comparison problems, we usually have a well separated bullet scores but for the land scores, there are usually some overlaps [Susan]. We propose beta distributions for those scores. Because the beta distribution is a well-used one in statistics to describe a quantity from 0 to 1 which is usually a probability for another distribution in Bayesian analysis. And it is very flexible to capture any unimodal shape from 0 to 1. However, it may not be adequate to explain a heavy tail or even a second mode. Thus, we consider the beta mixture distribution which is a more complex distribution than the beta distribution which is a special case of the beta mixture distribution. In the beta mixture distribution, we introduce a hierarchical structure with a prior probability to combine two beta distributions as one. The distributions are defined below.

The reported random forest scores are denoted as Y_{ij} for j^{th} LEA comparison within class i, where $i = 1$ is KM and $i = 2$ is KNM. Y_{ij}'s are considered independent and identically distributed within each class, i.e.

$$Y_{1j} \overset{iid}{\sim} Betamix(p_1, \mu_{11}, \phi_{11}, \mu_{12}, \phi_{12})$$
$$Y_{2j} \overset{iid}{\sim} Betamix(p_2, \mu_{21}, \phi_{21}, \mu_{22}, \phi_{22})$$

where μ_{ik} and ϕ_{ik} are distribution parameters for i^{th} class and k^{th} component, $k = 1$ or 2. And p_i is the prior probability for the first component in i^{th} class, thus, $1 - p_i$ is the prior probability for the second component in i^{th} class. Note that we are using the mean and precision parameterization of beta distributions which simplifies the math in calculation. It's equivalent to the usual α and β parameterization through the following transformation:

$$\mu = \frac{\alpha}{\alpha + \beta}$$
$$\phi = \alpha + \beta$$

LAPD data set and the estimated distributions

For the following sections of the paper, we will base our analysis on the LAPD data sets. This is a large data set of ... It is the first time such a large data set available to the researchers, which makes it possible for a statistical analysis for the distribution of similarity scores. (More! And number of land comparisons for different cases)

We consider cross correlation functions (CCF) of lands, which is produced by calculating the maximized CCF between two signatures extracted from a pair of bullet LEAs (refer to a paper, could be the random forest one or others).

The estimation was done using Nelder-Mead algorithm in R [reference]. This is a general purpose numeric method which works reasonably well for non-differentiable function. In our case, the non-differentiable objective function is the log likelihood function of the beta mixture distribution. Therefore, we finally found the maximum likelihood estimators (MLE). Instead of a simpler beta model, we start with the more complex two-component beta mixture model, see how well it fits the data and test if some components are necessary. We estimated beta mixture distributions for both KM and KNM as in Table 1 and Table 2.

Full Data KM Distribution Estimation						
Model	Component Prior Probability	μ	ϕ	logLik	p-value	BIC
3-comp	0.247	0.406	24.097	6278		-12479
	0.521	0.664	13.342			
	0.232	0.819	33.823			
2-comp	0.419	0.466	15.812	6243	0	-12438
	0.581	0.759	18.476			
Beta		0.635	6.529	5712	0	-11404

Table 1: Parameter estimations for the Beta distribution (1-component beta mixture distribution), 2-component and 3-component distributions for the KM CCF. "2-comp" refers to the 2-component beta mixture distribution, and the same for "3-comp". Column "logLik" is the maximized log likelihood for each distribution. Column "p-value" is the p-value for asymptotic likelihood ratio tests between the current model and one-step more complex model, where 0 indicates that we would reject the hypothesis that the current one is sufficient to decribe the data

The estimated Beta distributions are shown in the Figure 1 and the estimated two-component distribution are also shown in the Figure 2. As expected, the estimated distributions show the properties we desire for the similarity scores. The majorities of the estimated distributions are apart from each other, while the minority part between the two distributions has some overlap. It's worth to note that both curves have heavy tails to the farther boundaries and KM curve has heavier tails. We can see the two-component distributions fit the data very well, while the single beta distributions are not as good as the two-component distributions for both KM and KNM. The single beta distribution for KM clearly

Full Data KNM Distribution Estimation						
Model	Component Prior Probability	μ	ϕ	logLik	p-value	BIC
3-comp	0.650	0.345	35.017	78961		-157830
	0.314	0.490	24.979			
	0.032	0.667	18.454			
2-comp	0.674	0.358	39.908	78808	0	-157558
	0.336	0.494	13.324			
Beta		0.404	15.801	75349	0	-150675

Table 2: Parameter estimations for the Beta distribution (1-component beta mixture distribution), 2-component and 3-component distributions for the KNM CCF. "2-comp" refers to the 2-component beta mixture distribution, and the same for "3-comp". Column "logLik" is the maximized log likelihood for each distribution. Column "p-value" is the p-value for asymptotic likelihood ratio tests between the current model and one-step more complex model, where 0 indicates that we would reject the hypothesis that the current one is sufficient to decribe the data

failed to capture the potential second mode of the histogram indicating it is not sufficient. The single beta distribution for KNM is not too bad but still failed to capture the distributional information for some parts. This indicates that even though beta distributions are flexible for variables in 0 to 1, they are still restricted too much to be able to describe the cases here. The two-component beta mixture distributions look more promising.

