

**NISTIR 7730**

**Further Studies  
of Bootstrap Variability  
for ROC Analysis on Large Datasets**

*Jin Chu Wu  
Alvin F. Martin  
Raghu N. Kacker*

**NISTIR 7730**

**Further Studies  
of Bootstrap Variability  
for ROC Analysis on Large Datasets**

*Jin Chu Wu  
Alvin F. Martin  
Raghu N. Kacker*

October 2010



U.S. Department of Commerce  
*Gary Locke, Secretary*

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

# Further Studies of Bootstrap Variability for ROC Analysis on Large Datasets

Jin Chu Wu<sup>\*a</sup>, Alvin F. Martin<sup>a</sup> and Raghu N. Kacker<sup>b</sup>

<sup>a</sup>Information Access Division, <sup>b</sup>Applied and Computational Mathematics Division,  
Information Technology Laboratory,  
National Institute of Standards and Technology, Gaithersburg, MD 20899

**Abstract** – The nonparametric two-sample bootstrap is successfully applied to computing the measurement uncertainties in receiver operating characteristic (ROC) analysis on large datasets in areas such as biometrics, speaker recognition system, etc. To determine the number of bootstrap replications in our applications, the bootstrap variability related to standard error and two bounds of 95 % confidence interval was studied in a scenario where the statistic of interest was the true accept rate (TAR) of the genuine scores at a specified false accept rate (FAR) of the impostor scores. From the operational perspective, three more scenarios are of interest, in which the statistics are the TAR at a given threshold value, the FAR at a specified threshold value, and the equal error rate, respectively. Regarding the ROC analysis, the area under ROC curve is also of interest. In this article, the bootstrap variability was studied in all these five scenarios concerning both high- and low-accuracy matching algorithms. With the tolerance 0.02 of the coefficient of variation, which can be applied to all cases investigated, it is found that 2000 bootstrap replications are appropriate for ROC analysis on large datasets in order to reduce the bootstrap variance and ensure the accuracy of the computation.

**Index Terms** – Bootstrap, variability, ROC analysis, biometrics, speaker recognition, standard error, confidence interval, large datasets.

---

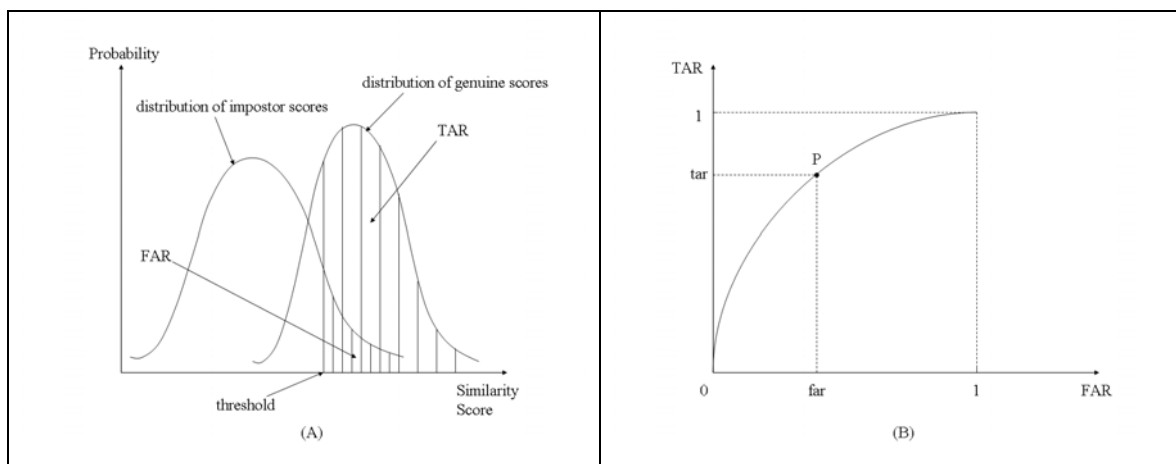
\* Tel: + 301-975-6996; fax: + 301-975-5287. E-mail address: [jinchu.wu@nist.gov](mailto:jinchu.wu@nist.gov).

## 1 Introduction

The receiver operating characteristic (ROC) analysis is employed in many applications as a useful statistical technique. Sampling variability can result in uncertainties of measures in ROC analysis. Thus, it is important to take account of the measurement uncertainties when evaluating and comparing the performance of algorithms. The nonparametric two-sample bootstrap is successfully applied to computing the measurement uncertainties in ROC analysis on large datasets in areas such as biometrics, speaker recognition system, etc [1-4].

Generally speaking, for instance in biometrics, genuine scores are created by comparing two different images of the same subject, and impostor scores are generated by matching two images of two different subjects. Both scores may be referred to as similarity scores. These two sets of similarity scores constitute two distributions, respectively, as schematically depicted in Figure 1 (A) for continuous similarity scores.

The cumulative probabilities of genuine and impostor scores from the highest similarity score to a specified similarity score (i.e., threshold) are defined as the true accept rate (TAR) and the false accept rate (FAR), respectively. Thus, in the FAR-and-TAR coordinate system, while the threshold moves from the highest similarity score down to the lowest similarity score, an ROC curve is constructed as drawn in Figure 1 (B).



**Figure 1 (A): A schematic diagram of two distributions of continuous genuine scores and impostor scores, showing three related variables: TAR, FAR, and threshold. (B): A schematic drawing of an ROC curve constructed by moving the threshold from the highest similarity score down to the lowest one.**

Any point P on an ROC curve has two coordinates FAR and TAR and is associated with a threshold through two distributions of genuine scores and impostor scores. The three variables, FAR, TAR, and threshold, are related to each other, as illustrated in Figure 1 (A) and (B). Any one of these three variables can determine the other two variables. In practice, it is never required that TAR be specified in the first place. Thus, from the operational perspective, the three scenarios are of interest, that are measuring the TAR at a specified FAR, measuring the TAR at a given threshold value, and measuring the FAR at a specified threshold value, respectively.

The equal error rate (EER) is defined to be  $1 - \text{TAR}$  (i.e., the probability of type I error) or FAR (i.e., the probability of type II error) when they are equal. As is well-known, these two error rates are traded-off of each other. Generally speaking, the smaller the EER is, the more apart the two distributions of genuine scores and impostor scores are, thus the higher the ROC curve is and the more accurate the matching algorithm is [5, 6]. Therefore, from the operational perspective, the fourth scenario that is of interest is measuring the EER.

In addition, an ROC curve can be measured by employing the area under the ROC curve (AURC), which corresponds to the probability of correctly identifying which of the two stimuli is more likely than the other, and measures the overall ROC curve [5, and references therein]. If it is computed using the trapezoidal rule, the AURC is equivalent to the Mann-Whitney statistic that is formed by genuine and impostor scores. Hence, the variance of the Mann-Whitney statistic can be utilized as the variance of the AURC. On the other hand, the standard error (SE) of AURC can also be computed using the nonparametric two-sample bootstrap [7]. Moreover, the metric AURC is widely employed in areas such as medical decision making, even though the sizes of datasets are far less than those in our applications. Hence, the fifth scenario that is of interest is measuring the AURC.

As extensively investigated [5], there is usually no underlying parametric distribution function for genuine and impostor scores; the distributions of genuine scores and impostor scores are considerably different in general; and the distributions vary substantially from algorithm to algorithm in a way that differentiates algorithms in terms of qualities. This suggests that nonparametric analysis is pertinent to evaluating matching algorithms on large-scale datasets.

An ROC curve is characterized by the relative relationship between the distributions of the genuine scores and the impostor scores [5, 6]. These two distribution functions are interrelated by the algorithm that generates them. The performance of a matching algorithm is determined not only by its ability of executing genuine matching but also by its ability of executing impostor matching. All statistics of interest are influenced by the combined impact of these two distributions. As a result, computing the measurement uncertainties in ROC analysis is a two-distribution issue rather than a one-distribution issue. Thus, the nonparametric two-sample bootstrap is employed to compute the uncertainties of measures, in terms of SEs and confidence intervals (CI), in all five scenarios stated above. The two samples are referred to as a set of genuine scores and a set of impostor scores.

One of the important parameters regarding bootstrap methods is the number of bootstrap replications. It is intrinsically related to bootstrap variability. As investigated in the literature [8-11], the substantial bootstrap variance is caused by the sampling variability as well as the bootstrap resampling variability. The former is because the sample size is finite and limited, and the latter is because the number of bootstrap replications is not infinite. The bootstrap variance results in the variances of, for example, the SE and the lower bound and upper bound of CI of the distribution formed by bootstrap replications of the statistic of interest. As a consequence, these variances can be functions of the sample size as well as the number of bootstrap replications. Inversely, the sample size and the number of bootstrap replications can be determined by studying the variances of SE and the two bounds of the CI of two-sample bootstrap distribution of the statistic of interest.

Regarding the sample sizes, in this article, the total number of genuine scores is a little over 60 000 and the total number of impostor scores is a little over 120 000. They are fixed based on our previous studies [12]. The research was carried out using Chebyshev's inequality on two statistics of interest, namely, the TAR at an operational FAR and AURC. It was found that if the numbers of similarity scores increased from what were used, the measurement accuracy would improve little.

In our applications, as pointed out above, it is inappropriate to assume the normality for distributions of similarity scores. The statistics of interest are probabilities such as TAR, FAR, EER, and AURC rather than a simple arithmetic mean. Moreover, the sizes of datasets are much larger than those encountered in other applications, such as medical application, etc [11]. Hence, in order to reduce the bootstrap variance and ensure the accuracy of computation, and subsequently determine the appropriate number of two-sample bootstrap replications in our applications, the bootstrap variability was studied in Ref. [1]. However, it was carried out only in one scenario where the statistic of interest was the TAR at a specified FAR. In this article, the bootstrap variability studies will be conducted in all other four scenarios as well.

As is well-known, the bootstrap method assumes that an independent and identically distributed (i.i.d.) random sample of size  $n$  is drawn from a population with its own probability distribution. Our large government data bases used for developing similarity scores in fingerprint technology were randomly collected from real practice rather than using multiple acquisitions and thus had no dependencies. Moreover, our studies showed that our data bases have no dependencies [7]. Thus, the random sample is assumed to be i.i.d. in our work.

With the i.i.d. assumption, the bootstrap objects are individuals in the sample. Otherwise, the bootstrap objects are the subsets of the sample into which the sample is grouped based on data dependencies caused by multiple biometric acquisitions [11, 13, 14]. This can preserve the dependencies among the data. However, everything else in the bootstrap method remains intact. Of course, how the sample is grouped into subsets will have impact on the bootstrap results. As a matter of fact, from the statistical point of view, the sample should be collected as randomly as possible.

All similarity scores are converted to integer scores if they are not already [5]. Thus, the probability distribution functions of similarity scores are all discrete, and the ROC curve is not a smooth curve. Since there is usually no underlying parametric distribution function for similarity scores, the empirical distribution is assumed for each of the observed scores.

As opposed to continuous distribution some concepts and definitions need to be established and modified accordingly. For instance, first, ties of genuine scores and/or impostor scores can often occur on large fingerprint data sets and thus must be taken into account while computing the estimated TAR at a specified FAR. Second, when computing the cumulative discrete probability at a score, the probability of this score must be taken into account [15]. Third, generally speaking there does not exist such a similarity score (range) at which the probabilities of the type I error and the type II error are exactly equal. All related formulas for computing statistics of interest in this article can be found in Refs. [2, 5, and references therein].

The methods involving the nonparametric two-sample bootstrap and bootstrap variability studies are presented in Section 2. The results of bootstrap variability studies in all five scenarios for both high- and low-accuracy algorithms<sup>1</sup> are provided in Section 3, which determine the number of bootstrap replications for ROC analysis on large datasets. The conclusions and discussion can be found in Section 4. Tables and figures are shown in Appendices 1 and 2, respectively.

## 2 Methods

### 2.1 The formulations of discrete distribution functions of similarity scores

Without loss of generality, the similarity scores are expressed inclusively using the integer score set  $\{s\} = \{s_{\min}, s_{\min}+1, \dots, s_{\max}\}$ , running consecutively from the lowest score  $s_{\min}$  to the highest score  $s_{\max}$ . The genuine score set and the impostor score set are denoted as

$$\mathbf{G} = \{ m_i \mid m_i \in \{s\} \text{ and } i = 1, \dots, N_G \}, \quad (1)$$

and

$$\mathbf{I} = \{ n_i \mid n_i \in \{s\} \text{ and } i = 1, \dots, N_I \}, \quad (2)$$

where  $N_G$  and  $N_I$  are the total numbers of genuine scores and impostor scores, respectively. These two sets,  $\mathbf{G}$  and  $\mathbf{I}$ , constitute two discrete probability distribution functions of genuine scores and impostor scores, respectively.

### 2.2 An algorithm for the nonparametric two-sample bootstrap

The nonparametric two-sample bootstrap [8, 11] is employed to compute the estimates of measurement uncertainties in all five scenarios. The algorithm is as follows.

#### *Algorithm 1* (Nonparametric two-sample bootstrap)

- 1: **for**  $i = 1$  **to**  $B$  **do**
- 2:     select  $N_G$  scores randomly WR from  $\mathbf{G}$  to form a set  $\{\text{new } N_G \text{ genuine scores}\}_i$
- 3:     select  $N_I$  scores randomly WR from  $\mathbf{I}$  to form a set  $\{\text{new } N_I \text{ impostor scores}\}_i$
- 4:      $\{\text{new } N_G \text{ genuine scores}\}_i$  &  $\{\text{new } N_I \text{ impostor scores}\}_i \Rightarrow \text{statistic } \hat{T}_i$
- 5: **end for**
- 6:  $\{\hat{T}_i \mid i=1, \dots, B\} \Rightarrow S\hat{E}_B$  and  $(\hat{Q}_B(\alpha/2), \hat{Q}_B(1-\alpha/2))$
- 7: **end**

where  $B$  is the number of two-sample bootstrap replications and WR stands for “with replacement”. The original genuine score set  $\mathbf{G}$  in Eq. (1) and impostor score set  $\mathbf{I}$  in Eq. (2) are generated by a matching algorithm. As shown from Step 1 to 5, this algorithm runs  $B$  times. In the  $i$ -th iteration,  $N_G$  scores are randomly selected WR from the original genuine score set  $\mathbf{G}$  to form a new set of  $N_G$  genuine scores,  $N_I$  scores are randomly selected WR from the original

---

<sup>1</sup> Specific hardware and software products identified in this report were used in order to adequately support the development of technology to conduct the performance evaluations described in this document. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

impostor score set  $\mathbf{I}$  to form a new set of  $N_I$  impostor scores, and then from these two new sets of similarity scores the  $i$ -th bootstrap replication of the estimated statistic of interest,  $\hat{T}_i$ , is generated.

The  $\hat{T}_i$  are different in different scenarios discussed in Section 1. In Scenario 1,  $\hat{T}_i = T\hat{A}R_i(f)$  at a specified  $f = \text{FAR}$ . In Scenario 2,  $\hat{T}_i = T\hat{A}R_i(t)$  at a given threshold  $t$ . In Scenario 3,  $\hat{T}_i = F\hat{A}R_i(t)$  at a given threshold  $t$ . In Scenario 4,  $\hat{T}_i = E\hat{E}R_i$ . And in Scenario 5,  $\hat{T}_i = A\hat{U}R_i$ . The formulas for computing all these five statistics of interest can be found in Refs. [2, 5, and references therein].

Finally, as indicated in Step 6, from the set  $\{\hat{T}_i \mid i=1, \dots, B\}$ , the estimator of the SE, denoted by  $S\hat{E}_B$ , i.e., the sample standard deviation of the  $B$  replications, and the estimators of the  $\alpha/2$  100 % and  $(1 - \alpha/2)$  100 % quantiles of the bootstrap distribution, denoted by  $\hat{Q}_B(\alpha/2)$  and  $\hat{Q}_B(1 - \alpha/2)$ , at the significance level  $\alpha$  can be calculated [11]. Definition 2 of quantile in Ref. [16] is adopted. That is, the sample quantile is obtained by inverting the empirical distribution function with averaging at discontinuities. Thus,  $(\hat{Q}_B(\alpha/2), \hat{Q}_B(1 - \alpha/2))$  stands for the estimated bootstrap  $(1 - \alpha)$  100 % CI. If 95 % CI is of interest, then  $\alpha$  is set to be 0.05.

### 2.3 An algorithm for empirical studies of nonparametric two-sample bootstrap variability

As pointed out in Section 1, it is important to re-study the variances of the SE and the two bounds of the CI of two-sample bootstrap distribution of the statistic of interest in our applications. To take into account the impact of the mean value, the coefficients of variation (CV) rather than just variance is employed. The empirical studies of bootstrap variability will be carried out in all five scenarios, respectively. Here is an algorithm for studies of bootstrap variability:

#### **Algorithm II (Bootstrap variability)**

```

1: for  $i = 1$  to  $L$  do
2:   for  $j = 1$  to  $B$  do
3:     select  $N_G$  scores randomly WR from  $\mathbf{G}$  to form a set {new  $N_G$  genuine scores} $_j$ 
4:     select  $N_I$  scores randomly WR from  $\mathbf{I}$  to form a set {new  $N_I$  impostor scores} $_j$ 
5:     {new  $N_G$  genuine scores} $_j$  & {new  $N_I$  impostor scores} $_j \Rightarrow$  statistic  $\hat{T}_{ij}$ 
6:   end for
7:    $\{\hat{T}_{ij} \mid j=1, \dots, B\} \Rightarrow S\hat{E}_{Bi}$  and  $(\hat{Q}_{Bi}(\alpha/2), \hat{Q}_{Bi}(1 - \alpha/2))$ 
8: end for
9:  $\{S\hat{E}_{Bi}, \hat{Q}_{Bi}(\alpha/2), \hat{Q}_{Bi}(1 - \alpha/2) \mid i = 1, \dots, L\} \Rightarrow C\hat{V}_{B,L}(\kappa), \kappa = S\mathbf{E}_{B,L}, \mathbf{Q}_{B,L}(\alpha/2), \mathbf{Q}_{B,L}(1 - \alpha/2)$ 
10: end

```

where  $L$  is the number of Monte Carlo iterations and  $B$  is the number of bootstrap replications. As indicated from Step 1 to 8, Algorithm II runs  $L$  iterations for a specified  $B$ . The part from



Step 2 to 7 is equivalent to the nonparametric two-sample bootstrap Algorithm I, which generates the  $i$ -th  $\hat{S}E_{B,i}$ ,  $\hat{Q}_{B,i}(\alpha/2)$  and  $\hat{Q}_{B,i}(1 - \alpha/2)$  of a statistic of interest in the  $i$ -th iteration for a given  $B$ . The statistics of interest in five scenarios are specified, respectively, in Section 2.2.

As shown in Step 9, for a specified  $B$ , after  $L$  iterations of executing two-sample bootstrap algorithm, the following three sets are generated,

$$\begin{aligned} \mathbf{SE}_{B,L} &= \{ \hat{S}E_{B,i} | i=1, \dots, L \}, \\ \mathbf{Q}_{B,L}(\alpha/2) &= \{ \hat{Q}_{B,i}(\alpha/2) | i=1, \dots, L \}, \\ \mathbf{Q}_{B,L}(1 - \alpha/2) &= \{ \hat{Q}_{B,i}(1 - \alpha/2) | i=1, \dots, L \}. \end{aligned} \quad (3)$$

Thereafter, from these three sets, three estimated  $\hat{C}V$  s of SE, lower-bound and upper-bound of CI, can be obtained, respectively,

$$\hat{C}V_{B,L}(\kappa) = \frac{\sqrt{\hat{V}AR_{B,L}(\kappa)}}{\hat{E}_{B,L}(\kappa)}, \text{ where } \kappa = \mathbf{SE}_{B,L}, \mathbf{Q}_{B,L}(\alpha/2), \mathbf{Q}_{B,L}(1 - \alpha/2). \quad (4)$$

It is clear that the three estimated  $\hat{C}V$  s are functions of the number of bootstrap replications  $B$  and the number of Monte Carlo iterations  $L$ , as well as the significance level  $\alpha$ . Therefore, the number of bootstrap replications  $B$  can be determined by the tolerable CVs. Then, the question is: How many iterations  $L$  are sufficient for a specified  $B$  to guarantee the accuracy of the Monte Carlo computation?

#### 2.4 Determine the number of Monte Carlo iterations $L$

Two fingerprint-image matching algorithms are employed. Algorithm 1 is of high accuracy, and Algorithm 2 is of low accuracy. The significance level is set to be 5 %. As discussed in Section 1, the total number of genuine scores is a little over 60 000 and the total number of impostor scores is a little over 120 000. With these sample sizes, in order to have statistical significance, in Scenario 1 the operational FAR is specified to be 0.001 [6, 12].

To show the operational significance for each algorithm, in Scenarios 2 and 3 the system threshold yielding a FAR 0.001 is chosen [2]. Thus, for Algorithm 1 employing integer scores, the threshold was set to be 455; for Algorithm 2 using real-number scores in  $[0.0, 1.0)$ , the threshold was set to be 0.634030. The estimates of the three CVs for the SE, lower bound and upper bound of 95 % CI are denoted by  $\hat{C}VSE$ ,  $\hat{C}VLB$ , and  $\hat{C}VUB$ , respectively.

In all five scenarios as classified in Section 2.2 and for each algorithm, the number of replications  $B$  was set to range from 200 up to 1000 at intervals of 200. Then, for each specified  $B$ , the number of Monte Carlo iterations  $L$  ran from 100 up to 1000 at intervals of 100, and thus 10 estimates of  $\hat{C}VSE$ s,  $\hat{C}VLB$ s, and  $\hat{C}VUB$ s were generated, respectively. From these 10 estimates, the minimum, maximum, and range of  $\hat{C}VSE$  s,  $\hat{C}VLB$ s, and  $\hat{C}VUB$  s for each specified  $B$  were obtained.

As shown in the odd numbered tables in Appendix 1, in general, both minimal  $\hat{CV}$ s and maximal  $\hat{CV}$ s decrease and the range between these two gets smaller as the number of replications  $B$  increases. For instance, in Table 1, for the SE, the ranges of 10 estimated  $\hat{CVSE}$ s vary from about 0.007 down to 0.002. For the lower bound and upper bound of 95 % CIs, the maximal  $\hat{CVLB}$ s and  $\hat{CVUB}$ s are less than 0.00007, and the ranges are not greater than 0.000008.

As a result, to obtain the estimates of CVs at a fixed number of replications  $B$  higher than 1000, the number of Monte Carlo iterations  $L$  does not need to run from 100 up to 1000 at intervals of 100. As estimating CVs, to save tremendous computing time, while the number of replications  $B$  varied from 1200 up to 2000 at intervals of 200, the number of Monte Carlo iterations  $L$  was fixed at 500. All the estimates of the corresponding CVs are shown in the even numbered tables in Appendix 1.

### 3 Results

All estimated  $\hat{CVSE}$ s,  $\hat{CVLB}$ s, and  $\hat{CVUB}$ s are presented in Table 1 through Table 20 in Appendix 1, and depicted in Figure 2 through Figure 16 in Appendix 2, respectively. In the cases where the number of replications  $B$  was set to be from 200 up to 1000 at intervals of 200, only the maximal  $\hat{CVSE}$ s,  $\hat{CVLB}$ s, and  $\hat{CVUB}$ s, as shown in the odd numbers of tables in Appendix 1, are employed. It shows that all estimated  $\hat{CVSE}$ s,  $\hat{CVLB}$ s, and  $\hat{CVUB}$ s decrease as the number of replications  $B$  increases. In order to determine the number of bootstrap replications  $B$ , the tolerances of CVs in different situations must be set.

#### 3.1 Set the tolerances of CVs

As defined in Eq. (4), the CV is a ratio of the SE to the mean, and thus its estimator is affected by both values. For the distributions of the SEs and lower bounds and upper bounds of 95 % CIs created by Monte Carlo iterations as shown in Eq. (3), respectively, the magnitudes of the estimated means are quite different with respect to different statistics of interest that are stated in Section 2.2, and thus the magnitudes of the estimated  $\hat{CV}$ s are also quite different. As a result, the tolerances of the CVs in different situations shall be set accordingly.

For instance, in Scenario 1 where the statistic of interest is the TAR at a given FAR, for high-accuracy Algorithm 1, the estimated  $\hat{SE}$ s of distributions of SEs, lower bounds and upper bounds of 95 % CIs, generated by 500 Monte Carlo iterations while the number of bootstrap replications  $B$  was set to be 2000 as discussed in Section 2.4, are 0.0000053, 0.0000198, and 0.0000192, respectively. It is demonstrated that the distribution of SEs is of less dispersion than the distributions of lower bounds and upper bounds of 95 % CIs, respectively. This is because in the tail of the distribution fewer samples occur [11].

The estimated means of the corresponding distributions are 0.000331, 0.992617, and 0.993913, respectively. Thus, the corresponding estimated CVs are 0.016040, 0.000020, and 0.000019, respectively, as presented in the last column of Table 2 in Appendix 1. It is noticed that the estimated mean for SEs is much less than 1, while on the contrary the estimated means for the

two bounds of 95 % CIs are close to 1. This is why the estimated  $\hat{CVSE}$  is much larger than the estimated  $\hat{CVLB}$  and  $\hat{CVUB}$  in Scenario 1 as presented in Table 1 through Table 4. Therefore, generally speaking, the tolerance for CVSE needs to be set larger than those for CVLB and CVUB [1].

The estimated  $\hat{CV}$ s for low-accuracy Algorithm 2 are in general greater than the corresponding  $\hat{CV}$ s for high-accuracy Algorithm 1, except for CVLB and CVUB in Scenario 4 where the statistic of interest is EER. This is due to the combined impact of the magnitudes of the estimated means and  $\hat{SE}$ s in Scenario 4. Another exception is that the estimated  $\hat{CVLB}$ s and  $\hat{CVUB}$ s in Scenarios 3 and 4 where the statistics of interest are FAR at a given threshold value and EER, respectively, are larger than those in other scenarios. This is because the magnitudes of the estimated means in these two scenarios are quite small.

All the tolerances of CVs in different situations can be found in Figure 2 through Figure 16 in Appendix 2. Some tolerances of CVs are larger than others. Nonetheless, the largest tolerance of CV is 0.02, which is set for all CVSEs in five scenarios for both high-accuracy Algorithm 1 and low-accuracy Algorithm 2.

### **3.2 The number of nonparametric two-sample bootstrap replications**

All estimated  $\hat{CVSE}$ s of Algorithms 1 and 2 in five scenarios are depicted from Figure 2 to Figure 6 in Appendix 2. As indicated in Section 3.1, the tolerance for all CVSEs is set to be 0.02. With this 0.02 tolerance, for instance, in Scenario 1 where the statistic of interest is TAR at a given FAR, as shown in Figure 2, 1400 two-sample bootstrap replications are sufficient for high-accuracy Algorithm 1, and 1800 replications are enough for low-accuracy Algorithm 2. In all other four scenarios for both Algorithms 1 and 2 as depicted in Figure 3 through Figure 6, with the tolerance 0.02, 1400 bootstrap replications are sufficient.

All estimated  $\hat{CVLB}$ s and  $\hat{CVUB}$ s in five scenarios are illustrated from Figure 7 to Figure 16, where each figure shows the results of one algorithm. For instance, the estimated  $\hat{CVLB}$  and  $\hat{CVUB}$  in Scenario 1 for Algorithm 1 are shown in Figure 7. As discussed in Section 3.1, the tolerances for CVLB and CVUB should be set smaller. Hence, if the tolerance is set to be at 0.000025, as indicated in Figure 7, 1400 replications can meet the requirement. Those for Algorithm 2 are depicted in Figure 8. As pointed out in Section 3.1, the tolerance for low-accuracy algorithms should be set larger. Thus, if the tolerance is set to be at 0.000450, 1400 replications can satisfy the restriction.

Regarding CVLB and CVUB in all other cases as depicted in Figure 9 through Figure 16 in Appendix 2, all tolerances set for CVs are less than 0.008 that is set in Scenario 3 where the statistic of interest is FAR at a given threshold value as shown in Figure 11 and Figure 12. For such tolerance settings, the least 1200 or 1400 or 1600 bootstrap replications can meet the constraints, respectively, depending on individual case.

It is clear that the largest tolerance employed so far is 0.02. Although the tolerance set for the CVSE is larger than those for CVLB and CVUB, the tolerance 0.02 for CVs is acceptable concerning our applications [11]. To reconcile numbers of bootstrap replications for all five scenarios where different statistics of interest are used as well as different qualities of matching algorithms, and further to be more conservative, it is suggested that 2000 nonparametric two-sample bootstrap replications be required in order to reduce the bootstrap variance and achieve statistical accuracy of the computation.

#### 4 Conclusions and discussion

In our applications, the normality assumption for distributions of similarity scores cannot be made, the statistics of interest are all probabilities rather than a simple sample mean, and the datasets are very large. Therefore, the bootstrap variability needs to be re-studied to determine the appropriate number of two-sample bootstrap replications in order to reduce the bootstrap variance and ensure the accuracy of the computation.

The nonparametric two-sample bootstrap variability related to SE and two bounds of 95 % CI of bootstrap distributions was empirically studied in five scenarios for ROC analysis, where the statistics of interest are TAR at a specified FAR, TAR at a given threshold value, FAR at a given threshold value, EER, and AURC, respectively. In addition, the bootstrap variability studies were conducted on both high-accuracy matching algorithm and low-accuracy matching algorithm.

To take into account the impact of the mean value, the CV rather than just variance is employed. All estimated  $\hat{CVSE}$ s,  $\hat{CVLB}$ s, and  $\hat{CVUB}$ s in five scenarios for both high-accuracy and low-accuracy algorithms were computed, and are presented in 20 tables and depicted in 15 figures accordingly. They decrease as the number of replications  $B$  increases.

The largest tolerance of CVs set so far is 0.02, which is acceptable in our applications. With this tolerance, to reconcile all cases and to be more conservative, it is suggested that the appropriate number of nonparametric two-sample bootstrap replications for ROC analysis on large datasets be 2000. It is worth mentioning that such extensively empirical studies of bootstrap variability involved numerous computations and thus took weeks of CPU time.

As pointed out in Section 1, the variance of two-sample bootstrap is also caused by the sample size. The sample sizes employed in this article were based on our previous studies [12]. Certainly, the 2000 number of bootstrap replications can also be applied to the cases where the sample sizes are smaller than what is used in this article. However, if for some reason the number of bootstrap replications needs to be reinvestigated, the empirical methods for studying the bootstrap variability developed in this article should remain the same.

In this article, TAR at a specified FAR and TAR at a given threshold value were discussed. In some literature [17], the false non-match rate (FNMR), which is equal to  $1 - TAR$ , at a given FAR or threshold value was employed. It is trivial to show that with respect to the same two new sets of similarity scores randomly selected WR from the two original sets of scores, respectively, as shown in the two-sample bootstrap Algorithm I in Section 2.2, the SE of FNMR is equal to the SE of TAR. But the lower bound and upper bound of 95 % CI for FNMR can be obtained by

interchanging two bounds for TAR and subtracting them from 1, respectively. Subsequently, the two bounds of 95 % CIs of FNMR are quite close to 0 as opposed to 1 that is what TAR does. Hence, if switching from TAR to FNMR, the CVSE will remain the same but the CVLB and CVUB of FNMR will be larger than those of TAR.

For instance, in Scenario 1 for both Algorithms 1 and 2 when the number of bootstrap replications  $B$  was set to be 2000 as shown in the last columns of Table 2 and Table 4, respectively, the estimated  $\hat{C}VLB$  and  $\hat{C}VUB$  of high-accuracy Algorithm 1 changed from 0.000020 and 0.000019 to 0.003152 and 0.002687, respectively; and those of low-accuracy Algorithm 2 changed from 0.000318 and 0.000389 to 0.001595 and 0.001196, respectively. However, they are all less than the tolerance 0.02. Hence, the assertion that the number of two-sample bootstrap replications be 2000 is still valid if FNMR is employed. In the meantime, it is worth pointing out that the CVLB and CVUB increase greatly if using FNMR instead of TAR.

In some applications, such as speaker recognition evaluation, etc, the statistic of interest is a detection cost function that is defined as a weighted sum of probabilities of type I error and type II error [4]. The variance of this kind of cost function involves a covariance term (i.e., the cross term) of correlated probabilities of type I error and type II error. Such a metric is not dealt with in this article. However, the probability of type I error (i.e., the FNMR at a given threshold value) and the probability of type II error (i.e., the FAR at a specified threshold value) have been coped with individually in this article.

## References

1. J.C. Wu, Studies of operational measurement of ROC curve on large fingerprint data sets using two-sample bootstrap, NISTIR 7449, National Institute of Standards and Technology, September, 2007.
2. J.C. Wu, Operational measures and accuracies of ROC Curve on large fingerprint data Sets, NISTIR 7495, National Institute of Standards and Technology, May, 2008.
3. J.C. Wu, A.F. Martin, R.N. Kacker, C.R. Hagwood, Significance test in operational ROC analysis, in Biometric Technology for Human Identification VII, Proceedings of SPIE Vol. 7667, 76670I (2010).
4. J.C. Wu, A.F. Martin, C.S. Greenberg, R.N. Kacker, Measurement uncertainties in speaker recognition evaluation, NISTIR 7722, National Institute of Standards and Technology, September, 2010.
5. J.C. Wu, C.L. Wilson, Nonparametric analysis of fingerprint data on large data sets, Pattern Recognition 40 (9) (2007) 2574-2584.
6. J.C. Wu, M.D. Garris, Nonparametric statistical data analysis of fingerprint minutiae exchange with two-finger fusion, in Biometric Technology for Human Identification IV, Proceedings of SPIE Vol. 6539, 65390N (2007).
7. J.C. Wu, A.F. Martin, R.N. Kacker, Validation of two-sample bootstrap in ROC analysis on large datasets using AURC, NISTIR 7733, National Institute of Standards and Technology, October, 2010.
8. B. Efron, Bootstrap methods: Another look at the Jackknife. Ann. Statistics, 7:1-26, 1979.

9. P. Hall, On the number of bootstrap simulations required to construct a confidence interval, *Ann. