NISTIR 7740

Comparison of Confidence Intervals for Large Operational Biometric Data by Parametric and Non-parametric Methods

Su Lan Cheng
Ross Micheals
Z. Q. John Lu

National Institute of Standards and Technology
U.S. Department of Commerce

NISTIR 7740

Comparison of Confidence Intervals for Large Operational Biometric Data by Parametric and Non-parametric Methods

Su Lan Cheng
Ross Micheals
Information Technology laboratory
Information Access Division

Z. Q. John Lu
Information Technology laboratory
Statistical Engineering Division

November 2010

U.S. Department of Commerce
Gary Locke, Secretary

National Institute of Standards and Technology
Patrick D. Gallagher, Director

ABSTRACT

Receiver operating characteristic (ROC) or Detection Error Trade-off (DET) curves are used to measure the performance of a biometric verification or identification system. To go beyond reporting the ROC/DET and to enhance evaluation of a verification system we compute the confidence intervals of the False Accept Rate (FAR) and False Reject Rate (FRR).

In this paper, we validate the accuracy of variance estimators by comparing the estimators to variance computed over repeated experiments. The confidence intervals of the error rate using both parametric and non-parametric methods are evaluated. For the parametric approach, we calculate the confidence interval based on variance estimations from the survey sampling variance approach and the binomial distribution model approach. For the non-parametric approach, we use the bootstrap method to compute the confidence intervals directly. Two different datasets selected from the National Institute of Standards and technology (NIST) Proprietary Fingerprint Template Evaluation II (PFTII) program and several authentication systems are tested in the evaluation process. Then the confidence intervals from all three approaches are reported with different sample sizes.

What we found from the evaluation result is that there is no significant difference between the confidence intervals computed by all three methods. However, for very large data sets, the binomial model approach is computationally the most efficient among the three approaches, and we also argue that it can easily be extended in theory to evaluation problems with smaller data sets or extremely low error rates.

To enhance readability, we have chosen to use the familiar terms FAR and FRR rather than the more formal equivalent terms False Match Rate (FMR) and False Non-Match Rate (FNMR). Because the interest is using the operation of the matcher rather than that of the complete verification or identification system, we do not include other types of system errors such as Failure To Acquire (FTA), or Failure To Enroll (FTE).

1. Introduction

The confidence intervals for FAR and FRR based on smaller datasets of hundred subjects or less have been studied and researched in papers by using both parametric and non-parametric estimation. The parametric approach estimates the confidence intervals based on the binomial distribution. The non-parametric approach, bootstrap technique, has been proposed to estimate the confidence intervals of the error rate [1, 2, 3]. Note that [3] is based on much larger dataset with 60K (thousand) subjects.

We use both methods to compute confidence intervals for very large datasets. In addition to these two common methods, we take a different approach by using the sampling variance theory to estimate the confidence interval. Based on [4, 5], simple random sampling, stratified sampling and cluster sampling, have been investigated extensively. The variance estimated from a single trial agrees with the average sample variance from repeating experiments. This research came from the realm of biometric facial recognition. In the present study, we adapt simple random sampling (SRS) to compute the variance, then comparing with the results in [1, 3] of the binomial and bootstrap estimates. Confidence intervals from each method are described in detail. The resulting confidence intervals based on several data sets are then compared.

Indeed, to strengthen our results, datasets of different sample size with different genuine scores and impostor scores have been explored with the same procedures.

In this report, we include the following sections:

- Datasets and Terminology in section 2
- Variance estimation by random sampling with partitions and by binomial estimator in section 3
- Confidence Interval in section 4
- Conclusion and findings in section 5
- Reference in section 6
- Appendix A & B in section 7

2. Datasets and Terminology

The datasets used in this study are taken from the NIST PFTII program. PFTII is part of a NIST ongoing program to measure the performance of fingerprint matching software utilizing vendor proprietary fingerprint templates[1].

[1] PFTII web site can be found in http://www.nist.gov/itl/iad/ig/pftii.cfm

PFTII contains three sample datasets: two two-finger datasets, and one ten-print dataset. We select one two-finger dataset (dataset 1) of sample 120,000 subjects which are randomly selected from 3.5 million, and the ten-print dataset (dataset 2) of sample 120,000 subjects from 2 million.

The similarity scores used in the study are from evaluating three algorithms (A, B, C) from NIST PFT[2] tests for finger 02 (right index) of dataset 1, and finger 01 (right thumb) of dataset 2.

2.1 Dataset Configuration

The sample data are randomly selected and grouped into one gallery set and two probe sets with the following structure:

- The gallery set $\mathcal{G}$ contains M subjects.
- The first probe set $\mathcal{P}_m$ contains M **m**ates to the M subjects of set $\mathcal{G}$.
- A second probe set $\mathcal{P}_{nm}$ contains N **n**on-**m**ates to the M subjects in $\mathcal{G}$.

where M = 120,000, and N = 120,000 with the following constraints:

- $\mathcal{P}_m$ contains exactly the same subject IDs as $\mathcal{G}$, but with different images that have been acquired at a different time.
- All the IDs from datasets are consolidated,

  i.e. there are no common IDs between $\mathcal{G}$ and $\mathcal{P}_{nm}$, from which it necessarily follows that there are no common IDs between $\mathcal{P}_m$ and $\mathcal{P}_{nm}$.

  The consolidation came from the results of evaluating the NIST PFT test, other NIST Automated Fingerprint Identification System(AFIS) testing and confirmation by the examiners.
- Fingerprint matching always involved the same finger position (e.g., right index) for both the probe and the gallery.

The fingerprint verification protocol of PFTII is one-to-one matching which is shown in Figure 1. Therefore the matching process produces M genuine scores and N (= M) impostor scores.

[2] PFT web site is in http://www.nist.gov/itl/iad/ig/pft.cfm

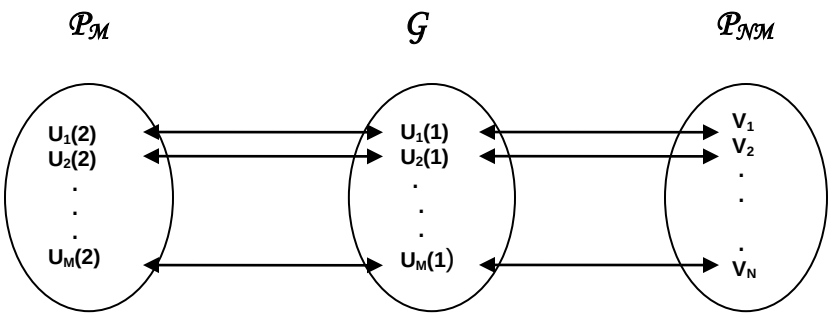

Figure 1

2.2 Terminology

Let **X** = {X_1,.. X_N} be a sample of *impostor* scores from matching subjects between sets { $\mathcal{Pnm}$, $\mathcal{G}$ } and **Y** ={Y_1,.. Y_M} be a sample of *genuine* scores set from matching subjects between {$\mathcal{Pm}$, $\mathcal{G}$}.

For the set of impostor scores, assume that **x** is a sample of N numbers drawn from a population with distribution function F, that is, F(x) = Prob(X ≤ t). For the set of genuine scores Y, assume that Y be a sample of N numbers drawn from a population with the distribution G(y) = Prob(Y ≤ t).

