Forensic Science International 326 (2021) 110912

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Statistical models for firearm and tool mark image comparisons based on the congruent matching cells (CMC) method

Nien Fan Zhang

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

ARTICLE INFO

Article history:
Received 23 January 2021
Received in revised form 1 July 2021
Accepted 14 July 2021
Available online 20 July 2021

Keywords:
Binomial distribution
Correlated binomial distribution
Error rates estimation
Maximum likelihood estimator
Nonlinear regression

ABSTRACT

In the branch of forensic science known as firearm evidence identification, various similarity scores have been proposed to compare firearm marks. Some similarity score comparisons, for example, congruent matching cells (CMC) method, are based on pass-or-fail tests. The CMC method compares the pairwise topography images of breech face impressions, from which the similarity score is derived for quantifying their topography similarity. For an image pair, the CMC method determines a certain number of correlated cell pairs. Next, each correlated pair is determined to be a congruent match cell (CMC) pair, or not based on several identification parameters. The number of CMC pairs as a threshold is required so that the two images of surface topographies can be either identified as matching or determined to be non-matching. To reliably estimate error rates or evaluate likelihood ratio (LR), the key is to find an appropriate probability distribution for the frequency distribution of the observed CMC results. This paper discusses four statistical models for CMC measurements, which are binomial and three binomial-related probability distributions. In previous studies, for a sequence of binomial distributed or other binomial-related distributed random variables (r.v.), the number of Bernoulli trials N for each r.v. is assumed to be the same. However, in practice, N(the number of cell pairs in an image pair) varies from one r.v. (or one image pair) to another. In that case, the term, *frequency function,* of the CMC results is not appropriate. In this paper, the generalized frequency function is introduced to depict the behavior of the CMC values and its limiting distribution is provided. Based on that, nonlinear regression models are used to estimate the model parameters. The methodology is applied to a set of actual CMC values of fired cartridge cases.

Published by Elsevier B.V.

1. Introduction

For firearm image comparisons based on breech face impressions, the parts of the firearm that make forcible contact with the bullets or cartridge cases when fired create characteristic tool marks or ballistic signatures on their surface. These signatures can be used for firearm evidence identifications [1]. In general, tool marks have so called class characteristics that are common to certain brands or models of firearms and individual characteristics arising from random variation in firearm manufacturing and wear. Fig. 1 (from Ref. [2]) shows topography images of breech face impressions obtained from a pair of cartridge cases ejected from the same firearm. The image pair has several features in common.

In the area of firearm image comparisons, various similarity scores have been proposed to compare firearm marks, for example, the value of the normalized cross-correlation function for a pair of images [2,3], and the value from the congruent matching cells (CMC) method [2]. From a statistical point of view, both these similarity scores are based on pass-or-fail tests. In this paper, we focus on the CMC method, which deals with pairs of topography images of breech face impressions for which the similarity or matching has been quantified. For a topography image pair, the CMC method divides one image, which is designated to be a reference image, into an array of rectangular cells as shown in Fig. 2 (from Ref. [2]). For each reference cell, a search for a matching cell is then conducted on the compared image to yield a correlated cell pair. Whether the cell pair is correlated or not is determined by the maximum cross-correlation function [2].

Next, each correlated pair is determined to be a congruent match cell (CMC) pair, or not based on several identification parameters. A certain number of CMC pairs as a threshold is required so that the two images of surface topographies can be identified as matching. Some discussion on how to determine a threshold for CMC comparisons can be found in Section 2 of [2].

E-mail address: nien-fan.zhang@nist.gov.

https://doi.org/10.1016/j.forsciint.2021.110912
0379-0738/Published by Elsevier B.V.

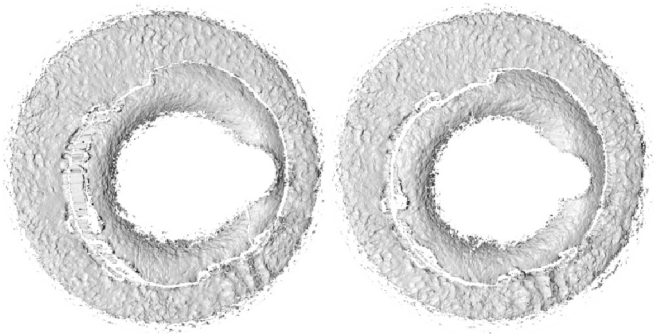

Fig. 1. Topography images of breech face impressions obtained from a pair of cartridge cases fired from the same firearm.

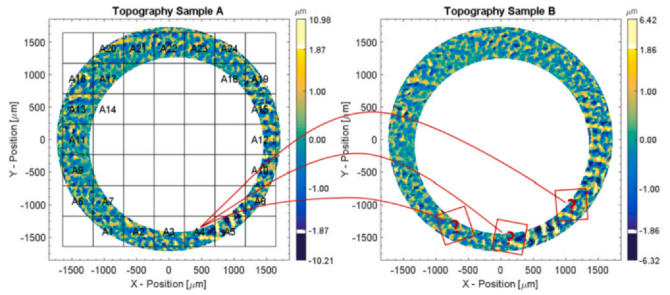

Fig. 2. Conceptual diagram of a topography image overlaid by a 7 × 7 grid, dividing the image into cells. Only cells with a sufficient fraction of measured pixels are used. Also shown is a schematic diagram of the automated search procedure to find an area in the compared image (right) that has a strong correlation with one of the cells in the reference image (left).

