Using correlated binomial distribution in estimating error rates for forensic firearm identification

Nien Fan Zhang

National Institute of Standards and Technology, 100 Bureau Dr,
Gaithersburg, MD 20899

Abstract

Estimating error rates for firearm evidence identification is a fundamental challenge in forensic science. The recently developed Congruent Matching Cells (CMC) method provides applications to firearm evidence identification. To estimate error rates, appropriate statistical models are needed for the CMC values. In this paper, in addition to the binomial probability distribution the correlated binomial distribution is proposed. For an image comparison, correlated binomial distribution can be applied to the cell pairs from CMC method. An application to an actual data set demonstrates that the correlated binomial distribution fits the relative frequency distribution of CMC values much better than the binomial distribution.

Key Words: Ballistic signatures, Bernoulli trials, forensic science, maximum likelihood estimate, nonlinear regression

1. Introduction

In firearm evidence identification, when bullets and cartridge cases are fired or ejected from a firearm, the parts of the firearm create characteristic tool marks called ballistic signatures. In general, tool marks have so called “class characteristics” that are common to certain brands of firearms and individual characteristics arising from random variation in firearm manufacturing. Recently, a quantitative approach known as the Congruent Matching Cells (CMC) method was developed to improve the accuracy of ballistic identifications and provide a base to estimating error rates [1]. This paper proposes statistical models and the corresponding methodology for estimating the model parameter and error rates. In Section 2, the CMC method and the proposed parameter estimation based on the binomial distribution is described. In Section 3, the correlated binomial distribution based on dependent Bernoulli trials is introduced. In Section 4, maximum likelihood estimators of the parameters of correlated binomial distribution is proposed and applied to an actual data set. In Section 5, the correlated binomial is combined with the beta distribution to have a compound probability distribution called the beta-correlated binomial distribution followed by the conclusions.

2. CMC methods for ballistic identification

The CMC method deals with pairs of measured 2D or 3D topography images of breech face impressions whose similarity we wish to quantify. The CMC method divides each image into a rectangular array of cells. For each cell, a search for a matching cell is then made on a compared image [2]. Figure 1 shows the correlated and uncorrelated cell pairs in an image pair. The cell-by-cell analysis is done to determine whether each cell pair is a correlated call pair or not.

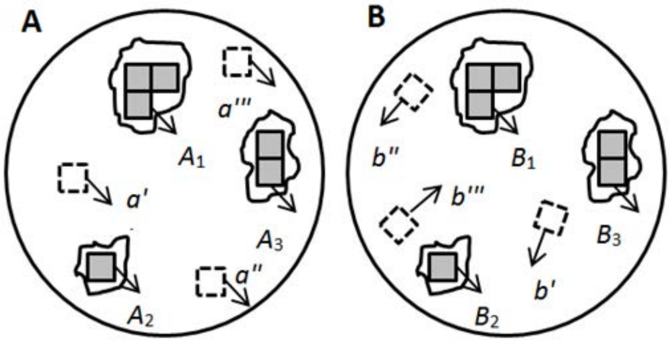

Figure. 1. Schematic diagram of topographies A and B originating from the same firearm and registered at the position of maximum correlation. The six solid cell pairs in each image are located in three valid correlated regions (A_1, B_1), (A_2, B_2), and (A_3, B_3). The dotted cell pairs (a', b'), (a'', b''), and (a''', b''') are located in the invalid correlation region [2].

Congruent matching cell pairs, or CMCs are determined by three sets of identification parameters for quantifying both the topography similarity of the correlated cell pairs and the pattern congruency of the cell distributions. From a statistical point of view, the CMC method is based on pass-or-fail tests of individual cell pairs comprising an image pair of breech face impression. For a pair of images of breech face impressions, N represents the number of correlated cell pairs in the image pair. For a given correlated cell pair, a random variable X represents the outcome of the CMC method applied to the correlated cell pair. When the CMC method determines that the cell pair is a congruent matching cell pair, e.g., (A_1, B_1), then X = 1; otherwise X = 0. Symbol P represents probability in general and the symbol p represents the probability that X = 1. That is $P(X=1)=p$, and $P(X=0)=1-p$.

An approach was developed in [2] for estimating the expected error rates of ballistic identifications based on the CMC method. Error rates are discussed in detail in [2]. To estimate error rates, a key is to find the best probability distribution for the relative frequency distribution of the observed CMC values.