Let x denote the imposter score from any chosen case, and y is the genuine score from any chosen case:

- F(x) = Prob(X ≤ x) is the theoretical cumulative distribution function of impostor scores **X**.
- G(y) = Prob(Y ≤ y) is the cumulative distribution function of genuine scores **Y**.
- The unbiased estimate of F(t) at x = t (threshold) and

  the unbiased estimate of G(t) at y = t are:

$$\hat{F}(x = t) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\{X_i \le t\}} = \frac{1}{N}(\#X_i \le t)$$

$$\hat{G}(y = t) = \frac{1}{M}\sum_{i=1}^{N} \mathbf{1}_{\{Y_i \le t\}} = \frac{1}{M}(\#Y_i \le t)$$

These two estimates are random variables.

From the definition FAR and FRR at threshold t are:

$$\mathrm{FAR}(x) = (\#Xi > t)/N = 1 - (\#Xi \le t)/N \quad (1)$$
$$\mathrm{FRR}(y) = (\#Yi \le t)/M$$

Thus

$$\mathrm{FAR}(x) = 1 - \hat{F}(x)$$
$$\mathrm{FRR}(y) = \hat{G}(y)$$

Since N=M, we use M throughout the report.

3. Variance Estimation

First we will compute the empirical variance by the random partition samples and from binomial distribution theory.

There is a relationship between SRS and i.i.d. data: see [4, 5] in the survey sampling. SRS implies that every element has the same probability of being selected. If the sample data (impostor scores and genuine scores) are i.i.d., we can estimate the variance of FAR and FRR from a single trial from repeatable experiments by SRS.

Micheals [6] shows that the unbiased variance for the simple random sampling is:

$$\sigma_\theta^2 = V_{SRS}(\bar{\bar{X}}) = \mu_\theta (1 - \mu_\theta) / M \quad (2)$$

where μ_θ and σ_θ represent the true mean and variance of θ, and θ is the probability of success of the random variable X. Details about the distribution of θ can also be found in [5, 6].

3.1 Random Sampling using Partitions & Procedure

In order to get multiple estimates of the true empirical variance and to determine qualitatively if the choice of partition significantly affects the result, we use random, non-overlapping partitioning of the sample to simulate a 'hold-out' style cross-validation of the experiment.

To simplify the mechanics of the randomizing process, we begin by randomly generating L exhaustive partitions of the set {1, 2, 3, …, M}, where each partitioning $\boldsymbol{\mathcal{P}}_k$ ($1 \le k \le L$) consists of m sets each of size n, such that n x m = M. We note that the first set in each partition $\boldsymbol{\mathcal{P}}_k$ is selected by random sampling without replacement from set {1, …, M}, the second set selected by random sampling without replacement from the residue left by the selection of the first set, and so on; the m^{th} such set finally exhausts {1, …, M}.

Let $\boldsymbol{\mathcal{P}} = \{(X_1,Y_1),\cdots (X_M,Y_M)\}^T$ be the set of pairs of impostor score and genuine score from a specific matcher. From this baseline of sample scores, we empirically determine a set of thresholds of interest: $\boldsymbol{\mathcal{T}} = \{t_1, t_2, t_3, t_4, t_5\}$ that corresponding to the following set of target Far values:
$\{10^{-4}, 5x10^{-4}, 10^{-3}, 5x10^{-3}, 10^{-2}\}$ respectively.

By using the partitioning $\mathcal{P}_k$ we generate sets of random sampled pairs [without replacement] from set $\mathcal{P}$. We do this L times to form new sets { $\mathcal{P}_1$, $\mathcal{P}_2$, ⋯ $\mathcal{P}_L$}. Each set $\mathcal{P}_i$, is thus partitioned into m subsets of size n, where M = m x n, shown as in Figure 2.

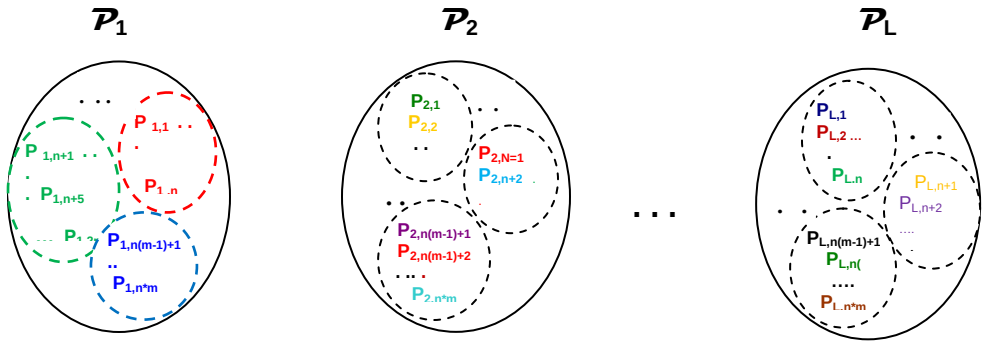

Figure 2

Let $\mathcal{P}_i = (\mathcal{X}_i, \mathcal{Y}_i)$, the i-th random sample set of $\mathcal{P}$ where

$$\mathcal{X}_i = \{X_1, \ldots X_n, .X_{n+1}, .. X_{2n}, \quad \ldots \quad X_{(m-1)n}, .. X_{mn}\}_i^T \text{ (i-th imposter set)}$$

$$= \{ \underbrace{\mathcal{X}_{i1},}_{} \underbrace{\mathcal{X}_{i2},}_{} \ldots \underbrace{\mathcal{X}_{i\,m}}_{} \}^T ; \quad (T - \text{Transpose})$$

$$\mathcal{Y}_i = \{Y_1, \ldots Y_n, .Y_{n+1}, .. Y_{2n}, \ldots . \quad Y_{(m-1)n}, .. Y_{mn}\}_i^T \quad \text{(i-th genuine set)}$$

$$= \{ \underbrace{\mathcal{Y}_{i\,1},}_{} \underbrace{\mathcal{Y}_{i\,2},}_{} \ldots \underbrace{\mathcal{Y}_{i\,m}}_{} \}^T$$

3.2 Sampling Variance of FAR and FRR

First, compute the FAR($\mathcal{X}_i$) and FRR($\mathcal{Y}_i$) from each $\mathcal{P}_i$ by the following procedure:

For each (i = 1 … L):

- Calculate FAR($\mathcal{X}_{i1,}$), …. , FAR($\mathcal{X}_{i\,m}$) from each subgroup { $\mathcal{X}_{i1,}$ …. $\mathcal{X}_{i\,m}$}
- Calculate FRR($\mathcal{Y}_{i1,}$), …. , FRR($\mathcal{Y}_{i\,m}$) from each subgroup {$Y_{i1,}$ …. $\mathcal{Y}_{i\,m}$}
- Compute the sample variance of FAR($\mathcal{X}_i$), V(FAR($\mathcal{X}_i$)) from

$$V_i(x) = \frac{\sum_{j=1}^{m}(x_{ij} - \bar{x}_i)^2}{m-1} \qquad (3)$$

where x_{ij}= FAR($\mathcal{X}_{i\,j}$) and $\bar{x}_i = \sum_{j=1}^{m} (\text{FAR}(\mathcal{X}_{i\,j}))/m$, (i= 1,.. L).
Apply FRR($\mathcal{Y}_i$) to equation (4) to compute V(FRR($\mathcal{Y}_i$)).