(This paragraph doesn't match the current results and we are investigating the estimations for now) It's also helpful to see how the individual components look like in the beta mixture distributions [figure]. The KNM and KM distributions seem to share a common component while keep the other components far apart from each other. The separated components represent the ideal cases where the bullet land engraved areas preserved the information of source well and result clear distinct separation. The shared components represent the cases where the KM results lower scores because of some degree of tank rash, pitting, breakoff or other damages on the bullets and the KNM results higher score because of the random identification effect. These properties together well explained the observed empirical distribution of similarity scores in our cases. Particularly, the heavier tail of the KM comparisons is explicitly included the form of the model by one of the components.

We need to do formal statistical tests to verify some of the points mentioned before. We would like to know if simpler beta distributions are sufficient to model the data. We would also quantify the variability of estimations of the distribution through parametric bootstrap.

(test, bootstrap inference, and report the model for the full data)

Evaluate the error rates

(report theoretical error rates)

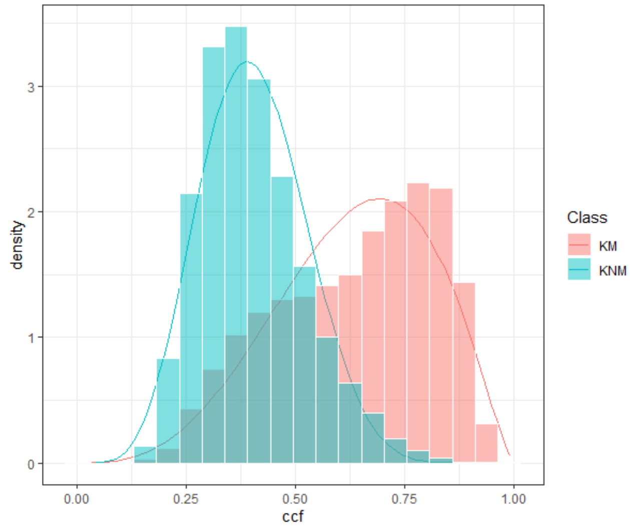


Figure 1: Estimated two-component distributions for full data

Estimations with changing sample size

(quantify the variation, reproduce some results in the previous sections with changing sample size)

Conclusion

```
ggplot() +
  geom_line(data = fau330_combined_noA_simulated,
            mapping = aes(x = ccf, y = y, color = id)) +
  geom_rug(data = fau330_combined_A_observed,
           mapping = aes(x = ccf, color = id), alpha = 0.3) +
  ggtitle("FAU330: rugs with A, densities without A (ccf)")
```

```
ggplot() +
  geom_line(data = fau330_combined_noA_rf_simulated,
            mapping = aes(x = rfscore, y = y, color = id)) +
  geom_rug(data = fau330_combined_A_rf_observed,
           mapping = aes(x = rfscore, color = id), alpha = 0.3) +
  ggtitle("FAU330: rugs with A, densities without A (RF)")
```

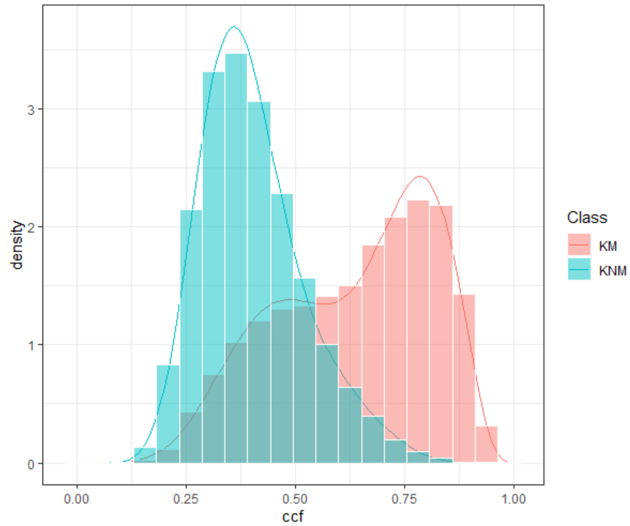
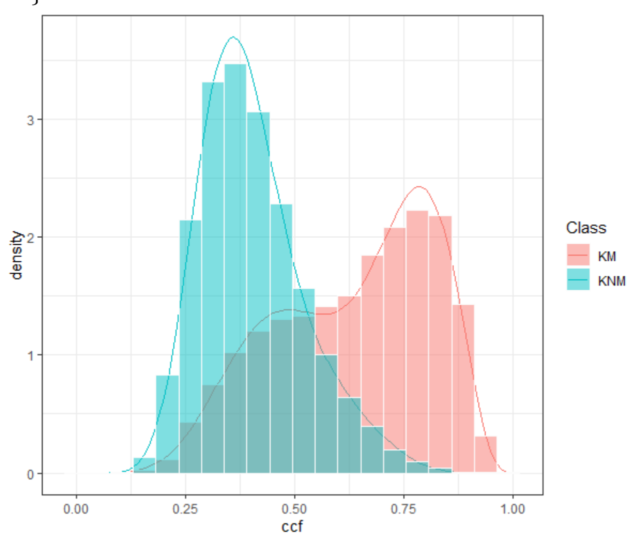


Figure 2: Estimated two-component distributions for full data

```r
knitr::include_graphics("../code/organized code/generated figures/two-component-betamix.png'
```

\begin{figure}



{

}

\caption{rf_full_2components_combined} \end{figure}

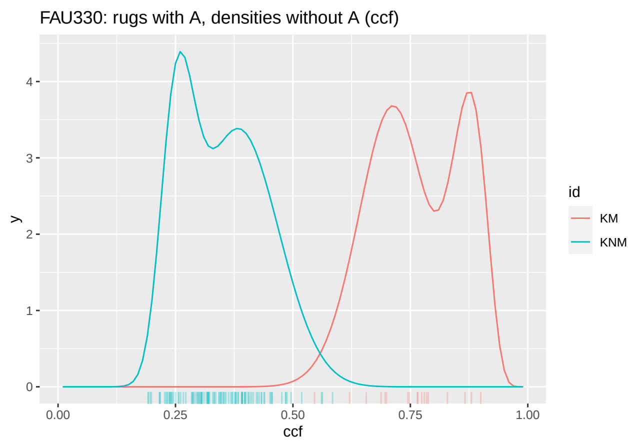


Figure 3: Densities from training, rugs from test (ccf)

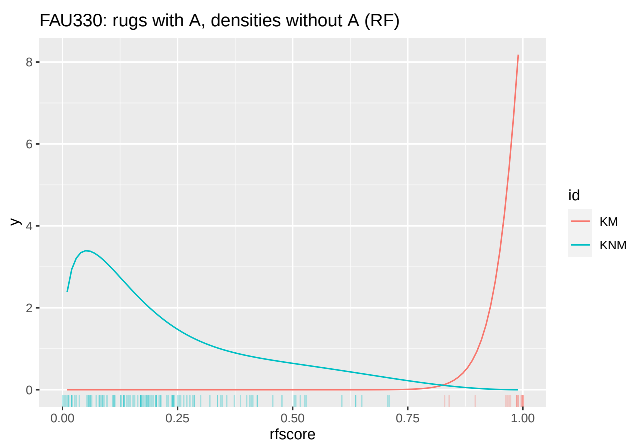


Figure 4: Densities from training, rugs from test (rfscore)

The following content is help information of the Elsevier template (keep for reference for a while)

The Elsevier article class

Installation. If the document class *elsarticle* is not available on your computer, you can download and install the system package *texlive-publishers* (Linux) or install the LaTeX package *elsarticle* using the package manager of your TeX installation, which is typically TeX Live or MikTeX.

Usage. Once the package is properly installed, you can use the document class *elsarticle* to create a manuscript. Please make sure that your manuscript follows the guidelines in the Guide for Authors of the relevant journal. It is not necessary to typeset your manuscript in exactly the same way as an article, unless you are submitting to a camera-ready copy (CRC) journal.

Functionality. The Elsevier article class is based on the standard article class and supports almost all of the functionality of that class. In addition, it features commands and options to format the

- document style
- baselineskip
- front matter
- keywords and MSC codes
- theorems, definitions and proofs
- lables of enumerations
- citation style and labeling.

Front matter

The author names and affiliations could be formatted in two ways:

(1) Group the authors per affiliation.
(2) Use footnotes to indicate the affiliations.

See the front matter of this document for examples. You are recommended to conform your choice to the journal you are submitting to.

Bibliography styles

There are various bibliography styles available. You can select the style of your choice in the preamble of this document. These styles are Elsevier styles based on standard styles like Harvard and Vancouver. Please use BibTeX to generate your bibliography and include DOIs whenever available.

References

Hare, E., Hofmann, H., Carriquiry, A., 2016. Automatic matching of bullet land impressions. Annals of Applied Statistics 11, 2332–2356.

National Research Council, 2009. Strengthening forensic science in the united states: A path forward, Strengthening Forensic Science in the United States: A Path Forward. National Academies Press. doi:10.17226/12589

Song, J., Vorburger, T.V., Chu, W., Yen, J., Soons, J.A., Ott, D.B., Zhang, N.F., 2018. Estimating error rates for firearm evidence identifications in forensic science. Forensic Science International 284, 15–32. doi:10.1016/j.forsciint.2017.12.013