Estimating error rates for bullet comparisons in forensic science

Yawei Ge[a,b], Heike Hofmann[a,b]

[a] *Department of Statistics, Iowa State University,*
[b] *Center for Statistics and Applications of Forensic Evidence, Iowa State University*

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
Thanks for including the color, that is helpful! Next steps:

1. for any image that you include with include_graphics("image.png") write a ggsave("image.png") in the code
2. Let's use darkgrey for different source and darkorange for same source comparisons
3. Switch from known match and known non-match to (known) same source and (known) different source
4. Whenever you show results - number estimates or figures, make sure to include in the description what bullets/barrels these results are based on.
5. in the section on the LAPD data define what you mean by 'full data' and what sampling schemes you use later on.

1. Introduction

Firearm examination is an important topic in forensic science to help identify if two pieces of firearm evidence are from a same source or not by examining the characteristics on the evidence left by the firearms. Currently, those examinations are executed in forensic labs by Firearms and Toolmark Examiners (FTEs) under the regulation of AFTE (AFTE Criteria For Identification Committee, 1992). However, this process is in its nature based on human decisions and therefore subjective. The National Research Council criticized this subjectivity in the conventional firearm identification process and called for it to be reduced or to be complemented by more objective procedures (National Research Council, 2009). The President's Council of Advisers on Science and Technology (PCAST) also emphasized the importance to establish the scientific validity and reliability

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

of subjective forensic feature-comparison methods by blind empirical testing (President's Council of Advisors on Science and Technology (PCAST), 2016). Additionally, the PCAST report recognized the importance of developing of objective computer-based algorithms in its following addendum (President's Council of Advisors on Science and Technology (PCAST), 2017).

In response to this criticism, subsequent research on cartridge case comparisons introduced the method of congruent matching cells (CMC) (Song, 2015), which allows a quantitative assessment of the similarity of two breech face impressions. Theoretical error rates of CMC were analyzed by Song et al. (2018) and Zhang (2019) on the data set of Fadul Jr. et al. (2011) based on variants of a binomial model.

XXX xiao-hue Tai's paper should be mentioned.

XXX mention the data that all of these methods are based on.

In bullet comparisons, an automatic matching algorithm was proposed by Hare et al. (2016). Empirical error rates associated with this method are discussed by Vanderplas et al. (2020) based on three test sets . The most prominent single feature in the random forest model is the cross-correlation function, previously identified as important in other methods of similarity assessment (Chen et al., 2019; Chumbley et al., 2010; Krishnan and Hofmann, 2019; Vorburger et al., 2011). }

One less well investigated aspect of the cross-correlation measure, however, is its distribution.

Due to limitations of sample sizes and study designs, a sound distributional assumption about the resulting similarity scores is not well established in the bullet comparison field for automated matching algorithms. Error rates are generally estimated empirically from only dozens of bullets originally used to estimate the current lab practice's error rates. And many of those tests are also criticized for the closed-set design and potential information provided to the examiners as pointed out by PCAST (President's Council of Advisors on Science and Technology (PCAST), 2017, 2016). It is important to understand the underlying distributions of the similarity scores to estimate the theoretical error rates of any automatic algorithm. The distributions also play a key role in likelihood ratio analysis to quantify the strength of various similarity scores and unify the measurement of strength of evidence. And that also helps later to choose proper thresholds for the methods to reduce the overall error rates instead of simply choosing a threshold to distinguish between same and different sources.

XXX Let's end the introduction here with a paragraph on the structure of the rest of the paper.

In section 2... In section 3...

2. Background

A sketch of a bullet and a barrel would be good here - showing rifling (grooves and lands) and striation marks. Introduce rifling, Describe LEA and GEA, and structure of LEAs

Generally, in forensic science, we consider the problem of determining for two pieces of evidence whether they originate from the same source or from different sources. Specifically, when bullet evidence is compared, it is of interest if two bullets were fired from the same gun. For firearms, there is rifling machined into the internal of the barrels aiming to improve the aerodynamic stability of bullets. During the firing process, microimperfections in a barrel leaves markings on a bullet. These markings are thought to be unique and are used as the basis to determine the source of the bullets. For bullet evidence, in particular, striations on land engraved areas (LEAs) are used by firearms and toolmarks examiners to determine same source of two pieces of evidence (AFTE Criteria For Identification Committee, 1992). For computer based algorithms, 3D scans of the LEAs are used. Following the process described in Hare et al. (2016), we extract **signatures** XXX pictures! from these scans. Various scores are used in the literature to describe the similarity between two signatures (Hare et al., 2016; Ma et al., 2004; Nichols, 2003, 1997). Here, we will employ the cross-correlation function to measure similarity.

This provides us with a rich tool box and help the field advance, but it comes with issues. For the practitioners not in statistical major who don't fully understand the differences among those algorithms, a same number represents different strength of evidence is extremely confusing. And even in the academic field, we need a unified way to quantify the strength of evidence considering different behaviors of different scores. The score based on likelihood ratios is proposed in forensic science to measure the strength of evidence. And it represents how much one result is more likely than the other one and comes with great statistical properties in terms of minimizing losses and achieving large sample properties. The key of making use of likelihood ratios is to understand the underlying theoretical distributions of each particular type of scores. Intuitively, without knowing how a type of score is distributed, it makes no sense to make any claim for a particular value. For example, the random forest method in the bullet comparison gives very separated scores, either close to one or close to zero, and it actually achieves good classification results. But it doesn't mean it won't make an error, it makes error even when the score close to zero which we might expect to be different source pairs. But when CCF achives that score close to 0, we hardly make wrong claims. It's because the difference of distributions of each score and the absolute value itself could be misleading. And the prediction error rates should be evaluated and controlled at the same time.

3. The distributional forms of similarity scores

There are two types of distributions we have to consider in the case of bullet comparisons. One is called known-match (KM) or same-source distribution for similarity scores from same-source pairs of LEA scans. The other is called known-non-match distributions (KNM) for those actually not matched lands.When we make any decisions based on any quantitative measurement, we are actually making a choice between those two potential distributions. The strength of any identification process is also measured by the disparities of those distributions.

However, those two distributions are only available in experimental settings where we know the ground truth. And the distributions could be different for different scores and different combination of firearm and bullets. In particular, we consider the cross-correlation function (CCF) for bullet signatures as a typical similarity score with a practical range of 0-1. Although CCF could theoretically be negative values, that is almost never observed for LEA comparisons in practice (none of our 121992 cases). And it behaves well under a distribution in 0-1 and we can always convert it to 0-1 by simple transformation if there are rare cases. So we treat CCF with the range 0-1. The cross-correlation function (CCF) is calculated after extracting the signatures of bullet lands by maximizing the cross correlation of two sequence of signatures by horizontally adjusting the signatures (Vorburger et al., 2011). And CCF is the most prominent single feature in the automatic random forest algorithm (Hare et al., 2016). So, it is selected as a representative of similarity scores with range from 0 to 1 in this paper for further analysis. As the name indicated, the higher the similarity score is, the stronger evidence is to support the same source assumption.

We can see from the Figure ... (a figure of nonparametric fit of the CCF?), which is a typical one we usually have for the similarity scores, that the distributions of KM and KNM 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 (Vanderplas et al., 2020). We propose beta distributions for those scores. Because the beta distribution is a well-used distribution in statistics to describe a quantity from 0 to 1 which is usually a probability or proportion quantifying our knowledge for another distribution in Bayesian analysis. And it is very flexible to capture unimodal asymmetric shapes in 0 to 1. However, it may not be adequate to explain a heavy tail or even a second mode. Thus, we further consider the beta mixture distribution which is a more complicated distribution than the beta distribution as a special case. In the beta mixture distribution, we introduce a hierarchical structure with a prior probability to combine a few beta distributions as one. The beta mixture distributions are defined below.

Assuming a random variable Y follows a k-component beta mixture distribution, it has the distribution expressed as follows:

$$Y \sim Betamix(p, \mu, \phi)$$

where p, μ and ϕ are parameter vectors defined as:

$$\begin{aligned}
p &= (p_1, p_2, p_3, ..., p_k), && p_l \in [0,1], l = 1, 2, 3, ..., k \text{ and } \sum_{l=1}^{k} p_l = 1 \\
\mu &= (\mu_1, \mu_2, \mu_3, ..., \mu_k), && \mu_l \in (0,1), \quad l = 1, 2, 3, ..., k \\
\phi &= (\phi_1, \phi_2, \phi_3, ..., \phi_k), && \phi_l \in (0, +\infty), \quad l = 1, 2, 3, ..., k
\end{aligned}$$

then, the $Betamix(p, \mu, \phi)$ can be explicitly written as:

$$f(x; p, \mu, \phi) = p_1 f_1(x; \mu_1, \phi_1) + p_2 f_2(x; \mu_2, \phi_2) + ... + p_k f_k(x; \mu_k, \phi_k)$$

where $f_l(\cdot; \mu_l, \phi_l)$, $l = 1, 2, 3, ..., k$ are beta distributions with probability densities as:

$$f_l(x; \mu_l, \phi_l) = \frac{\Gamma(\phi_l)}{\Gamma(\mu_l \phi_l)\Gamma((1-\mu_l)\phi_l)} x^{\mu_l \phi_l - 1}(1-x)^{(1-\mu_l)\phi_l - 1}$$

In this definition, μ_l and ϕ_l are distribution parameters for l^{th} component. And p is the prior probability for components in the beta mixture distribution and is considered as a fixed real value vector. Note that we are using the mean and precision parameterization of beta distributions which simplifies the math in calculation and is more intuitive (μ_l is the mean, and the variance is roughly proportional to the reciprocal of ϕ_l). It's equivalent to the usual α_l and β_l parameterization through the following transformation:

$$\mu_l = \frac{\alpha_l}{\alpha_l + \beta_l}$$
$$\phi_l = \alpha_l + \beta_l$$

Note that in our applications, we assume different beta mixture distributions for KM and KNM cases. And within each case, the scores are assumed independent and identically distributed.

4. 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 on the data. Is there any formal documentation for LAPD? And number of land comparisons for different cases should be discussed here)

We consider cross correlation functions (CCF) for land comparisons, which is produced by calculating the maximized CCF between two signatures extracted from a pair of bullet LEAs.

The estimation was done using Nelder-Mead algorithm in R [reference]. This is a general purpose numeric method which works reasonably well for multidimensional optimization problems. In our case, the objective function is the log likelihood function of the beta mixture distribution. Therefore, we finally found the maximum likelihood estimates (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 respectively. For the beta distributions, the KM distribution has a mean at 0.635 and the KNM distribution has a mean at 0.404. We can also see that for the two component beta mixture distributions for both KM and KNM, there are components with mean around 0.5 and the other components more separated from each other. This indicates the model successfully accounted for different situations of comparisons involving tank rash and random identification etc. This is an ideal property we would like to see and make use in explaining the similarity scores. Another point about the estimates that worth mentioning is that the ϕ of the KM beta distribution is much smaller than any of the two components of the two-component KM beta mixture distribution. This indicates that the two component distribution could be better because of smaller variances for components.