The error rates for firearm evidence identification have been discussed [2,4,5]. The CMC method is based on the pass-or-fail tests of individual cell pairs from an image pair of breech face impressions. When the CMC method is applied to a set of cartridge cases, the result, in general, includes certain known matching (KM) image pairs and certain known nonmatching (KNM) image pairs. The false positive error rate is the rate of pairs of images incorrectly judged as matches. It represents the expected frequency or probability of obtaining an erroneous result of identification (declared match) when the comparing samples are from different sources (KNM). On the other hand, the false negative error rate is the rate of pairs of images incorrectly judged as nonmatches. It represents the probability of obtaining an erroneous result of exclusion (declared nonmatch) when the comparing samples are from the same source (KM).

To reliably estimate error rates, the key is to find an appropriate probability distribution for the frequency distribution of the observed CMC values. In addition, the importance of the choice of an appropriate statistical model in evaluating likelihood ratio (LR) for firearm evidence identifications is also emphasized in [5]. In Section 2, four statistical models for CMC values and their properties are discussed. In Section 3, estimations of the parameters of the statistical models for the case of image pairs with variable cell pair number, in particular the use of nonlinear regression models are discussed followed by discussions and conclusions in Section 4.

2. Statistical models for CMC values

2.1. Binomial distribution

A binomial probability distribution was proposed in [2] for the distribution of CMC values from an image pair comparison. Two assumptions are made: (1) The comparisons between cell pairs in an image pair are statistically independent from each other, (2) each cell pair comparison within an image pair has the same matching probability, p. Under these assumptions, for a given cell pair, a random variable (r.v.) X represents the outcome of the CMC method applied to the cell pair. When the CMC method determines that a cell pair is a matching cell pair, then $X = 1$; otherwise, $X = 0$. Denote the probability that $X = 1$ by p. That is, $P(X = 1) = p$, and $P(X = 0) = 1 - p$. For an image pair with N cell pairs, we have a sequence of Bernoulli trials, $X_1, .,X_N$ [6]. The sum $Y = \sum_{i=1}^{N} X_i$ is the number of the congruent matching cell pairs for the image pair. Under these two assumptions, $Y \sim Bin(N, p)$ is a binomially distributed r.v. with the probability mass function (pmf) given by

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

where $\binom{N}{k} = N!/[k!(N - k)!]$. We discuss the error rates defined in Introduction. The false positive error rate represents the probability that an image pair is declared to be a match when the image pair is actually from a set of known nonmatching (KNM) image pairs. For the KNM set, denote the probability of the outcome of the CMC method applied to a cell pair in an image pair in the KNM set to be matching by p_{KNM}. That is, for the r.v., X, $p_{KNM} = P(X = 1)$ for the KNM set. Denote the number of CMC cell pairs (Y) as a threshold for which the image pair is identified as matching by $C (C \geq 0)$. Namely, when $Y \geq C$, the corresponding image pair is determined to be matching. Assuming for a particular image pair from the KNM set, there are N_0 cell pairs in total. Denote the number of CMC cell pairs in this image pair by Y. As stated in Introduction, the false positive error rate is the probability of pairs of images incorrectly judged as matches. Namely, the false positive error rate E_1 for this image pair is given by

$$\begin{aligned} E_1 &= P(C \leq Y \leq N_0 | \text{the image pair} \in \text{KNM set}) \\ &= \sum_{k=C}^{N_0} P(Y = k | \text{the image pair} \in \text{KNM set}). \end{aligned} \tag{2}$$

In particular, when $Y \sim Bin(N_0, p_{KNM})$, from (1)

$$\begin{aligned} E_1 &= \sum_{k=C}^{N_0} P_{[1]}(Y = k) \\ &= \sum_{k=C}^{N_0} \binom{N_0}{k} p_{KNM}^k (1 - p_{KNM})^{N_0-k}. \end{aligned} \tag{3}$$

In practice, the probability p_{KNM} is unknown. Under the assumption of a binomial distribution of CMC values for the KNM set, a maximum likelihood estimate (MLE) $\hat{p}_{KNM}$ of p_{KNM} is obtained (from (11)) when the CMC values are from the KNM set. From (3), an estimate of false positive error rate is obtained by replacing the probability in (3) by its estimate.

Conversely, the false negative error rate represents the probability that an image pair is declared to be nonmatching when the image pair is actually from a set of known matching (KM) image pairs. For a particular image pair with N_0 cell pairs in total, similar to (2), the false negative error rate E_2 is given by

$$\begin{aligned} E_2 &= P(0 \leq Y < C | \text{the image pair} \in \text{KM set}) \\ &= \sum_{k=0}^{C-1} P(Y = k | \text{the image pair} \in \text{KM set}). \end{aligned} \tag{4}$$

For a KM set, denote the probability of the outcome of the CMC method applied to a cell pair to be matching by p_{KM}. That is, for the r.v., X, $p_{KM} = P(X = 1)$ for the KM set. For a particular image pair from the KM set with N_0 cell pairs in total, the CMC number, $Y \sim Bin(N_0, p_{KM})$. Similar to (3),

$$E_2 = \sum_{k=0}^{C-1} P_{[1]}(Y=k) = \sum_{k=0}^{C-1} \binom{N_0}{k} p_{KM}^k (1-p_{KM})^{N_0-k}. \quad (5)$$

Similar to the false positive error rate, an estimate of the false negative error rate can be obtained by replacing p_{KM} by an estimate $\hat{p}_{KM}$ based on the CMC values from the KM set.