To simplify notation, let $\bar{x}_i = \overline{\mathrm{FAR}_i(x)}$, and $\bar{y}_i = \overline{\mathrm{FRR}_i(y)}$

3.3 Estimated Variance of Binomial Estimator

Given data $X=\{x_1, x_2, \cdots x_M\}$ as described in 2.2 is a binomial random variable with the probability of success in M trials, i.e.,

$$\hat{F}(t) = \sum_{i=1}^{M} \mathbf{1}_{\{\mathbf{x_i} \le \mathbf{t}\}} \cong B(M, F(t))$$

then the sample proportion $Z = X/M$ is an unbiased estimator of F(t), with the expectation:

$$E(\hat{F}(t)) = F(t)$$

and the variance:

$$\tilde{\sigma}^2(t) = V\left(\hat{F}(t)\right) = \frac{1}{M}\tilde{F}(t)\left(1 - \tilde{F}(t)\right) \qquad (4)$$

We compute the variance for the binomial estimator from (4). Same applies to $V(\hat{G}(t))$. Details of this approach can be found in [2].

The binomial model for a sum of independent and identically distributed 0,1-valued random variables, or Bernoulli trials is standard, but the model applies to more general situations such as when the individual trials have different probability distributions [10]. To explain this important extension, we introduce some generic notations. Let p_i denote the probability of success at the i^{th} trial, and S be the number of successes in M independent trials. Then the variance of S is maximum when $p_1=p_2=\ldots = p_M=p$. This can be interpreted as follows: using the binomial model $B(M,(\sum_{i=1}^{M} P_i)/M)$ to compute the variance provides the correct upper bound for Var (S) even when the error rates of individual components are not equal (Theorem 3 of [10]). Furthermore, the confidence intervals based on the above binomial model are still valid even when the individual trials have unequal probabilities (cf Theorem 5 of [10]). When M is very large, the standard approximation using the normal distribution or Poisson approximation (for very small p's) of the Binomial distribution can be used to simplify binomial probability calculations.

For each random partition samples $\mathbf{P_i}$, we computer the variance:

$$\sigma^2_{\bar{x}_i} = V(\bar{x}_i) = \frac{\overline{\mathrm{FAR}_i(x)}\left(1 - \overline{\mathrm{FAR}_i(x)}\right)}{n} \qquad (5)$$

where n is the sample size of $\mathbf{P_i}$.

Therefore

$$V_{BIN}(\bar{x}_i) = \frac{\overline{FAR_i(x)}\left(1 - \overline{FAR_i(x)}\right)}{n}$$

and

$$V_{BIN}(\bar{y}_i) = \frac{\overline{FRR_i(y)}\left(1 - \overline{FRR_i(y)}\right)}{n}$$

for (i= 1,.. L).

The resulting variance from those two computations (3) and (5) and the mean of the sample variance of both FAR and FRR are displayed in Figure 3 and Figure 4 for matcher Algorithm A at the threshold where FAR = 10^{-4}, L = 1000.

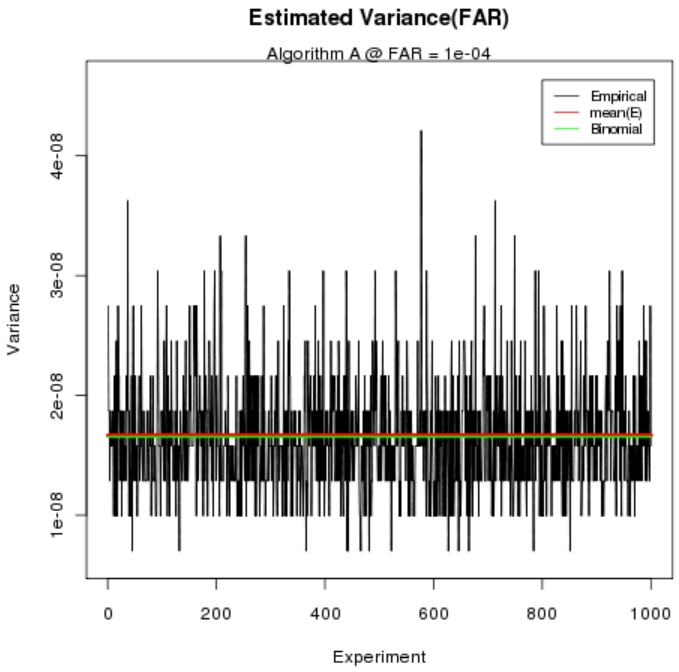

Figure 3

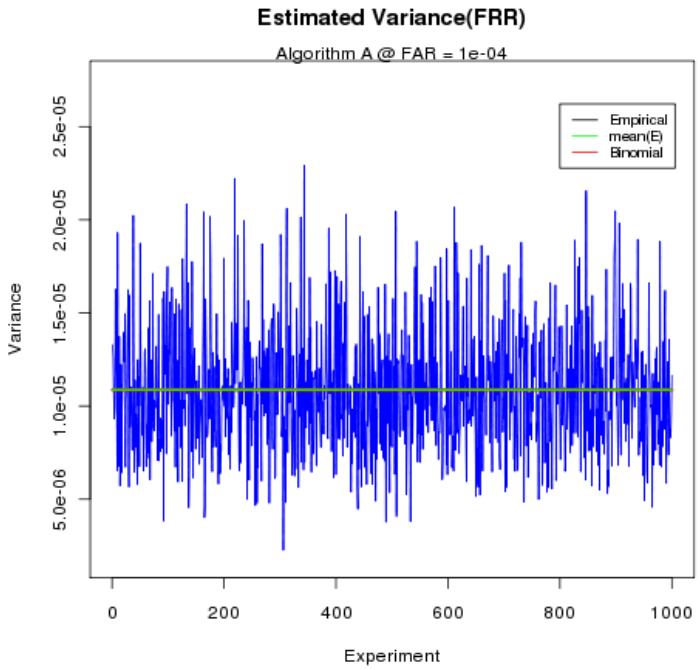

Figure 4

4. Confidence Intervals

Applying the normality tests Anderson Darling and Wilk-Shapiro, to each set {FAR($\boldsymbol{\mathcal{X}}_{i1}$), …. , FAR($\boldsymbol{\mathcal{X}}_{i\ m}$)} and to { FRR($\boldsymbol{\mathcal{Y}}_{i1}$), …. , FRR($\boldsymbol{\mathcal{Y}}_{i\ m}$) }. The result show that FAR($\boldsymbol{\mathcal{X}}_i$) and FRR($\boldsymbol{\mathcal{Y}}_i$) are both normally distributed when FAR and FRR are not too small (> 10^{-4}) from the p-values listed in Table 4 in appendix A. Furthermore, applying the multivariate Wilk-Shapiro test to FAR($\boldsymbol{\mathcal{X}}_i$) and FRR($\boldsymbol{\mathcal{Y}}_i$) indicates that these sets are Bivariate Normal distributed (p-values are shown in Appendix A). The statistical summary of correlation coefficients (ρ) indicates that the correlation of FAR and FRR is centered around ρ=0. (The minimum, twenty-five percentile, median, seventy five percentile and maximum of ρ values are shown in Appendix A Table 5). Further investigation and research will be needed to establish the confidence region from the correlation coefficients.

However the individual normal distribution 95 % confidence interval of FAR and FRR are:

a. For each $\boldsymbol{\mathcal{P}}_i$:

$$\left(\bar{x} - 1.96 * \frac{\hat{\sigma}}{\sqrt{m}}\ , \bar{x} + 1.96 * \frac{\hat{\sigma}}{\sqrt{m}}\right) \qquad (6)$$

where $\bar{x} = \left(\overline{\mathrm{FAR_i(x)}}\right)$; $\bar{y} = \left(\overline{\mathrm{FRR_i(y)}}\right)$

$$\hat{\sigma} = \sqrt{V(\overline{\mathrm{FAR}_i(x)})}; \quad \hat{\sigma} = \sqrt{V(\overline{\mathrm{FRR}_i(y)})}$$

b. For the binomial estimate, the confidence interval is derived from the 95th percentile of the normal distribution for large M:

$$(\bar{x} - 1.96 * \hat{\sigma}(\mathrm{x}), \bar{x} + 1.96 * \hat{\sigma}(\mathrm{x})) \qquad (7)$$

where

$$\hat{\sigma}(\mathrm{x}) = \sqrt{\frac{\overline{\mathrm{FAR(x)}}\left(1 - \overline{\mathrm{FAR(x)}}\right)}{\mathrm{M}}}$$

c. For the non-parametric bootstrap, the 95 % confidence interval is:

$$(\ \mathrm{FAR}_{\ q(\alpha/2)}\ (\ \mathrm{x}^*)\ ,\ \ \mathrm{FAR}_{\ q(1-\alpha/2)}\ (\ \mathrm{x}^*)\ )$$

where α = 0.05, and q(α/2) and q(1- α/2) are the 2.5 % and 97.5 % quantiles of bootstrap estimate of $\mathrm{FAR}(\mathrm{x}^*)$. In this report, a NIST non-parametric program from [3] was adopted here to estimate the confidence interval. The number of bootstrap replication in the program is 2000.

d. Calculate the confidence interval from the mean of $\mathrm{FAR}_i(\mathrm{x})$ by applying (6) with:

$$\bar{\bar{x}} = \overline{\overline{\mathrm{FAR(x)}}}; \quad \hat{\sigma} = \sqrt{\mathrm{V}(\overline{\overline{\mathrm{FAR(x)}}})}; \ \ \mathrm{V}\left(\overline{\overline{\mathrm{FAR(x)}}}\right) = \frac{\mathrm{V}\left(\overline{\mathrm{FAR(x)}}\right)}{L}$$

In calculation the confidence interval for the binomial distribution B(n, p), we approximate, using the normal distribution N(np, np(1-p)). This works well when n is very large and p is not too small. However, when n is large and p is very small, B(n, p) can be approximated by the Poisson distribution P(λ= np). As a rule of thumb, the Poisson approximation can be applied if n ≥ 20, p ≤ 0.05 or n ≥ 100, np ≤ 10.

The details of the confidence intervals of FAR and FRR from the each experiment (i), i = 1…L at FAR = 10^{-3} , for binomial, bootstrap, and confidence interval from the mean variance are shown in the Figure 5, and 6.

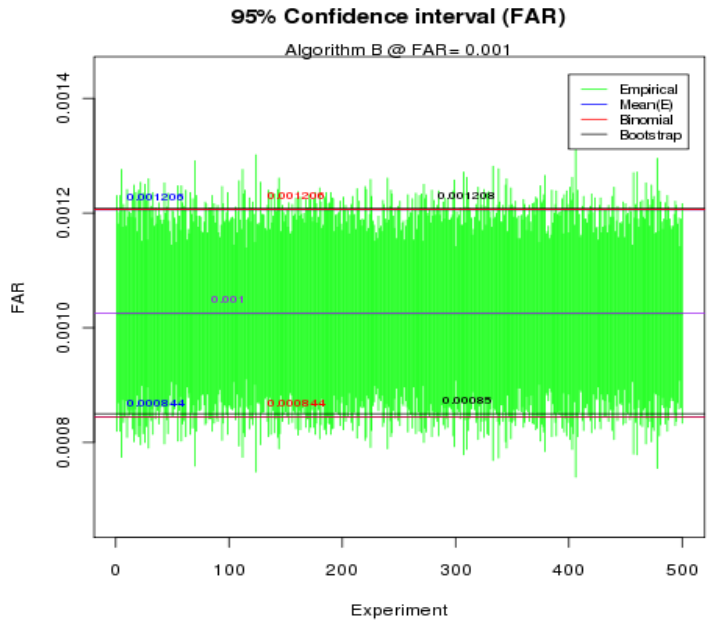

Figure 5

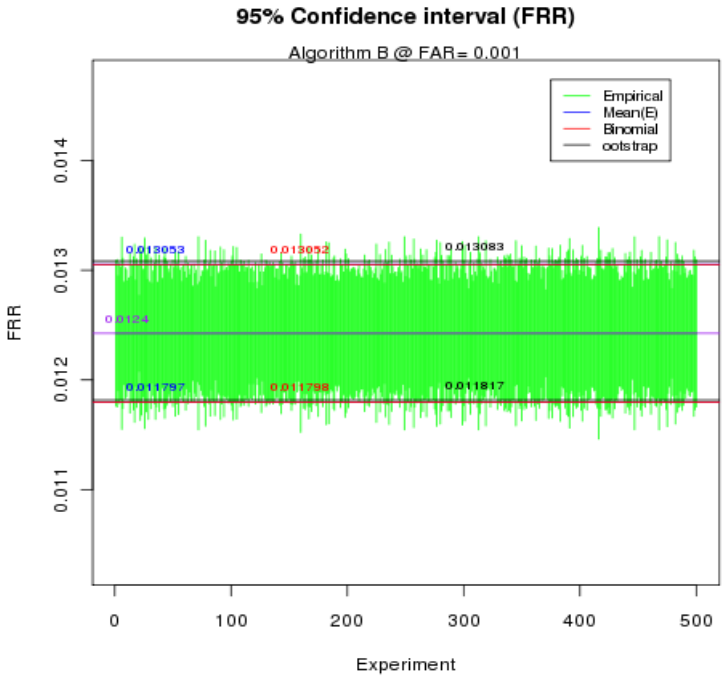

Figure 6

Figure 7 and 8 represent the CI of FAR and FRR from all five target threshold set $\mathcal{T}$, by the variance mean, binomial, and bootstrap estimation for Algorithm C.