3. The binomial distribution based on dependent Bernoulli trials

In [2], for CMC method, the random variable X is assumed to be a Bernoulli trial. Namely, the trials results $\{X_1 ..., X_N\}$ are N dichotomous items. Namely, the comparisons between cell pairs are independent from each other and with a common probability p. Denote the sum of the CMC values for the comparisons of the first image pair by Y_1 with N cell pairs. $Y_1 = \sum_{i=1}^{N} X_{1i}$. In probability, $Y_1 \sim Bin(N, p)$ is a binomial distributed random variable (r.v.) with the probability mass function

$$P_{[1]}(Y = k) = C_N^k p^k (1-p)^{N-k} \text{ for } k = 0,1,...,N \tag{1}$$

Similarly, for the M image pairs, we have $Y_1,...Y_M$ correspondingly. When $\{Y_j \, j = 1,...,M\}$ are independent from each other, we have a sequence of independently binomial distributed r.v.'s. That is, $Y_j = \sum_{i=1}^{N} X_{ji}$ and $j = 1,...,M$ and $Y_j \sim Bin(N, p)$. For observed values of $\{y_j, j = 1,...,M\}$ the maximum likelihood estimator of p is given by $\hat{p} = \sum_{j=1}^{M} y_j \Big/ MN$.

However, the assumption of independence among cell pair comparisons may be invalid. For various reasons, for example, the physical similarity between the cell pairs may lead that the comparisons considered are not be statistically independent with each other in general. In addition, two cell pairs may have a duplicate cell, for example, (A_1, B_1) vs. (A_2, B_4). In this case, we consider a model for dependent Bernoulli trials proposed by Bahadur [3], which sometimes is called Bahadur-Lazarsfed model. Similar to the Bernoulli trials, for a sequence of $\{X_1 ..., X_N\}$, with each X_i equal to 0 or 1 with $P(X = 1) = p$, and $P(X = 0) = 1 - p$ for $i = 1,..., N$. However, $\{X_1 ..., X_N\}$ may not be mutually independent. We define the second order correlation between X_i and X_j by

$$r_{ij} = \frac{\mathrm{Cov}[X_i, X_j]}{\sigma^2} = \frac{E[(X_i - \mu)(X_j - \mu)]}{\sigma^2}, \tag{2}$$

where μ is the marginal mean and σ is the marginal standard deviation of X_i. Higher order correlations are similar. For simplicity, we only consider the second order correlation and assume that the correlations are symmetric [3]. In this case, $r_{ij} = r(2)$ for $i, j = 1,..., N$. As discussed in [3], the probability mass function of the sum of $\{X_1 ..., X_N\}$ denoted by Y can be approximated by

$$P_{[2]}(Y) = P_{[1]}(Y)[1 + r_{(2)} g_2(Y, p)] \tag{3}$$

where $r(2)$ is the second order correlation. $P_{[1]}(Y)$ is the probability mass function when $\{X_i, i = 1,..., N\}$ is a sequence of Bernoulli trials given in (1), and the function $g_2(Y, p)$ is a second order polynomial in Y,

$$g_2(Y) = \frac{(Y - Np)^2 - (1-2p)(Y - Np) - Np(1-p)}{2p(1-p)}. \tag{4}$$

In this case, we say the r.v. Y has a correlated binomial distribution. The details of $g_2(Y,p)$ in (4) and the case for higher order, for example the third order approximation can be found in [4]. Figure 2 shows the probability mass functions of $P_{[1]}(Y)$ and $P_{[2]}(Y)$ with $N=26$, $p=0.6$, and $r(2)=0.02$. The two r.v.'s have the same mean = 15.6 while the r.v. with a correlated binomial distribution has a variance = 9.36, which is larger than the variance of 6.24 for the r.v. with the binomial distribution.

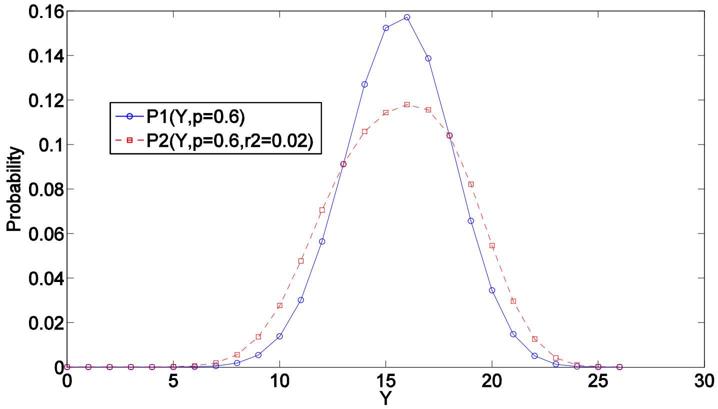

Figure 2. Probability mass functions for binomial (blue) and correlated binomial distribution (red) with $N=26$.