Full Data KM Distribution Estimation						
Model	Component Prior Probability	μ	ϕ	logLik	p-value	BIC
3-comp	0.247 0.521 0.232	0.406 0.664 0.819	24.097 13.342 33.823	6278		-12479
2-comp	0.419 0.581	0.466 0.759	15.812 18.476	6243	0	-12438
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 Figure 1 and the estimated two-component beta mixture distributions are also shown in 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 the KM curve has the heavier tail compared to KNM. 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 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.

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

Three candidate distributions for each of KM and KNM are considered. For the increasing of complexity, we have single beta, two-component beta mixture and three-component beta mixture distributions. And they are also nested in that order. Naturally, we first look at the maximized log-likelihood of each model, and we can do asymptotic log-likelihood ratio chi-square tests for single beta against two-component beta mixture distributions, and two-component against three-component beta mixture distributions. The p-values of those test are shown in the Table 1 and Table 2 as the column “p-value”. Surprisingly (or not), we found all those p-values are 0 which strongly suggests a more complex model when there is one. However, considering the size of the data we are using to fit these models, we can expect the statistical significance will be easily achieved since any small difference of the sufficiency will be detected. So we have to take the sample size effect into consideration. As Bayesian information criterion (BIC) is a well used criterion which takes the model complexity, sufficiency and the sample size (by a log function) into account. As shown in the column “BIC” in the tables, the BICs for the single beta distributions are larger than that of the two-component beta mixtures by relatively large proportions. And the BICs for the two-component beta mixture distributions are a little larger than that of the three-component beta mixture distributions for both KM and KNM. Obviously, we would prefer the two-component beta mixture distributions, but we would cast a doubt when it comes to the three-component beta mixture distributions. We still prefer the two-component one instead of the three-component one. The reasons are: 1) these differences of BICs between two and three component distributions are really small in proportion (by 0.3% for KM and 0.017% for KNM), 2) the estimation cost is not accounted, which will be higher for a more complex model, 3) the BICs still don’t take the sample size effect fully into account since the log function for sample size goes to flat when the sample sizes

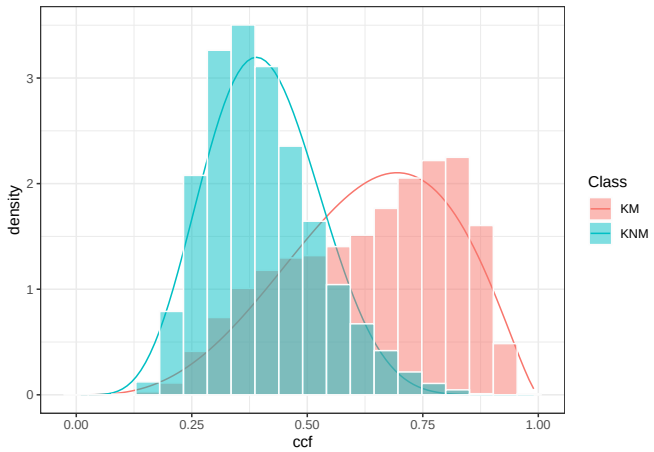

Figure 1: Estimated beta distributions for full data

are large. By simple calculation, we can see that the log-likelihood increased by a multiplicative factor more than 10 when the sample sizes increased from KM case to KNM case, however, at the same time, the log of sample size only roughly increased from 9 to 11. So we choose the two-component beta mixture distributions for both KM and KNM. And also as we have seen, these distributions have potential good forensic interpretations.

It's also helpful to see how the individual components look like in the beta mixture distributions as in Figure 3. The KNM and KM distributions seem to share a common component while keeping 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 in clearly distinct separation. The shared components represent the cases where the KM results in lower scores because of some degree of tank rash, pitting, breakoff or other damages on the bullets and the KNM results in higher score because of the random identification effect. According to the estimated prior probabilities, both distributions put less weight on the common component while putting larger weight on the the components characterizing the differences of KM and KNM respectively, which agrees on our expectation that majorities of the distributions are separated while the minorities overlapped. 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 in the form of the model by one of the components.

5. Application: Determining source in the LAPD data

move the description of the database construction into this section

In the current practice, FTEs fire two or three test bullets from a suspected firearm using ammunition that matches the evidence found on the crime scene

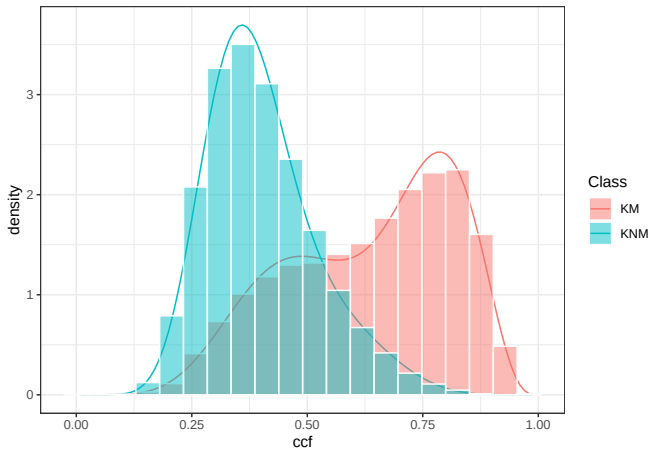

Figure 2: Estimated two-component beta mixture distributions for full data

as closely as possible. Test fires are checked for consistency of striation marks between fires. Generally, the bullet with the more well expressed striations is used to compare to the evidence. With an automated algorithm in these cases as a tool helping the examiners, we will make use of all those available test bullets and lands to generate similarity scores. And those scores will be used to get score-based likelihood ratios which will quantify the strength of evidence supporting either hypothesis from the prosecution side or the defendant side. The critical step of producing likelihood ratio is to get estimations of the distributions of both KM and KNM comparisons.

Here, we distinguish between two scenarios: 1) use scores from comparisons of test fires to estimate densities for same source and different source, and 2) use scores with known provenance from a database of similar ammunition and firearms. The second scenario is where our analysis of the theoretical distributions of the relevant population of similarity scores stand out. For a similar combination of firearms and ammunition, the similarity scores extracted through a same algorithm are considered identically distributed and comprise the relevant population, which is not a bad idea. In real cases, as stated, we can only make use of similarity scores from two or three test fires. This is good enough to produce valid comparisons but likely not good enough to estimate the distributions which generate the likelihood ratios. Besides the likelihood ratio calculation, the reference distributions are also important in deciding thresholds and control the error rates. For thresholds specifically, we choose the point where the likelihood ratio is 1 which has no preference for either hypothesis.

In either of the two scenarios, we derive scores by comparing test fires and the questioned bullet. The pipeline of the process is as follows: 1) given a questioned bullet and a suspected firearm, 2) conduct test fires with the suspected firearm and collect three test fired bullets, 3) for every pair of bullets (including the questioned), we conduct all possible land comparisons and get

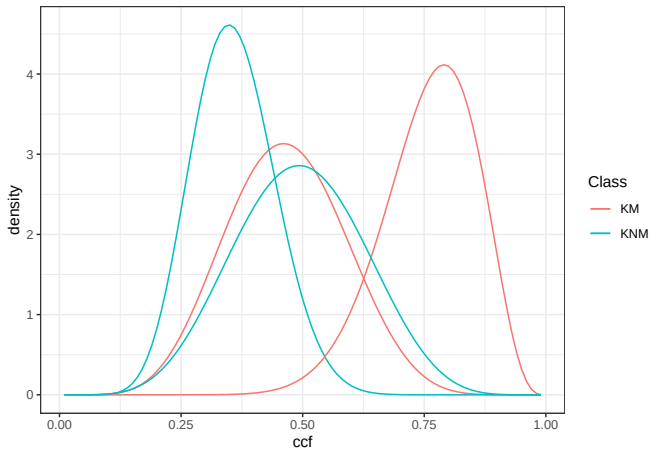

Figure 3: Estimated-components

36 land comparison scores (6 land engraved areas on each bullet), 4) among 36 comparisons within a pair of bullets, we select the six comparisons as KM comparisons (for comparisons among test fired bullets) or questioned land comparisons (for comparisons between the questioned bullet and any test fired bullet) by SAM [reference] due to the nature order of land engraved areas produced by a same barrels. Then, for the first scenario, the KM and KNM distributions will be estimated with the KM and KNM similarity scores among the test fired bullets respectively. While for the second scenario, the distributions are estimated with a database of the similarity scores from similar firearms and ammunition.

To show the usefulness of the relevant population and the better distribution estimations by using a database, we design an experiment with our LAPD data set. We randomly split our data into the database sets and case sets. For the database sets, we conduct every comparison within each set (among four bullets), and estimate the KM and KNM reference distribution as two component beta mixture distributions with all available data. For the case sets, we randomly choose one bullet out of each set (four bullets a set) as a questioned bullets and leave the other three as test fired bullets. By compare the questioned bullet with the other three from the sets, we get known-matched bullet comparisons. By randomly matching the questioned bullets with the test fired bullets (except its own set and simple exchanges of bullets between two sets), we get known-non-matched bullet comparisons. One important question is how many sets of bullets should be assigned as database sets since we would like to keep it small but sufficient and put more data into case sets to evaluate the performance. And also, this is an important question in general that how many test fires do we need in general firearm forensic practice. We will dig into this issue in the later section. But for now, we split the 442 sets into 342 database sets and 100 test sets.

Following the above procedure, we have 100 questioned bullets (same bullets in same and different source cases), 100 same source question-test pairs, 100 different source question-test pairs. For each questioned bullet in either case (same or different source), we have 18 questioned land comparisons. For the database scenario, we have 342x36 = 12312 known matched land comparisons, and 342x240 = 82080 known non-matched land comparisons. For the individual case scenario, we have 18 known matched land comparisons (among test bullets), and 125 known non-matched land comparisons. [may be summarized in a table, and for the specific counts, we will provide details of that when introducing the LAPD data set]

The results of the experiment show us how the database scenario with well established density estimations improve the over-all accuracy. As shown in Figure 4, the left boxplot represents the the individual case scenario where everything is calculated and predicted based on one questioned bullet and three test fired bullets. The right boxplot represents the database scenario where the similarity scores among one questioned bullet and three test fired bullets are evaluated based on a relevant database. Figure 4 is based on the matched bullets, in which case all the comparisons of lands are actually matches. For each boxplot, there are one hundred points (number of questioned bullets), and each point represents the proportion of correct predictions (i.e. matched lands) of one questioned bullet and the corresponding three test fired bullets based on total eighteen land-land comparisons as shown in y-axis. Figure 5 is formed similar to Figure 4, but all the land-land comparisons are non-matched comparisons based on non-matched bullets. And in Figure 5 the y-axis is the proportion of correctly predicted non-matched lands. The results of the experiments are very strong support of making use of a relevant database for automated algorithms. In Figure 4, the evaluation based on an individual set are a little better than using of the database because of its more aggressive and data specific way of calculating likelihood ratio and selection of thresholds. But that improvement is little comparing with the significant loss of the accuracy in correctly identifying the non-matched land-land comparisons. Figure 5 implies that it makes no sense of using the individual case calculation to do prediction for non-matched bullet comparisons, while the database scenario works much better and even better than the two results of the matched bullet case, which meets our expectation of a more concentrated KNM distribution than the KM distribution.

[boxplot for KM and KNM, two scenarios, error rates]

[overall confusion table]

[compare this section with nex section, for KM, they are pretty close]

[what about thresholds and likelihood ratios themselves]

In this paper, we use the CCF as the similarity score..

We evaluate the accuracy under each scenario by using a comparable set of tests with questioned bullets of known origin.

As test set we are using bullets from the LAPD bullet set: 626 barrels with four fired bullets each.

Should we consider a binomial decison rule to get a bullet comparison results? if that works well here...

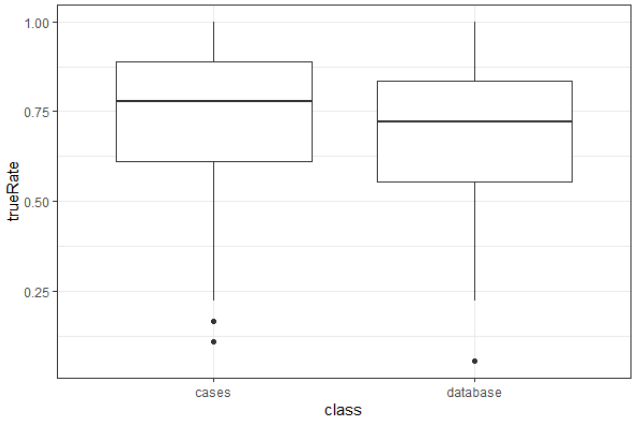

Figure 4: matched-km

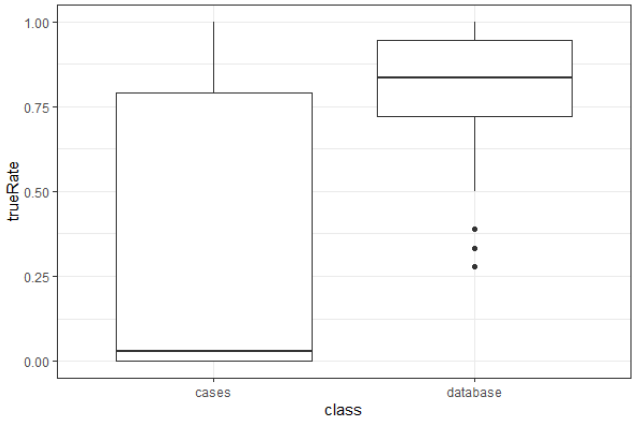

Figure 5: non-matched

6. Evaluate the error rates

- Given the iid assumption, could we develop a further model for the bullet level distribution and estimate the corresponding error rates?

Estimating the underlying theoretical error rates is a fundamental challenge in the forensic science (President's Council of Advisors on Science and Technology (PCAST), 2017, 2016). The estimation process becomes more straightforward after we finished the distribution estimation. There are many types of error rates can be defined. We focused on the four most used ones, i.e. 1) False positive rate (FPR), 2) False negative rate (FNR), 3) False identification rate (FIR), 4) False exclusion rate (FER), as defined in Table 3. FPR and FNR are considered source-specific assessment. They are both probabilities given

the ground truth to make false conclusions, which is useful especially when we evaluate an algorithm. FIR and FER are considered decision-specific assessment. They are both probabilities given the algorithm decisions to false claim the actual source, which is useful in real cases when jury is presented with the forensic conclusions to make judgment.

	Ground Truth		
	Match	Non-Match	Error Rate
Identification	Correctly judged as matches, True Positives (TP)	Incorrectly judged as matches, False Positive (FP)	False Identification Rate (FP/(TP + FP))
Elimination	Incorrectly judged as non-matches, False Negative (FN)	Correctly judged as non-matches, True Negative (TN)	False Exclusion Rate (FN/(TN + FN))
Error Rate	False Negative Rate (FN/(FN+TP))	False Positive Rate (FP/(FP + TN))	

Table 3: xxxx

The error rates are also affected by the threshold selected to make decisions. In an extreme example, if we are very conservative in making identifications, we can set the threshold to be large enough, say CCF >= 0.99, then the FPR will be 0. But at the same time, the FNR will be 1. It's clear that there are trade-offs between the error rates. So, to make these error rates comparable and consistent, for the current paper, we refer the threshold as the value where the probability densities of KM and KNM distributions equal with each other. This is not a bad choice, since it has good statistical meaning that at this value, the likelihood ratio equals 1 and any deviation from 1 would give preference to a certain decision (as if a threshold). For the two component beta mixture distributions we estimated in last section, this threshold is 0.529. In the application of the algorithms, the threshold need to be carefully chosen where the study of associated error rates will in turn be essential guiding that process.

The four theoretical error rates are directly associated with the distributions we estimated as shown in the Figure 6 and Figure 7. The estimated values of the error rates are reported in Table 4. The FPR is 0.148, while the FNR is 0.301 which is much larger. This means for the given threshold value 0.529, the algorithm has probability 0.148 to misclssify an actual different-sourced pair of land engraved areas (LEAs) into identifications. And the algorithm has a probability 0.301 to misclassify an actual same-sourced pair of land engraved areas (LEAs) into eliminations. The fact that FNR is larger than FPR implies that for the current threshold, we are more likely to misclassify an pair of KM LEAs instead of KNM LEAs. This is partly due to the fact the KM distribution is more spread than the KNM distribution. The FPR will decrease rapidly when increasing the threshold, while the FNR will be small only when the threshold is very low at great cost of FPR. And similar pattern can be observed from FIR

and FER. However, these error rates are explained in a totally different view. FIR is 0.212 means when the algorithm reports an identification (to the jury in court potentially), the algorithm could have made wrong claim at probability 0.212. As we aforementioned, the FPR and FNR are used to evaluate algorithms overall while the FIR and FER are used to evaluate a certain result reported.

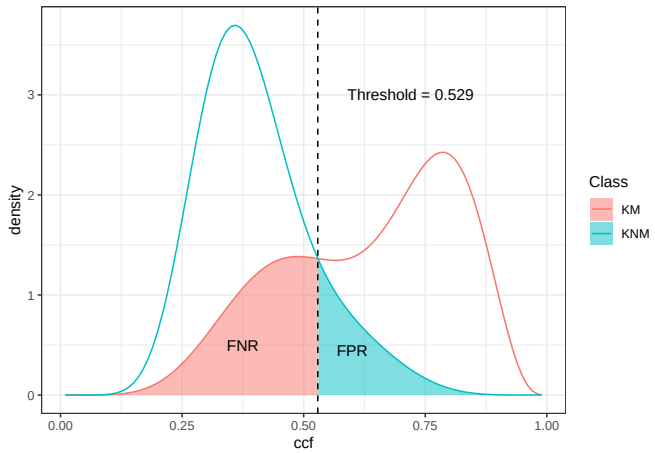

Figure 6: False Negative Rate(FNR) and False Positive Rate(FPR)

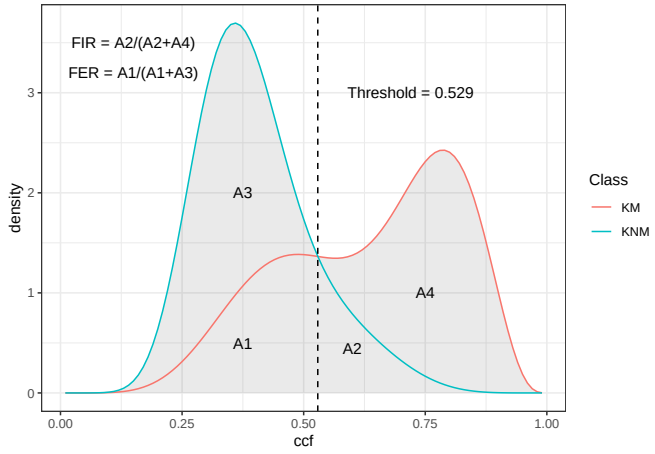

Figure 7: False Indentification Rate(FIR) and False Exclusion Rate(FER)

Overall, the error rates are not small enough to give confidence to any conclusions. But remember the current comparison discussed is in the land level. In real cases, we will need to combine the land level conclusions to bullet level ones. And this process of combining multiple pieces of evidence will improve the performance greatly. One of the common ways in computer-based bullet analysis is to calculate the sequential average maximum (SAM) [susan's paper and the

Type of Error Rates	Estimated Values
False positive rate (FPR)	0.148
False negative rate (FNR)	0.301
False identification rate (FIR)	0.212
False exclusion rate (FER)	0.354

Table 4: xxxx

original paper of SAM?] of land scores. Next, we are going to making use of the KM and KNM distributions in the land level to explore the error rates of the bullet level.

7. Estimations with changing sample size

To address the question that how many test fires we need to reach good estimations of the distributions in a database and finally, a good classification result with controlled error rates, we will investigate the estimations and error rates in a changing sample size setting. There are two ways we might evaluate the estimations. First, from a statistical view, we can quantify the variation of the estimated curves, and find the minimum requirement to make the estimation converge. From a practical view, we look at the error rate and find a sample size when the error rates converge. Then, we can make sure to get reliable estimations of the densities and reliable performance of making predictions with likelihood ratios.

We design a study to evaluate the effect of data sizes. As in the previous sections, we are using the LAPD data sets. And set the unit of data size as a set of four same source bullets as in the case of LAPD and in general, the way of forming a data base. It's worth noting that any meaningful statistical results rely on replications. But we must be careful that as the used sample size in a single trial increases, the replications can require very large data sets. Otherwise, the heavy reuse of data could cause high dependence among the results and lead to false conclusions. We are using a approach that is adopted in machine learning field of dealing with this issue. Instead of using different samples at each fixed sample size to achieve replications, we are adding new data to the existing pool of used data to achieve replications by sequentially fitting the model. This also meets the sequential convergence pattern in a statistical sense perfectly. When the addition of new data doesn't significantly change some adopted measurements of our goal, then we find the minimum number of sample size reaching that stage. And additionally, we also repeat the above process with ten randomly selected sequences, which further investigates the data specific effect in the process and it is harder to see all the sequences converged uniformly to a same reasonable small interval. So we also introduce different levels of convergence to formulate the conclusions.

Figure 8 shows the Bhattacharyya distance between each estimated distribution and the full data distribution in both KM and KNM cases for the ten

sequence of randomly selected data. Bhattacharyya distance is associated with the the amount of overlap of two distributions. The lower it is, the more overlapped area there is for the two distributions. From Figure 8, the Bhattacharyya distance decreases rapidly with the increase of the size of the used data at the beginning, and it also reveals that those those distributional estimations with one or two sets of data just don't make sense in producing a stable results. But as the sample size increases to about 5 sets of data, the estimations are pretty stable and close the theoretical distributions, and when it increase to about 40 sets of data, the the estimations essentially converged. It's worth noting that the KNM distribution estimations do have outliers in the early stage, which could be the result of the using of the two-component distributions since the two-component distributions are better with large data as in the previous analysis but with less data, this improvement could be less the cost of estimating a more complex model.(So we will also need to check the single beta) So for the KNM distribution, we might need more data to achieve stable estimations of the two-components distributions, but we do have smaller distance for KNM than KM for estimations with more than 40 sets of data even though both of them are small.

Figure 9 shows the thresholds we calculated from using the pair of KM and KM distributions with increasing sample size and the red line is the threshold we calculated with the full data distributions (considerred as theoretical ones). Similar to the case of Bhattacharyya distance, the thresholds are restricted to a small interval after the extreme variation in the first couple of estimations. This is good since we can quickly reach the point to get reasonable thresholds. But it also reveals that the further convergence of the thresholds are difficult. Unlike the Bhattacharyya distance, the thresholds will converge to a smaller interval approximately from 0.52 to 0.55 but not a point. However, if we focus on each sequence (line), we can find that after 60 sets of data, they all almost converged to a point. This reveals the different level of convergence, i.e. the convergence of estimations for a particular data base and the convergence of estimations of different databases. The latter is obviously more demanding. But finally, these are all in a reasonable small interval. Another point worth pointing out is that the thresholds seem to require more data than the Bhattacharyya distance. It is expected that the statistical convergence (Bhattacharyya distance) is more strict than the practical convergence (the thresholds), but it is not the case. Since the thresohlds are calculated base on both distributions, the convergence of the thresolds depends on the convergence of both KM and KNM distributions. And it enlarges the variation.

Figure 10 shows the false negative error rates and the false positive error rates. Similar to the thresholds behavior, the error rate has two stages of converges at about 5 sets and 60 sets of data. But in this figure, the y-axis has the exact practical meaning of error rates and we can quantify how large that variation related to the sample sizes means to us. And after 60 sets of data, the error rates are in reasonably small intervals and the cost of further improvement is high. And we also stress the point of different levels of convergence, a database with 60 sets of data can give predictions almost converged to a number and tens of

sets of additional data won't change that significantly. The further convergence of different databases to a same number is more demanding and pretty slow. The range of the interval after 60 sets of data is about 0.045 which means that the most extreme difference of the error rates achieved with different database is 0.04 (larger in false positive and smaller in false negative error rates) for each land to land comparison.

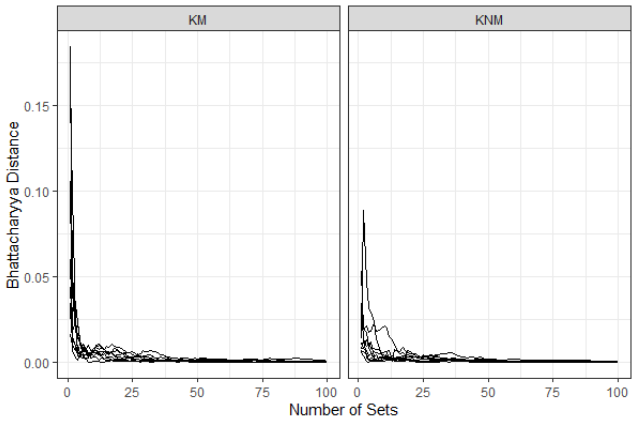

Figure 8: Bhattacharyya Distance

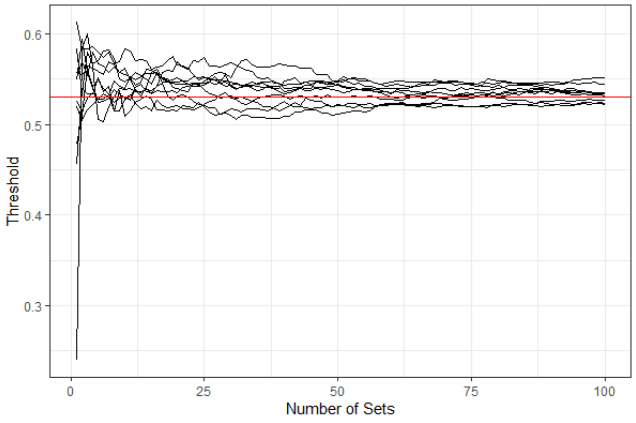

Figure 9: Threshold

To conclude the section, we might ask for 5 sets of bullets as the minimum requirement of a meaningful database. Further, 60 sets of bullets could significantly improve the performance. But further improvement can be very expensive.

(try different model for KNM? like one component?)

(In the study of the both types of convergence, we also study the behaviors of the two components and evaluate the functional form of the distributions as

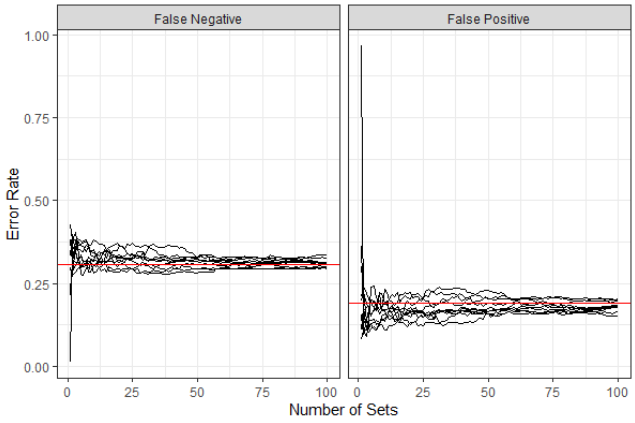

Figure 10: Error

well.)

(The convergence of thresholds depends on the convergence of both curves, more demanding)

(We will reproduce the some results in the previous sections: 1. application section?)

8. Conclusion

We did a distributional analysis of the similarity scores to guide the use of likelihood ratios in quantifying the strength of forensic evidence. In particular, we picked the CCF as a typical similarity score and established two-component beta mixture distributions as the theoretical forms for both known-matched and known-nonmatched land-land comparisons. Then we showed how effectively the use of those theoretical forms from a relevant population could help in improving the performance of the prediction in forensic practice with likelihood ratios. The false negative error rates based on individual cases are a little bit lower than the ones based on the relevant database, but it comes with extreme false positive error rates. As expected, the predictions based on individual cases face the problem similar to overfitting issues. This result strongly supports using a relevant database for the relevant population in likelihood ratio analyses for forensic science. We analyzed the associated underlying theoretical error rates in using likelihood ratios with the distributional forms we established. Ideally, we would expect to achieve those error rates if the relevant population is accessible. Thus, we could theoretically control the error rates in the practice. Further, we analyzed the amount of data necessary to establish a reasonable database for the relevant population, and finally achieve the target error rates. We introduced different level of convergence for the required data sizes. Intuitively, the requirement for a particular database to produce stable results is lower than the requirement of all databases to produce a same stable result. And we could conclude that for a particular database, the ideal size would be 60 sets of bullets.

References

AFTE Criteria For Identification Committee, 1992. Theory of identification, range striae comparison reports and modified glossary definition afte criteria for identification committee report. AFTE Journal 24, 336–340.

Chen, Z., Chu, W., Soons, J.A., Thompson, R.M., Song, J., Zhao, X., 2019. Fired bullet signature correlation using the Congruent Matching Profile Segments (CMPS) method. Forensic Science International 305, 109964.

Chumbley, L.S., Morris, M.D., Kreiser, M.J., Fisher, C., Craft, J., Genalo, L.J., Davis, S., Faden, D., Kidd, J., 2010. Validation of tool mark comparisons obtained using a quantitative, comparative, statistical algorithm. Journal of Forensic Science 55, 953–961.

Fadul Jr., T., Hernandez, G., Stephanie, S., Sneh, G., 2011. Empirical Study To Improve the Scientific Foundation of Forensic Firearm and Tool Mark Identification Utilizing 10 Consecutively Manufactured Slides. National Institute of Justice.

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

Krishnan, G., Hofmann, H., 2019. Adapting the chumbley score to match striae on land engraved areas of bullets. Journal of Forensic Sciences 64, 728–740. doi:10.1111/1556-4029.13950

Ma, L., Song, J.-F., Whitenton, E.P., Vorburger, T.V., Zhou, J., Zheng, A., 2004. NIST Bullet Signature Measurement System for RM (Reference Material) 8240 Standard Bullets 49(4).

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

Nichols, R., 2003. Firearm and toolmark identification criteria: A review of the literature, part II. Journal of forensic sciences 48, 318–27. doi:10.1520/JFS14149J

Nichols, R.G., 1997. Firearm and Toolmark Identification Criteria: A Review of the Literature. Journal of Forensic Science 42, 466–474. doi:10.1520/JFS14149J

President's Council of Advisors on Science and Technology (PCAST), 2016. Report to the president, forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods. Executive Office of the President of the United States, President's Council ....

President's Council of Advisors on Science and Technology (PCAST), 2017. An Addendum To The PCAST Report On Forensic Science In Criminal Courts.

Song, J., 2015. Proposed "Congruent Matching Cells (CMC)" Method for Ballistic Identification and Error Rate Estimation 47, 9.

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

Vanderplas, S., Nally, M., Klep, T., Cadevall, C., Hofmann, H., 2020. Comparison of three similarity scores for bullet lea matching. Forensic Science International 308. doi:10.1016/j.forsciint.2020.110167

Vorburger, T.V., Song, J.F., Chu, W., Ma, L., Bui, S.H., Zheng, A., Renegar, T.B., 2011. Applications of cross-correlation functions. Wear, The 12th International Conference on Metrology and Properties of Engineering Surfaces 271, 529–533. doi:10.1016/j.wear.2010.03.030

Zhang, N.F., 2019. The use of correlated binomial distribution in estimating error rates for firearm evidence identification. Journal of Research of the National Institute of Standards and Technology 124. doi:10.6028/jres.124.026