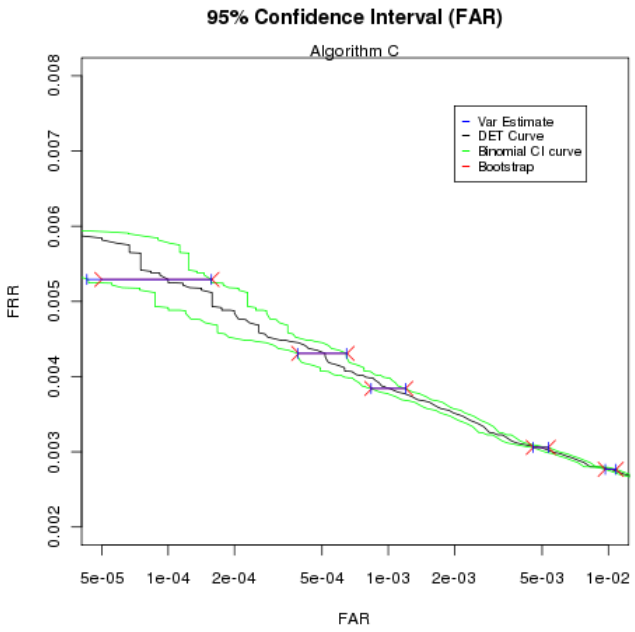

Figure 7

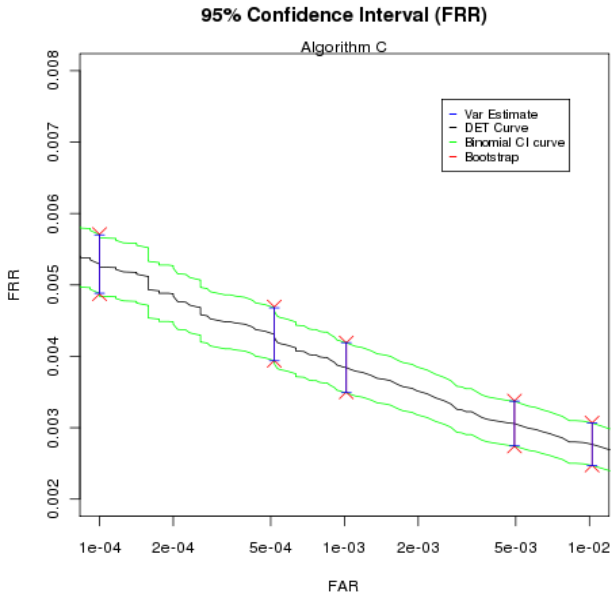

Figure 8

Data of the CIs from Algorithm A, B and C are shown by Table 1, 2 and 3. Note that these results are based on the scores from dataset 1 of finger 02.

Table 1: Algorithm A - Dataset 1 finger 02

A	FAR	Variance	Binomial	Bootstrap	FRR	Variance	Binomial	Bootstrap
CI^{+} [3]	0.0001	0.000156	0.000157	0.000158	0.070142	0.071587	0.071587	0.071608
CI-		0.000044	0.000043	0.000050		0.068697	0.068697	0.068729
CI^{+}	0.0005	0.000636	0.000635	0.000635	0.054058	0.055345	0.055337	0.055317
CI-		0.000380	0.000381	0.000381		0.052771	0.052779	0.052758
CI^{+}	0.0010	0.001241	0.001242	0.001242	0.045408	0.046590	0.046586	0.046550
CI-		0.000875	0.000874	0.000874		0.044226	0.044230	0.044221
CI^{+}	0.0050	0.005483	0.005485	0.005485	0.033567	0.034593	0.034586	0.034613
CI-		0.004683	0.004681	0.004681		0.032541	0.032548	0.032525
CI^{+}	0.0100	0.010624	0.010623	0.010623	0.028925	0.029880	0.029873	0.029858
CI-		0.009492	0.009493	0.009493		0.027970	0.027977	0.027950

Table 2: Algorithm B

B	FAR	Variance	Binomial	Bootstrap	FRR	Variance	Binomial	Bootstrap
CI^{+}	0.0001	0.000157	0.000157	0.000158	0.019234	0.020024	0.020011	0.020038
CI-		0.000043	0.000043	0.000050		0.018444	0.018457	0.018479
CI^{+}	0.0005	0.000635	0.000635	0.000642	0.014042	0.014708	0.014708	0.014683
CI-		0.000381	0.000381	0.000383		0.013376	0.013376	0.013350
CI^{+}	0.0010	0.001206	0.001206	0.001208	0.012425	0.013053	0.013052	0.013083
CI-		0.000844	0.000844	0.000850		0.011797	0.011798	0.011817
CI^{+}	0.0048	0.005189	0.005191	0.005175	0.009883	0.010445	0.010443	0.010442
CI-		0.004411	0.004409	0.004408		0.009321	0.009323	0.009358
CI^{+}	0.0109	0.011497	0.011496	0.011500	0.008683	0.009208	0.009208	0.009208
CI-		0.010319	0.010320	0.010350		0.008158	0.008158	0.008158

Table 3: Algorithm C

C	FAR	Variance	Binomial	Bootstrap	FRR	Variance	Binomial	Bootstrap
CI^{+}	0.0001	0.000157	0.000157	0.000158	0.005292	0.005701	0.005703	0.005708
CI-		0.000043	0.000043	0.000050		0.004883	0.004881	0.004875
CI^{+}	0.0005	0.000647	0.000646	0.000650	0.004308	0.004678	0.004679	0.004692
CI-		0.000387	0.000388	0.000392		0.003938	0.003937	0.003942
CI^{+}	0.0010	0.001198	0.001197	0.001200	0.003842	0.004192	0.004192	0.004188
CI-		0.000836	0.000837	0.000842		0.003492	0.003492	0.003500
CI^{+}	0.0049	0.005342	0.005339	0.005333	0.003058	0.003371	0.003370	0.003375
CI-		0.004542	0.004545	0.004558		0.002745	0.002746	0.002742
CI^{+}	0.0102	0.010822	0.010820	0.010846	0.002767	0.003064	0.003064	0.003067
CI-		0.009678	0.009680	0.009683		0.002470	0.002470	0.002475

[3] CI^{+} is the upper bound of the confidence intervals, CI- is the lower bound.

5. Conclusion and Findings

We can summarize with the following findings:

- Figures 3, and 4 illustrate that the mean variance calculated from the simple random sample method are almost identical to the variance from binomial estimates.
- Figures 5, and 6 illustrate that the CI computed from mean variance of simple random sample and the binomial estimate almost identical. Thus this validates equation (2) and (4), i.e., the mean of empirical variance is the binomial estimate variance.
- Figures 7 and 8 illustrate the 95 % CIs of FAR and FRR for all target thresholds from algorithm C. The figures show the CI from bootstrap coincides with the others.

As predicted by statistical asymptotic theory, we found no significant difference in the confidence intervals of FAR and FRR among all three approaches: the mean of estimated variance, the binomial estimation, and the bootstrap. In Appendix B, Tables 6-8 show the CI from dataset 2 of evaluation results on finger 01 with similar results, i.e., without any significant difference.

Therefore, we can adopt the Binomial approach by

- locating the FRR from a specific FAR from DET curve of the test scores,
- applying FRR to equation (7) to obtain the CI for FRR.

The binomial approach avoids the long computing time by the bootstrap simulations.

Nevertheless, there are some questions which arise from the above analysis.

1. Can we apply the binomial estimated confidence interval equation to a much smaller sample size dataset? Will there be any difference between parametric and non-parametric estimates? If not, how much smaller can the dataset size be so that it still gives a good estimate?

Two subsets of the original 120K sample dataset with the size of 60K and 20K are selected randomly to compute their confidence intervals from the same procedures. The partition size n has been adjusted in the process from the different sample sizes. The resulting CIs of FRR for dataset 2 of finger 01 are shown from Figure 9 and 10.

Once again, Figures 9 and 10 show that there is no significant difference among the three estimates. Yes, binomial estimation can be applied to smaller data samples. But the measured error rate has to be at least 1/n, where n is the sample size.

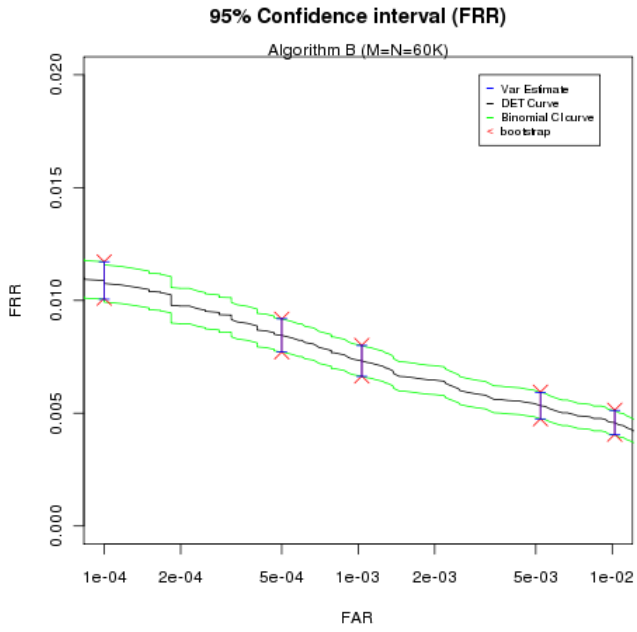

Figure 9

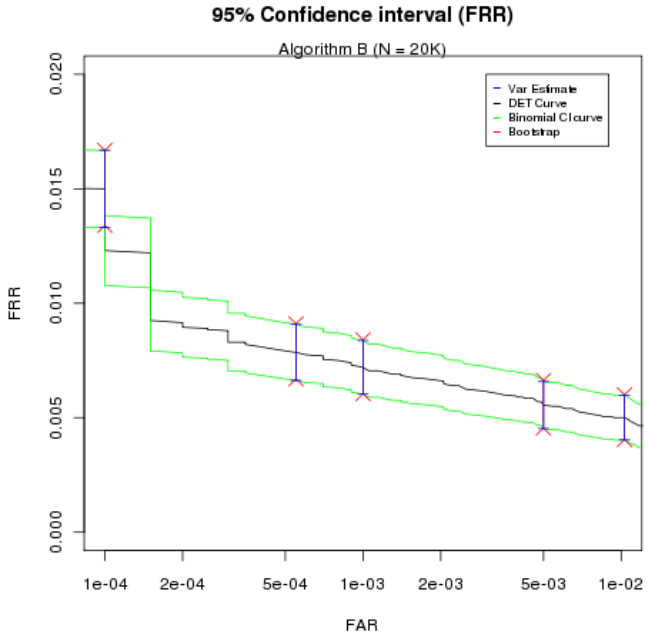

Figure 10

Plots of all the CI of FRR from three sample sizes: 120K, 60K and 20K along with their respective DETs are shown in Figure 11.

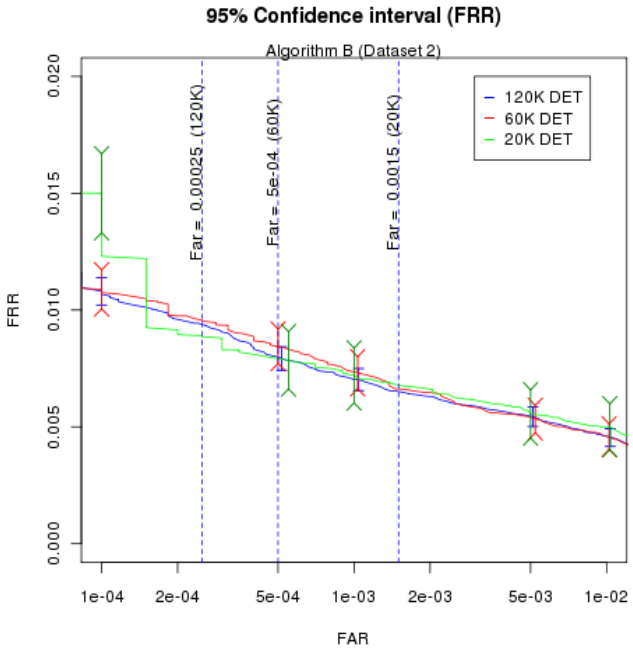

Figure 11

The confidence intervals do indeed converge from 20K sample to 120K sample but not symmetrically. The DET curves coincide each other until after the tail end of FAR value (< 2e-4) which is slightly greater than 0.01 %.

Minimum error for the sample size M (=120,000) is 1/M (= 8.3e-6) which is less than FAR = 1e-4 in Figure 11.

Adoption of the “Rule of 30” for the biometric test size was initially suggested in [7] and is supported in [8]. It gives the lower bounds of errors of a test for a given level of accuracy. For example, If the FAR is 0.01 %, then the sample size will need to be 300K to have at least 30 errors to comply with the rule. In Figure 11, three vertical dotted lines represent the minimal FAR values, {2.5e-4, 5e-4, 1.4e-3} from the respective sample sizes, {120K, 60K, and 20K} according to “Rule of 30”.

From the observation of Figure 11, there are too fewer data points at the tail end of the DET for the smaller size sample where the CI is far from that of the larger size sample. There are some research papers on the estimation tails of probability distributions and the confidence region around a very smaller error rate as in [8]. Further investigation of that aspect is warranted in future.

However, in our study, the experiment dataset sample size is 120K. If we follow this rule of 30, the guideline suggests that we can only guarantee the error rate

0.025 % for FAR instead of 0.01 % that we are estimating here. This leads us to the question “Do we really need a sample size of 300K in order to report FRR with a CI of 95 % at FAR = 0.01 % ?”

2. If by doubling the imposter scores but with the same number of genuine scores in the experiment with the same protocol, will the binomial estimator still give accurate estimate confidence intervals?

For example, we let M = 60K, and N =120K for sample scores sets, then the confidence intervals are shown in Figure 12 and 13.

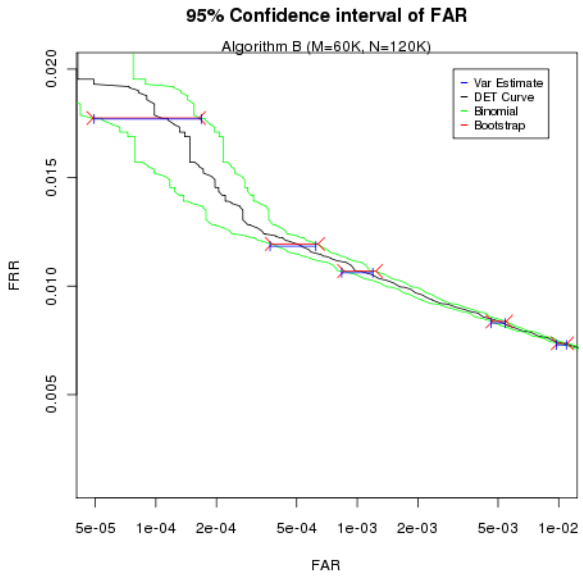

Figure 12

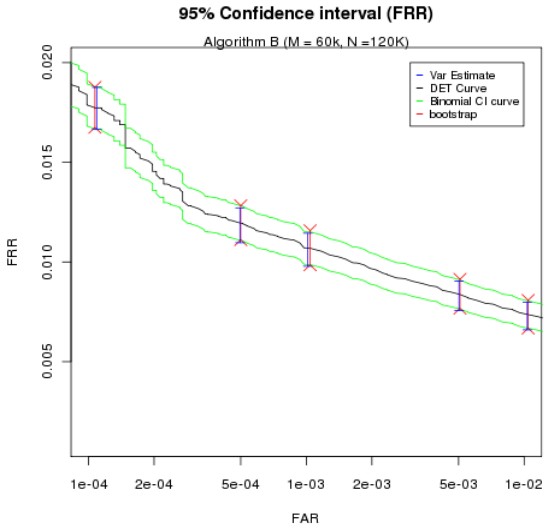

Figure 13

Once again, all three methods illustrate the similar confidence intervals. The

confidence intervals of FRR of the 60Kx120K (genuine, impostor sizes) showed the similar range of FRR of 60Kx60K. (In appendix B, table 9, and 10). We can therefore apply binomial estimation to compute the confidence interval which is based on the sample size and the corresponding FAR from the DET curve but with the different FRR value.

If we apply the same analogy to 120Kx120K and 120Kx300K, can we predict the accurate CI range for 120Kx300K from 120Kx120K? Will it look like Figure 11 where the FRR of 60KX60K approximately equal to 120Kx120K? It's possible, but unless we increase the imposter score to 300K, we don't know.

From the aspect of the "Rule of 30", Table 1-3 and 6-8 show that CIs are very compatible with all three methods where the sample size is 120K which only provides 12 errors for FAR = 10^{-4}. Therefore we do not find "Rule of 30" relevant for the PFTII tests.

In this study, we have selected a target set $\mathcal{T}$ of thresholds of interested FAR values to compute the CI of FRR. We will investigate further by selecting a set of interested FRR values to compute the confidence intervals of FAR by the same procedures.

6. References

[1] R. Bolle, N. Ratha, S Pankanti: Evaluating Authentication System Using Bootstrap Confidence Intervals. ProutoID'99 (1999)

[2] R. Bolle, N. Ratha, S Pankanti: Error analysis of pattern Recognition systems - the subset bootstrap. CVIU IEEE 2003.

[3] Jin Chu Wu: operational Measures and Accuracies of ROC curve on Large Fingerprint Datasets, NISTIR 7495, 2008

[4] R. Micheals, T. Boult: Improving variance Estimation in Biometric Systems. Computer Vision and pattern Recognition 2007 IEEE.

[5] R. Micheals: Biometric System Evaluation. A Thesis of Doctor of Philosophy in Computer Science, 2003, Lehigh University.

[6] R. Micheals, T. Boult: Is the Urn well-Mixed? Uncovering FaLse Cofactor Homogeneity Assumptions in Evaluation, NISTIR 7156, 2004.

[7] Doddington et al.: Sheep , Goats, Lambs, Wolfs: An Analysis of individual Differences in Speaker Recognition Performance, Proc. Int. Conf. on Speech and language processing, 1998

[8] A J Mansfield, J L Wayman: Best Practices in Testing and Reporting Performance of Biometric Devices, Version 2.01 2002. National Physical Laboratory Report, Centre for Mathematics and Scientific Computing, 14/02.

[9] R Smith; Estimating Tails of Probability Distributions, The Annals of Statitstics,1987,Vol. 15,No. 3, 1174-1207.

[10] W. Hoeffding: On the Distribution of the Number of Successes in Independent Trials. The Annals of Mathematical Statistics, Vol. 27, No.3 (Sep. 1956), pp.713-721.

7. Appendix A Table 4 Normality test P-Values

Algorithm	Dataset	FAR Value	Median FAR p-value (Shapiro)	Median FRR p-value (Shapiro)	Median FAR p-value (Anderson)	Median FRR p-value (Anderson)	Median Multivariate p-value (Shapiro)
A	2	0.0001	0.000105	0.000002	0.519552	0.000002	0.067715
A	2	0.0005	0.100611	0.074713	0.481129	0.074713	0.259513
A	2	0.0010	0.238208	0.191036	0.498212	0.191036	0.312403
A	2	0.0050	0.434978	0.409187	0.495715	0.409187	0.329778
A	2	0.0100	0.477359	0.458813	0.474295	0.458813	0.320820
A	1	0.0001	0.000089	0.000002	0.478007	0.000002	0.064373
A	1	0.0005	0.099951	0.072191	0.505644	0.072191	0.261747
A	1	0.0010	0.237872	0.196611	0.498333	0.196611	0.301616
A	1	0.0050	0.453799	0.431782	0.509553	0.431782	0.362946
A	1	0.0100	0.478621	0.449979	0.518082	0.449979	0.362283
B	2	0.0001	0.000082	0.000002	0.494205	0.000002	0.054991
B	2	0.0005	0.100256	0.070558	0.462484	0.070558	0.274883
B	2	0.0010	0.243477	0.208760	0.447962	0.208760	0.312299
B	2	0.0050	0.417540	0.387144	0.428931	0.387144	0.312714
B	2	0.0100	0.453561	0.427894	0.431700	0.427894	0.343890
B	1	0.0001	0.000105	0.000002	0.450043	0.000002	0.075009
B	1	0.0005	0.094934	0.073585	0.456605	0.073585	0.254219
B	1	0.0010	0.236580	0.189092	0.464645	0.189092	0.296973
B	1	0.0050	0.431710	0.418430	0.464428	0.418430	0.337072
B	1	0.0100	0.471834	0.448874	0.473857	0.448874	0.340417
C	2	0.0001	0.000116	0.000002	0.403654	0.000002	0.053667
C	2	0.0005	0.103226	0.075246	0.399971	0.075246	0.278797
C	2	0.0010	0.227611	0.187754	0.365675	0.187754	0.293249
C	2	0.0050	0.426359	0.419611	0.356065	0.419611	0.306368
C	2	0.0100	0.451815	0.434846	0.343314	0.434846	0.309760
C	1	0.0001	0.000078	0.000002	0.445476	0.000002	0.035026
C	1	0.0005	0.110294	0.080157	0.429163	0.080157	0.289135
C	1	0.0010	0.251130	0.197729	0.407773	0.197729	0.305428
C	1	0.0050	0.036931	0.015718	0.371635	0.015718	0.243925
C	1	0.0100	0.515979	0.483681	0.382362	0.483681	0.376075

The median p-values from 1000 experiments show that majority of the p-values from Wilks-Shapiro and Anderson Darling tests indicate that FAR and FRR are normally distributed. There are exceptions at FAR = 0.0001 which reject the normality of the data due to insufficient errors generated from the data or p being too small for the normal approximation while the Poisson distribution might give a better approximation.

Note: Anderson, Shapiro normal and multivariate tests are implemented from R software.

Table 5 – Correlation Coefficient ρ

Algorithm	Dataset	FAR mean	FRR mean	Minimum ρ	Twenty five Percentile ρ	Median ρ	Seventy five Percentile ρ	Maximum ρ
A	2	0.0001	0.0456	-0.7715	-0.1773	-0.0040	0.1570	0.6886
A	2	0.0005	0.0331	-0.6931	-0.1621	-0.0129	0.1605	0.5638
A	2	0.0010	0.0304	-0.6451	-0.1702	-0.0131	0.1580	0.7267
A	2	0.0050	0.0238	-0.6457	-0.1646	0.0014	0.1710	0.6704
A	2	0.0101	0.0214	-0.5796	-0.1512	-0.0017	0.1729	0.8304
A	1	0.0001	0.0701	-0.6612	-0.1682	-0.0083	0.1619	0.5968
A	1	0.0005	0.0541	-0.7198	-0.1568	0.0032	0.1823	0.6200
A	1	0.0011	0.0454	-0.7338	-0.1480	0.0034	0.1632	0.6880
A	1	0.0051	0.0336	-0.6514	-0.1562	0.0123	0.1906	0.6638
A	1	0.0101	0.0289	-0.6554	-0.1408	0.0180	0.1628	0.6926
B	2	0.0001	0.0108	-0.6731	-0.1607	-0.0029	0.1643	0.7789
B	2	0.0005	0.0079	-0.7463	-0.1683	0.0006	0.1702	0.6310
B	2	0.0010	0.0070	-0.6563	-0.1750	-0.0068	0.1488	0.6795
B	2	0.0051	0.0054	-0.7164	-0.1935	-0.0158	0.1621	0.6466
B	2	0.0103	0.0045	-0.7586	-0.1804	-0.0279	0.1334	0.6403
B	1	0.0001	0.0192	-0.6669	-0.1519	0.0051	0.1630	0.6448
B	1	0.0005	0.0140	-0.6260	-0.1591	0.0074	0.1707	0.6620
B	1	0.0010	0.0124	-0.7034	-0.1528	0.0110	0.1689	0.7123
B	1	0.0048	0.0099	-0.7131	-0.1606	-0.0031	0.1462	0.6663
B	1	0.0109	0.0087	-0.7330	-0.1498	-0.0014	0.1415	0.6310
C	2	0.0001	0.0034	-0.6199	-0.1365	0.0139	0.1630	0.7567
C	2	0.0005	0.0028	-0.7775	-0.1291	0.0239	0.1710	0.6400
C	2	0.0010	0.0026	-0.7323	-0.1499	0.0067	0.1483	0.6384
C	2	0.0055	0.0023	-0.6970	-0.1802	-0.0115	0.1530	0.6876
C	2	0.0096	0.0022	-0.6425	-0.1704	-0.0118	0.1582	0.6866
C	1	0.0001	0.0053	-0.6384	-0.1510	-0.0001	0.1673	0.7708
C	1	0.0005	0.0043	-0.7032	-0.1757	-0.0107	0.1410	0.6544
C	1	0.0010	0.0038	-0.6571	-0.1829	-0.0215	0.1438	0.6724
C	1	0.0049	0.0031	-0.7959	-0.3818	-0.2330	-0.0635	0.5865
C	1	0.0102	0.0028	-0.6596	-0.1774	-0.0195	0.1287	0.5976

Appendix B: Confidence Intervals for Dataset 2, finger 01 (120Kx120K)

Table 6 - Algorithm A

A	FAR	Variance	Binomial	bootstrap	FRR	Variance	Binomial	Bootstrap
CI^+	0.000100	0.000157	0.000157	0.000158	0.045600	0.046781	0.046780	0.046771
CI^-		0.000043	0.000043	0.000050		0.044419	0.044420	0.044483
CI^+	0.000517	0.000644	0.000646	0.000650	0.033083	0.034093	0.034095	0.034079
CI^-		0.000390	0.000388	0.000383		0.032073	0.032071	0.032033
CI^+	0.001000	0.001176	0.001179	0.001192	0.030375	0.031343	0.031346	0.031325
CI^-		0.000824	0.000821	0.000825		0.029407	0.029404	0.029408
CI^+	0.005042	0.005443	0.005443	0.005442	0.023792	0.024653	0.024654	0.024633
CI^-		0.004641	0.004641	0.004658		0.022931	0.022930	0.022958
CI^+	0.010150	0.010718	0.010717	0.010717	0.002145	0.022267	0.022270	0.022258
CI^-		0.009582	0.009583	0.009583		0.020633	0.020630	0.020667

Table 7 - Algorithm B

B	FAR	Variance	Binomial	bootstrap	FRR	Variance	Binomial	Bootstrap
CI^+	0.000100	0.000157	0.000157	0.000167	0.010883	0.011386	0.011385	0.011379
CI^-		0.000043	0.000043	0.000050		0.010214	0.010215	0.010221
CI^+	0.000500	0.000646	0.000646	0.000650	0.008450	0.008426	0.008427	0.008433
CI^-		0.000388	0.000388	0.000392		0.007424	0.007423	0.007450
CI^+	0.001033	0.001223	0.001225	0.001225	0.007317	0.007497	0.007498	0.007500
CI^-		0.000861	0.000859	0.000867		0.006553	0.006552	0.006558
CI^+	0.005217	0.005510	0.005511	0.005500	0.005333	0.005849	0.005849	0.005842
CI^-		0.004706	0.004705	0.004717		0.005017	0.005017	0.005021
CI^+	0.010233	0.010911	0.010914	0.010900	0.004583	0.004923	0.004922	0.004925
CI^-		0.009773	0.009770	0.009775		0.004161	0.004162	0.004183

Table 8 - Algorithm C

C	FAR	Variance	Binomial	bootstrap	FRR	Variance	Binomial	Bootstrap
CI^+	0.000100	0.000156	0.000157	0.000167	0.003442	0.003774	0.003773	0.003792
CI^-		0.000044	0.000043	0.000050		0.003110	0.003111	0.003108
CI^+	0.000517	0.000645	0.000646	0.000650	0.002758	0.003054	0.003055	0.003067
CI^-		0.000389	0.000388	0.000392		0.002462	0.002461	0.002487
CI^+	0.001000	0.001178	0.001179	0.001183	0.002625	0.002914	0.002915	0.002925
CI^-		0.000822	0.000821	0.000825		0.002336	0.002335	0.002350
CI^+	0.005508	0.005926	0.005927	0.005933	0.002283	0.002552	0.002553	0.002533
CI^-		0.005090	0.005089	0.005092		0.002014	0.002013	0.002004
CI^+	0.009558	0.010111	0.010109	0.010133	0.002208	0.002472	0.002474	0.002483
CI^-		0.009005	0.009007	0.009004		0.001944	0.001942	0.001958

Table 9: 60Kx60K (Subset of Dataset 1, finger 02)

B	FAR	Variance	Binomial	bootstrap	FRR	Variance	Binomial	Bootstrap
CI^+	0.000109	0.000180	0.000180	0.000177	0.010800	0.018727	0.018738	0.018771
CI^-		0.000020	0.000020	0.000032		0.016639	0.016628	0.016707
CI^+	0.000496	0.000700	0.000699	0.000710	0.007925	0.012671	0.012682	0.012855
CI^-		0.000334	0.000335	0.000339		0.010963	0.010952	0.011116
CI^+	0.001016	0.001273	0.001272	0.001323	0.007025	0.011390	0.011402	0.011563
CI^-		0.000761	0.000762	0.000806		0.009776	0.009764	0.009914
CI^+	0.005022	0.005357	0.005353	0.005831	0.005433	0.009127	0.009130	0.009101
CI^-		0.004243	0.004247	0.004677		0.007673	0.007670	0.007655
CI^+	0.010338	0.012640	0.012630	0.012726	0.004542	0.007856	0.007859	0.007964
CI^-		0.010894	0.010904	0.011000		0.006510	0.006507	0.006598

Table 10: 60Kx120K (Dataset 1, finger 02)

B	FAR	Variance	Binomial	bootstrap	FRR	Variance	Binomial	Bootstrap
CI^+	0.000100	0.000169	0.000168	0.000164	0.015000	0.018759	0.018758	0.018763
CI^-		0.000049	0.000050	0.000049		0.016647	0.016648	0.016756
CI^+	0.000550	0.000623	0.000622	0.000639	0.007850	0.012714	0.012707	0.012823
CI^-		0.000369	0.000370	0.000377		0.010968	0.010975	0.011116
CI^+	0.001000	0.001195	0.001196	0.001238	0.007201	0.011444	0.011435	0.011571
CI^-		0.000837	0.000836	0.000861		0.009786	0.009795	0.009857
CI^+	0.005000	0.005420	0.005422	0.005484	0.005553	0.009035	0.009023	0.009134
CI^-		0.004624	0.004622	0.004689		0.007559	0.007571	0.007679
CI^+	0.010300	0.010918	0.010910	0.011009	0.005005	0.007982	0.007971	0.008077
CI^-		0.009758	0.009766	0.009882		0.006598	0.006609	0.006680