4. Estimating the parameters of the correlated binomial distribution

When the CMC method applies to a set of cartridge cases, the result, in general, includes certain known matching (KM) image pair comparisons and certain known non-matching (KNM) image pair comparisons. In [2], statistical models are fitted to the cases of KM and KNM, respectively. In either case, we assume that for M image pairs, the random variables for the sums of the CMC values for each image comparison are denoted by $Y_1,...,Y_M$. As discussed in Section 3, we assume that $Y_1,...,Y_M$ are independent from each other while for each image comparison, we have a sequence of N dependent Bernoulli trials. Maximum likelihood is used to estimate the parameters of the correlated binomial distribution. The likelihood function for given p and $r(2)$ is given by

$$L=\prod_{i=1}^{M} p_{[2]}(y_i \mid p,r(2))=\prod_{i=1}^{M} C_N^{y_i} p^{y_i}(1-p)^{N-y_i}[1+r_{(2)}g_2(y_i,p)]. \qquad (5)$$

The maximum likelihood estimator (MLE) of p and $r(2)$ are obtained when the respected $\log(L)$ reaches the maximum. We evaluated the models on a set of cartridge cases created by Weller et al. [5]. The cartridge cases were obtained from a set of eleven slides produced by the same manufacturer. The data set includes 370 KM image pairs. For illustration, based on the KM data set with $N=47$, the MLE of the parameters of the correlated binomial distribution are obtained with $\hat{p}=0.7823$ and $r_{(2)}=0.0191$. In this case, For comparison, the parameter p for the binomial distribution is estimated by 0.7864. Figure 3 shows the relative frequency distribution of the observed CMC numbers and the probability mass functions for the binomial and correlated binomial distributions with the corresponding estimates of the parameters, respectively. It is clear that the correlated binomial is a much better fit to the CMC values than that for the binomial distribution.

In addition, nonlinear regression models can also be used to estimate the parameters of correlated binomial distributions [4].

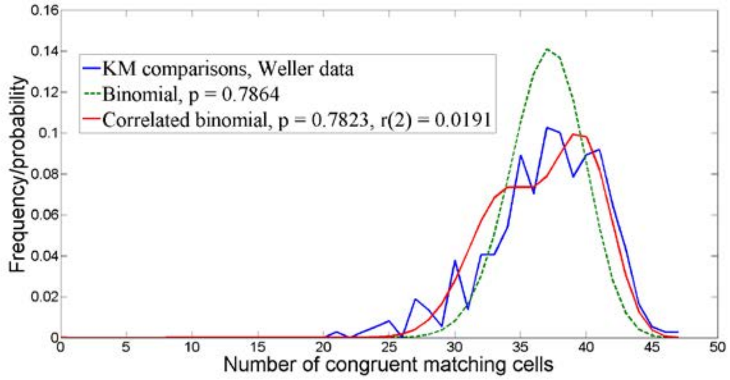

Figure 3. Relative frequency distribution in blue of CMC numbers for KM image pairs and the green and red curves represent the binomial and correlated binomial probability mass functions for the KM data

5. Use of the beta-correlated binomial distribution for CMC measurements

In [2], it was proposed to relax the assumption of a fixed probability of congruency for the binomial distribution when modeling the CMC measurements. This revised model allows one to vary p for different image pair comparisons. In this case, we assume that within one image pair comparison, the probability p for all the Bernoulli trials is the same while for different image pair comparisons, p varies. See [6]. As in the framework of Bayesian statistics, we assume that the parameter p is a random variable with a beta distribution. For the first image pair with N cell pairs, we have a sequence of Bernoulli trials: $X_{11},..., X_{1N}$, which are independent from each other and have a common probability of $p = p_1$. The sum of the CMC values for the first image pair is Y_1, which for given p_1 has a binomial distribution. Namely, $Y_1 \mid p_1 \sim Bin(N, p_1)$. In general, for M image pairs, we have $Y_i \mid p_i \sim Bin(N, p_i),\ i = 1,..., M$, where p has a beta distribution, i.e., $p \sim Beta(\alpha, \beta)$ with positive α and β as parameters to be fitted to the data. The probability mass function of the beta-binomial random variable Y for given N, α, and β is then given by

$$\begin{aligned} P(Y = k \mid N, \alpha, \beta) &= \int_0^1 \frac{P_{[1]}(k, p)}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1} dp \\ &= \frac{C_N^k}{B(\alpha, \beta)} \int_0^1 p^{k+\alpha-1}(1-p)^{N-k+\beta-1} dp \qquad (6) \\ &= C_N^k \frac{B(k+\alpha, N-k+\beta)}{B(\alpha, \beta)}, \end{aligned}$$

where $B(\alpha, \beta)$ is a beta function with parameters α and β , and $P_{[1]}(k, p)$ is the binomial probability mass function in (1) when $Y = k$.

In [2], comparisons of the fits of the beta-binomial probability model and the binomial probability model for data sets including the Weller data set for the cartridge cases were made. Here we need to emphasize that although using the beta-binomial distribution can relax the assumption for the same p for all image pairs, it still assumes that within each image pair, all cell pair comparisons are independent from each other. We check this assumption by considering correlations among cell pair comparisons.

Now instead of the independent Bernoulli trials, we assume that the cell pair comparisons within each image pair are dependent Bernoulli trials. The corresponding probabilities of the sum are approximated by $P_{[2]}(Y)$ as given by (3) when only the 2nd order correlation with a constant $r_{(2)}$ is assumed. Assume that p in the correlated binomial distribution is random with a beta distribution. Namely, $Y_i \mid p_i, r_{(2)} \sim corr.Bin(N, p_i, r_{(2)})$, $i = 1,...,M$ where p has a beta distribution, i.e., $p \sim Beta(\alpha,\beta)$. Similar to (6), the probability mass function of Y for given N, α, β, and $r_{(2)}$ is given by

$$\begin{aligned} &P(Y=k \mid N,\alpha,\beta,r_{(2)}) \\ &\quad = \int_0^1 \frac{P_{[2]}(k,p)}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}dp \\ &\quad = \frac{C_N^k}{B(\alpha,\beta)} \int_0^1 p^k(1-p)^{N-k}\{1+r_{(2)}g_2(k,p)\}p^{\alpha-1}(1-p)^{\beta-1}dp \\ &\quad = \frac{C_N^k}{B(\alpha,\beta)} \int_0^1 p^{k+\alpha-1}(1-p)^{N-k+\beta-1}\{1+r_{(2)}g_2(k,p)\}dp, \end{aligned} \tag{7}$$

where $g_2(k,p)$ is given by (4). In this case, the marginal probability $P(Y=k \mid N,\alpha,\beta,r_{(2)})$ for $k = 0,1,...,N$ has no explicit expression. However, it can be calculated by numerical integration. In this case, the random variable Y has a compound probability distribution called a beta-correlated binomial distribution.

6. Conclusions

Estimating error rates is an important part for firearm identifications. To evaluate error rates, a key is to determine the appropriate statistical model for the CMC values for image comparisons. The proposed correlated binomial distribution is reasonable for the settings of the actual image comparisons and demonstrates a good fit to the CMC data.

References

[1] Song J. 2015. Proposed “Congruent matching cells (CMC)” method for ballistic identification and error rate estimation, *AFTE J.* 47 (3). 177-85.

[2] Song J., T.V. Vorburger, W. Chu, J. Yen, J. A. Soons, D. B. Ott, and N. F. Zhang. 2017. Estimating error rates for firearm evidence identification in forensic science, *Forensic Science International.* (284). 15-32.

[3] Bahadur R. R. 1961. A representation of the joint distribution of the response to n dichotomous Items. In: H. Solomon (Ed.), *Studies in Item Analysis and Prediction* (Stanford: Stanford University Press). 158-68.

[4] Zhang, N. F. 2019. The use of correlated binomial distribution in estimating error rates for firearm evidence identifications, to be published in *Journal of Research of the National Institute of Standards and Technology*.

[5] Weller T. J., A. Zhang, R. Thompson, and F. Tulleners. 2012. Confocal microscopy analysis of breech face marks on fired cartridge cases from 10 consecutively manufactured pistol slides, *J. Forensic Sci*. 57 (4). 912-17.

[6] Wilcox R. R. 1981. A review of the beta-binomial model and its extensions, *Journal of Educational Statistics*. 6(1). 3-32.