2.2. Correlated binomial distribution

For the binomial distribution, the assumption of mutual independence among cell pair comparisons in an image pair, i.e., $\{X_1, .,X_N\}$, however, is most likely invalid in practice. For example, since the array of cells laid over the image is not precisely aligned with features in the image, neighboring cell pairs may share some individualizing features. Dependence among those neighboring cells is more likely to increase compared to cells situated relatively far apart. Thus, we need to consider some type of dependent Bernoulli trials for cell pair comparisons. Reference [7] proposes a general model for dependent Bernoulli trials, which sometimes is called Bahadur-Lazarsfeld model. The resultant distribution of $Y = \sum_{i=1}^{N} X_i$ is called a correlated binomial distribution. Reference [8] provides a comprehensive discussion on the correlated binomial distribution and the corresponding parameter estimation.

Consider a sequence of symmetric dependent Bernoulli trials, $\{X_1, .,X_N\}$ [8], where each X_i takes the value 0 or 1, with $P(X_i = 1) = p$, and $P(X_i = 0) = 1 - p$ for $i = 1, .,N$. Note that and $\mathrm{Var}[X_i] = p(1-p)$ for $i = 1, .,N$. The second-order correlation between X_i and X_j, where $j \neq i$, is given by $r_{(2)} = \mathrm{Cov}[X_i, X_j]/\sigma^2$ for $i = 1, .,N$, where $\mathrm{Cov}[X_i, X_j]$ is the covariance between X_i and X_j, and $\sigma = \sqrt{p(1-p)}$ is the standard deviation of X_i for $i = 1, .,N$. Similarly, the third and higher-order correlations, up to the Nth order correlation, are defined. The sum $Y = \sum_{i=1}^{N} X_i$ has a correlated binomial distribution with pmf $P(Y)$. From [8], $P(Y)$ can be approximated by the second order approximation given by

$$P_{[2]}(Y) = P_{[1]}(Y)\{1 + r_{(2)} g_2(Y, p)\} = \binom{N}{Y} p^Y (1-p)^{N-Y}\{1 + r_{(2)} g_2(Y, p)\}, \quad (6)$$

where $P_{[1]}(Y)$ is the pmf of Y when $\{X_i\}$ are independent from each other as given in (1). Namely, $P_{[1]}(Y)$ is the pmf of the corresponding binomial distributed r.v., Y. $g_2(Y, p)$ is a second degree polynomial in Y and also a function of p, which can be found in [7,8]. Ref. [8] discusses the properties of the second order approximation $P_{[2]}(Y)$. In particular, when $r_{(2)} = 0$, $P_{[2]}(Y)$ equals $P_{[1]}(Y)$. Ref. [8] uses an example to show that the second order approximation of a correlated binomial distribution model fits a practical KM data set much better than that based on the binomial distribution. For simplicity, in this paper, we call $P_{[2]}(Y)$ a correlated binomial distribution and denote Y~corr. $Bin(N, p, r_{(2)})$..

For the correlated binomial distribution, from (6) the expressions for the false positive error rate and false negative error rate are similar to those in the case of a binomial distribution.

$$E_1 = \sum_{k=C}^{N_0} \binom{N_0}{k} p_{KNM}^k (1-p_{KNM})^{N_0-k}\{1 + r_{(2)} g_2(k, p_{KNM})\} \quad (7)$$

and

$$E_2 = \sum_{k=0}^{C-1} \binom{N_0}{k} p_{KM}^k (1-p_{KM})^{N_0-k}\{1 + r_{(2)} g_2(k, p_{KM})\}. \quad (8)$$

Similar to the case of binomial distribution, the estimates of E_1 and E_2 can be obtained from the CMC values of KNM and KM sets, respectively.

2.3. Beta-binomial distribution

In [2], it is proposed to relax the assumption of the binomial distribution when modeling the CMC values by using a beta-binomial distribution. In this case, we have a sequence of CMC values of image comparisons $\{Y_1, .,Y_M\}$, which are independent from each other. Assume that within one image pair comparison, the probability p for all the independent Bernoulli trials is the same while for different image pairs, p varies and p is random with a beta distribution, i.e., p~$Beta(\alpha, \beta)$ with positive α and β [9]. The pmf of the beta-binomial r.v. Y for given N, α and β is given by

$$P(Y = k|N, \alpha, \beta) = \binom{N}{k} \frac{B(k+\alpha, N-k+\beta)}{B(\alpha, \beta)}, \quad (9)$$

where $B(\alpha, \beta)$ is a beta function [10] and $K = 0, 1, ...,N$. In [2], the beta-binomial distribution is applied to a sequence of CMC values. It shows that the beta-binomial distribution model fits a KM set much better than the model based on the binomial distribution.

For the beta-binomial distribution, the false positive error rate and false negative error rate and their estimates can be obtained in a similar manner to those for the binomial distribution and the correlated binomial distribution.

2.4. Beta-correlated binomial distribution

Although the use of the beta-binomial distribution relaxes the assumption of the same p for the binomial distribution for all image pair comparisons, it still assumes that within each image pair, all cell pair comparisons are independent from each other with the same p. Ref. [8] proposes to use dependent Bernoulli trials for the cell pair comparisons within each image pair. The corresponding probabilities of Y are approximated by $P_{[2]}(Y)$ as given by (6) where only the second-order correlation with a constant $r_{(2)}$ is assumed. As in the case of beta-binomial distribution for a sequence $\{Y_1, .,Y_M\}$, assume that the parameter p in the correlated binomial distribution is random with a beta distribution. Namely, $Y_j|p_j$, $r_{(2)}$~corr. $Bin(N, p_j, r_{(2)})$, for $j = 1, .,M$, where p has a beta distribution, i.e., p~$Beta(\alpha, \beta)$. From [8], the probability mass function of Y for given N, α, β, and $r_{(2)}$ is given by

$$P(Y = k|N, \alpha, \beta, r_{(2)}) = \frac{\binom{N}{k}}{B(\alpha, \beta)} \int_0^1 p^{k+\alpha-1}(1-p)^{N-k+\beta-1}\{1 + r_{(2)} g_2(k, p)\}dp, \quad (10)$$

where $g_2(k, p)$ as in (6) is a second degree polynomial in Y and also a function of p. The marginal probability $P(Y = k|N, \alpha, \beta, r_{(2)})$ 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 of beta-binomial and correlated binomial distributions called a beta-correlated binomial distribution.

For the beta-correlated binomial distribution, the estimates of false positive error rate and false negative error rate can be obtained in a similar manner to those for the binomial distribution and other two distributions in the above.

3. Estimating the model parameters for the case of image pairs with variable N

In the previous study in [8], for simplicity, the number of cell pairs (N) for a sequence of image pairs is assumed to be the same. However, in practice, N always varies from one image pair to another. We consider this general case.

A data set of CMC values, which is a sequence of the CMC results from M image pair comparisons, usually consists of two columns. The first one is a vector of the number of cell pairs (N) for each image

pair in a total of M image pairs. The second column is a vector of the corresponding CMC values for each image pair. From the practical image pair comparisons, it is not likely to have same cell pair number N for all M image pairs [2,5]. That is, the number of cell pairs for each image pair, i.e., N, varies from one image pair to another. Denote a sequence of M independent CMC values by $\{Y_{1,N_1}, .,Y_{j,N_j}, .,Y_{M,N_M}\}$, where $Y_{j,N_j}(j = 1, .,M)$ indicates the CMC value for the jth image pair comparison which consists of N_j cell pairs.

In practice, given a data set of CMC values, statistical models such as binomial, correlated binomial, beta-binomial, and beta-correlated binomial distributions discussed in Section 2 can be applied. Under the assumption of binomial distribution, from [8], the maximum likelihood estimator (MLE) of p is given by

$$\hat{p} = \sum_{j=1}^{M} Y_{j,N_j} / \sum_{j=1}^{M} N_j. \tag{11}$$

In [8], when a correlated binomial distribution is assumed, the maximum likelihood estimator of p and $r_{(2)}$ is discussed for the case that N can be different for different image pairs. For the beta-binomial or beta-correlated binomial distributions, the maximum likelihood estimators of the corresponding parameters can be obtained in a same manner.

In [8], nonlinear regression models are used to estimate p and $r_{(2)}$ under the assumptions: (1) a correlated binomial model, (2) the number of dependent Bernoulli trials N is the same for each Y. Here we discuss the general case that N may vary for different image pairs.

For a sample from the sequence $\{Y_{1,N_1}, .,Y_{j,N_j}, .,Y_{M,N_M}\}$, we call $f_M(k)$ a generalized frequency function given by

$$f_M(k) = \frac{\text{number of elements in the sample} = k}{M} = \frac{\sum_{j=1}^{M} \mathbf{1}_{Y_{j,N_j}=k}}{M}$$
$$\text{for } k = 0, ., \max\{N_j\} \quad (j = 1, .,M), \tag{12}$$

where $\mathbf{1}_A$ is the indicator function of an event A. That is, $\mathbf{1}_{Y_{j,N_j}=k} = 1$ if $Y_{j,N_j} = k$ and $\mathbf{1}_{Y_{j,N_j}=k} = 0$ if $Y_{j,N_j} \neq k$ for $k = 0, ., \max\{N_j\}$ and $j = 1, .,M$. Assume that there are L distinct N's in $\{N_1, .,N_M\}$, denoted by $\{Nd_l\}$ $(l = 1, .,L)$ i.e., $Nd_1 \neq Nd_2 \neq .\neq Nd_L$. Also assume that for each distinct Nd_l there are C_l indicators with the same Nd_l. Thus, $\sum_{i=1}^{l} C_l = M$. Here as an example, we assume Y has a correlated binomial distribution, i.e, Y~corr. $Bin(N, p, r_{(2)})$. Assume that when $M \to \infty$, $C_l \to \infty$ for every $l = 1, .,L$. In addition, assume that for each $l = 1, .,L$, when $M \to \infty$ and $C_l \to \infty$ $W_l = C_l/M$ remains fixed. From (12) by the law of large numbers [11], for every $Y = k$, it can be shown that when $M \to \infty$ and $C_l \to \infty$ $(l = 1, ...,L)$.

$$f_M(k) \to \sum_{l=1}^{L} w_l \text{ corr. } Bin(k, Nd_l, p, r_{(2)}) \text{ almost surely.} \tag{13}$$

In addition, the Glivenko-Cantelli theorem [11] extends the law of large number and gives uniform convergence. Note that $\sum_{l=1}^{L} w_l$ corr. $Bin(k, Nd_l, p, r_{(2)})$ is a proper probability mass function for $k = 0, ., \max\{N_j\}$. Namely, the generalized frequency function will uniformly approach a mixture of correlated binomial pmf's, which is a weighted mean of L correlated binomial pmf's. Eq. (13) and the uniform convergence also hold when Y either has a binomial distribution or a beta-binomial or a beta-correlated binomial distribution. Like Eq. 19 in [8], for a sample of $\{y_1, .,y_M\}$ we have a nonlinear regression model

$$f_M(k) = \sum_{l=1}^{L} w_l \text{ corr. } Bin(k, Nd_l, p, r_{(2)}) + \varepsilon, \tag{14}$$

where $f_M(k)$ for $k = 0, ., \max\{N_j\}$, is the generalized frequency function based on data from (12) and ε is a random error with zero mean. That is, $f_M(k)$ is approximated by a nonlinear function of p and $r_{(2)}$. This nonlinear regression model can be fitted to $\{f_M(k)\}$ to obtain optimal estimates of p and $r_{(2)}$ using Levenberg-Marquardt nonlinear least squares algorithm or other appropriate algorithms.

For illustration, an example is presented. The Weller data set of cartridge cases was obtained from a set of 11 firearm slides produced by the same manufacturer using the same process. The data set has 370 known matching (KM) image pairs and 4095 known nonmatching (KNM) image pairs of breech face impressions. The details of the data set can be found in [2]. The KM set consists of two vectors with lengths = 370, one for N and another for Y (CMC values). Denote the first vector by $\{N\}$. For different image pairs, the number of cell pairs N can be different. In fact, the M = 370 image pairs have L = 15 distinct N's, i.e., Nd = (28, 33, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 49). For each Nd_l $(l = 1, .,15)$, the number of image pairs with Nd_l cell pairs is C_l. That is, $\{C_l, l = 1, .,15\}$ = (7, 21, 17, 9, .,6, 15, 2, 2). From (13), the generalized frequency function will uniformly approach a pmf which is a mixture or a weighted mean of 15 pmf's of the underlying distributions when $M \to \infty$.

First, we assume that for each image pair, the distribution of CMC values is binomial, i.e., Y_j~$Bin(N_j, p)$ $(j = 1, .,370)$. To estimate the parameter p, from (11) the maximum likelihood estimator (MLE) can be applied with $\hat{p}$ = 0.908. Alternatively, assuming that the distributions of CMC values Y is correlated binomial, the nonlinear regression estimates from (14) are $\tilde{p}$= 0.9130 and $\tilde{r}_{(2)}$ = 0.0209. Further when the CMC value is assumed to be beta-binomial or beta-correlated binomial distributed, the nonlinear regression estimates of the corresponding parameters, α, β, and $r_{(2)}$, respectively, are obtained similarly. Fig. 3 demonstrates the mixture pmf's of each of the four probability distributions with corresponding estimated parameters, as well as the generalized frequency function from the KM set. Using the smallest sum of squares of differences between the generalized frequency function from the KM set and each of the four mixture pmfs as a performance criterion, we conclude that for Weller KM set, the mixture of correlated binomial, mixture of beta-binomial, and mixture of beta-correlated binomial models fit the data much better than the mixture binomial model. On the other hand, the other three pmfs perform similarly well as demonstrated in Fig. 3.

The Weller data has 4095 known nonmatching (KNM) image pairs. For each image pair, the number of cell pair N varies. In fact, the M = 4095 image pairs have L = 15 distinct N's. From (13), the generalized frequency function will uniformly approach a pmf which is a mixture or a weighted mean of 15 pmf's of the underlying distributions when $M \to \infty$.

We consider using the binomial, correlated binomial, and beta-binomial distributions to fit the KNM set of the Weller data with 4095 KNM image pairs. Under the assumption of binomial distribution, the MLE of $\hat{p}$= 0.0011209. Under the assumption of correlated binomial distribution, the MLE's are given by $\hat{p}$ = 0.001121 and $\hat{r}_{(2)}$ = 0.0001652 while for the beta-binomial distribution, the MLE's are $\hat{\alpha}$= 7.2594 and $\hat{\beta}$ = 6469.527. The generalized frequency function and the mixtures of the three pmf's when Y= 0, 1, 2, 3, 4 are listed in Table 1. Using the sum of squared difference (SSD) between the generalized frequency function and the mixtures of the three pmf'as a criterion, respectively, the three models fit the generalized frequency function data similarly well. Fig. 4 demonstrates the mixture pmf's of binomial and correlated binomial distributions for Weller KM set with the corresponding estimated parameters given in Fig. 3 as well as the generalized frequency functions from Weller KM and KNM sets, respectively. The curves of the mixture pmf's of binomial, correlated binomial, and beta-binomial probability distributions for the KNM set are overlapping the curve for the generalized function

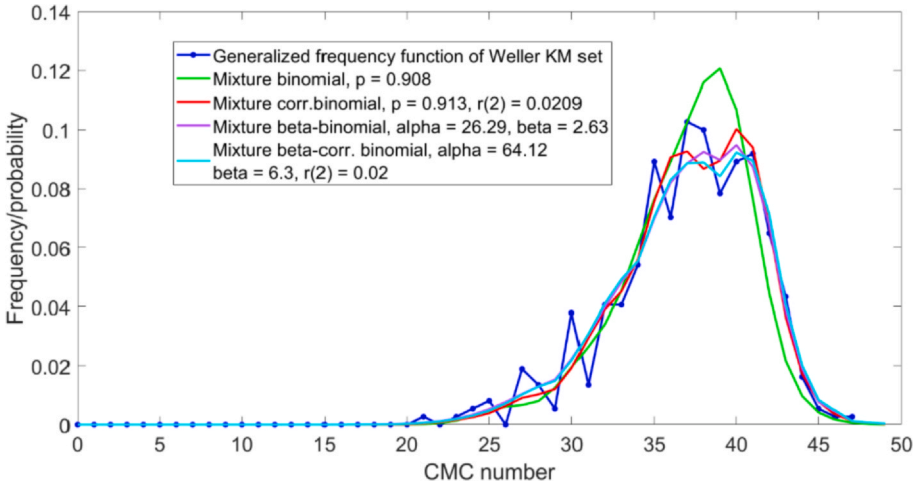

Fig. 3. Generalized frequency function of the CMC values from the KM image pairs of the Weller data with four mixture probability mass functions.

Table 1
Generalized frequency function of the Weller KNM set and pmf's of three mixture probability distributions.

Y (CMC value)	Generalized frequency function	Mixture Binomial pmf	Mixture Correlated binomial pmf	Mixture Beta-binomial pmf
0	0.954572	0.954410	0.954556	0.954552
1	0.044192	0.044553	0.0442641	0.044278
2	0.001212	0.001021	0.0011581	0.001147
3	0	0.000015	0.000022	0.000022
4	0	0	0	0
SSD		$1.93{\cdot}10^{-7}$	$9.05{\cdot}10^{-9}$	$1.27{\cdot}10^{-8}$

from the KNM data set (as indicated in Table 1) and thus are not shown. The curves for the generalized frequency functions from the KM and KNM data sets in Fig. 4 show a significant separation.

Regarding the error rates for the binomial distribution discussed in Section 2.1, the false negative error rate in (5) is estimated by replacing P_{KM} by $\hat{P}_{KM} = 0.908$. The threshold of the CMC values is chosen to be $C = 6$ and N_0 is assumed to be $N_0 = \text{median}\{N\} = 42$. From Eq. (5), the false negative error rate is estimated by

$$\hat{E}_2 = \sum_{k=0}^{5} \binom{42}{k} \hat{p}_{KM}^{k} (1 - \hat{p}_{KM})^{42-k} = 2.43{\cdot}10^{-33}. \tag{15}$$

Alternatively, when the correlated binomial distribution is assumed, the nonlinear regression estimates of the parameters are $\tilde{p} = 0.9130$ and $\tilde{r}_{(2)} = 0.0209$. From Eq. (8), $\hat{E}_2 = 4.53{\cdot}10^{-32}$. Similarly, for the beta-binomial, beta-correlated binomial with the corresponding parameter estimates shown in Fig. 3, $\hat{E}_2 = 1.80{\cdot}10^{-12}$ and

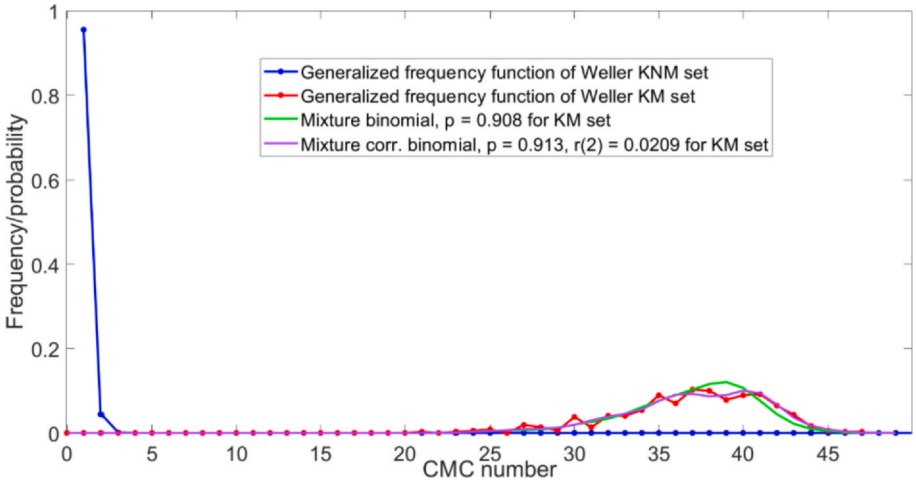

Fig. 4. Generalized frequency functions of the CMC numbers of KM and KNM images of the Weller data with two mixture probability mass functions for the KM images.

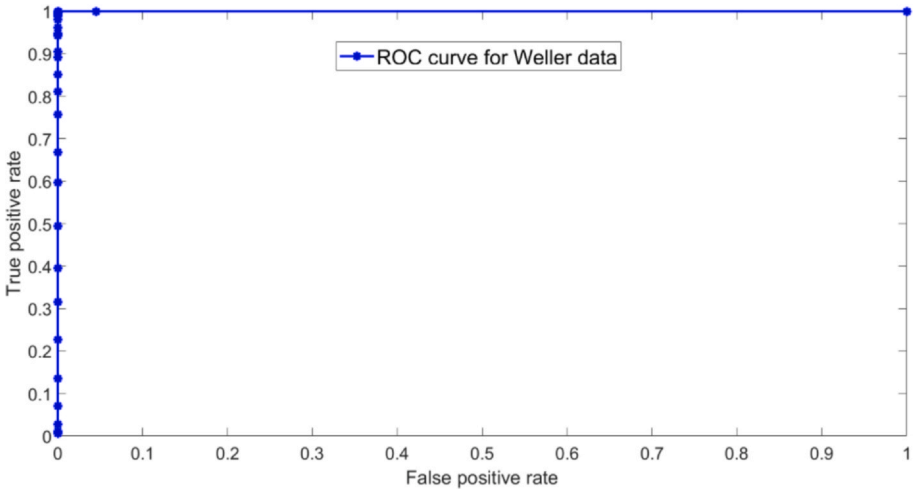

Fig. 5. The ROC curve for Weller data.

$\hat{E}_2 = 3.85{\cdot}10^{-17}$, respectively. Although the estimate of false negative error varies with the underlying statistical models as well as the different estimation approaches, the estimated negative error rates are all very small.

From (3), for the binomial distribution, the false positive error rate is estimated by replacing P_{KNM} by $\hat{P}_{\text{KNM}} = 0.001121$ with $C = 6$ and $N_0 = \text{median}\{N\} = 42$. That is,

$$\hat{E}_1 = \sum_{k=6}^{42} \binom{42}{k} \hat{p}_{\text{KNM}}^{k} (1 - \hat{p}_{\text{KNM}})^{42-k} = 1.01{\cdot}10^{-11}. \tag{16}$$

Alternatively, for the correlated binomial distribution, the MLE $\hat{P}_{\text{KNM}} = 0.001121$ and $\hat{r}_{(2)} = 0.000165$. The corresponding estimate of false positive error rate is $\hat{E}_1 = 3.12{\cdot}10^{-11}$. On the other hand, for the beta-binomial distribution with MLE $\hat{\alpha} = 7.2594$ and $\hat{\beta} = 6469.527$, the corresponding estimate of false positive error rate is $\hat{E}_1 = 5.48{\cdot}10^{-13}$.

The receiver operating characteristic (ROC) curve is often used for evaluating the performance of binary classification systems, which in our case is a system for classifying matching and nonmatching image pairs. An ROC curve provides a graphical representation of the classifier's performance by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings [12]. For Weller data including the KM and KNM image sets, the TPR and FPR can be calculated for a range of threshold. FPR and TPR are calculated based on KNM and KM image sets, respectively. Fig. 5 shows the ROC curve for the Weller data with the threshold C varies from 0,1,... to 46. On the other hand, each of the four statistical models can be applied to the Weller data to estimate the true positive rates and false positive rates. In fact, a false positive rate for a given N and a selected C here is the same as the false positive error rate for C from (2) and can be estimated, for example under the assumption of a binomial distribution as in (16) based on the KNM set. Similarly, the true positive rates can be estimated. Fig. 6 shows an ROC curve under the assumption of binomial distributions for $N = 46$ and with the corresponding MLEs of p given in the above. Specifically, for each threshold value from 0, 1,..., to 46, the FPR is estimated based on the KNM set for a binomial distribution with $\hat{p} = 0.0011209$ while the TPR is estimated based on the KM set for a binomial distribution with $\hat{p} = 0.908$. The ROC curves based on the binomial and other three statistical models with the corresponding estimated parameters are all overlapped. Note that the ROC curves in

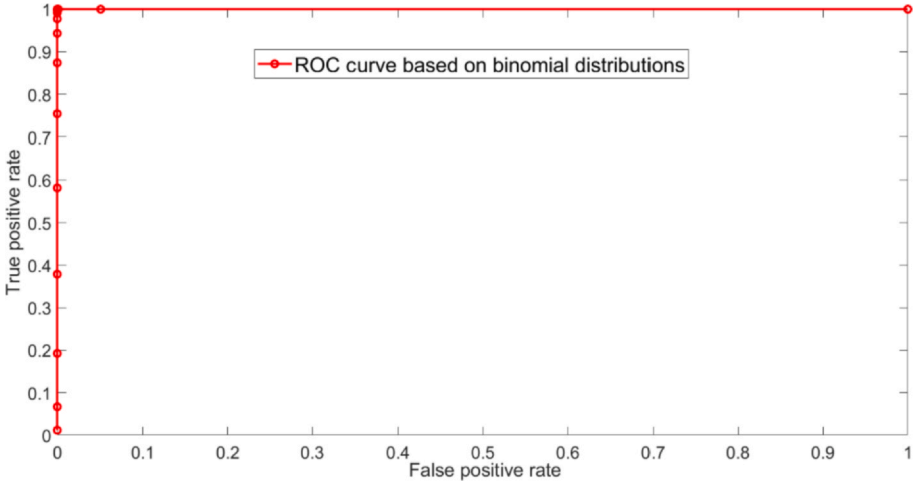

Fig. 6. The ROC curve based on binomial probability distributions with $N = 46$.

Figs. 5 and 6 are indistinguishable. They demonstrate almost perfect separations of KM and KNM distributions of the Weller data as well as that the statistical models fit the data well.

4. Discussions and conclusions

This paper discusses four statistical models for the results of similarity score comparisons in firearm and tool mark image comparisons, specifically the congruent matching cells (CMC) method, which is based on pass-or-fail tests. Because the assumption of independence among the cell pair comparisons from the CMC method is most likely invalid in practice, the correlated binomial distribution is proposed. On the other hand, although the beta-binomial distribution can relax the assumption of the same p for all image pair comparisons, it still assumes that within each image pair, all cell pair comparisons are independent from each other. To relax that assumption, the beta-correlated binomial as a compound probability distribution of the beta-binomial and correlated binomial distributions is proposed. Theoretically, the beta-correlated binomial, beta-binomial, and correlated binomial distributions are more complete than the binomial distribution because they contain fewer assumptions.

In the previous study in [8], the number of cell pairs (N) for a sequence of image pairs is assumed to be the same. However, in practice, N can change from one image pair to the next. In that case, the term, frequency function, of the CMC results is not appropriate, and the nonlinear regression estimator proposed in [8] cannot be applied. In Section 3, the generalized frequency function is introduced to depict the behavior of the CMC values for variable N values. By the law of large numbers, the generalized frequency function approaches a mixture probability mass function of the underlying probability distribution when the number of image pairs increases. Based on that, nonlinear regression estimators of the model parameters are proposed for the CMC data with variable N values. The methodology is applied to an actual CMC data set (Weller data) of fired cartridge cases with variable N to check the performance of the different statistical models based on the comparisons between the generalized frequency function data and the corresponding limiting probability mass functions. For the known matching (KM) set of the Weller data, the beta-correlated binomial, correlated binomial, and beta-binomial distributions fit the data much better than the binomial distributions. On the other hand, for the known nonmatching (KNM) set, each of the binomial, correlated binomial, and beta-binomial distributions fits the data well. Since for the KNM set, the estimate of correlation $r_{(2)}$ for the correlated binomial distribution is often very small and the pmf's of the correlated binomial and binomial distribution are almost indistinguishable, the binomial distribution may be good enough for fitting a KNM set. In addition, for Weller data the error rates, the ROC curve, and their estimations are discussed.

CRediT authorship contribution statement

Nien Fan Zhang received the M.S. and Ph.D degrees in statistics from Virginia Polytechnic Institute and State University, Blacksburg, in 1982 and 1985 respectively. He is a mathematical statistician in the Statistical Engineering Division of National Institute of Standards and Technology in Gaithersburg, MD, USA.

Declaration of interest

None.

Acknowledgments

The research was supported by the Special Programs Office (SPO) of NIST. The author thanks J. Yen, J. Song, T. V. Vorburger, J. Soons, and two reviewers for their helpful comments. The author also thanks M. Henn for his assistance on computation.

References

[1] G. Gerules, S.K. Bhatia, D. Jackson, A survey of image processing techniques and statistics for ballistic specimens, Sci. Justice 53 (2015) 236–250.
[2] J. Song, T.V. Vorburger, W. Chu, J. Yen, J.A. Soons, D.B. Ott, N.F. Zhang, Estimating error rates for firearm evidence identification in forensic science, Forensic Sci. Int. 284 (2018) 15–32.
[3] J. Song, E. Whitenton, D. Kelley, R. Clary, L. Ma, S. Ballou, SRM 2460/2461 standard bullets and cartridge cases project, J. Res. Natl. Inst. Stand. Technol. 109 (6) (2004) 533–542.
[4] J. Song, Proposed NIST ballistic identification system (NBIS) based on 3D topography measurements on correlated cells, J. Assoc. Firearm Tool-Mark. Exam. 45 (2) (2013) 184–189.
[5] J. Song, Z. Chen, T.V. Vorburger, J.A. Soons, Evaluating Likelihood Ratio (LR) for firearm evidence identifications in forensic science based on the Congruent Matching Cells (CMC) method, Forensic Sci. Int. 317 (2020) 110502.
[6] A. Papoulis, Probability, Random Variables and Stochastic Processes, 2nd ed., McGraw-Hill,, New York, 1984.
[7] R.R. Bahadur, A representation of the joint distribution of the response to n dichotomous items, in: H. Solomon (Ed.), Studies in Item Analysis and Prediction, Stanford University Press, Stanford, CA, 1961, pp. 158–168.
[8] N.F. Zhang, The use of correlated binomial distribution in estimating error rates for firearm evidence identification, J. Res. Natl. Inst. Stand. Technol. (2019) 124Article No 124026.
[9] N.L. Johnson, S. Kotz, N. Balakrishna, Continuous Univariate Distributions, 2nd ed., Wiley, 1995.
[10] R.A. Askey, R. Roy, Beta function, in: D.M. Lozier, Boisvert, C.W. Clark (Eds.), NIST Handbook of Mathematical Functions, Cambridge University Press, 2010.
[11] A.W. van der Vaart, Asymptotic Statistics, Cambridge University Press, Cambridge, UK, 1998, pp. 265–266.
[12] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (8) (2006) 861–874.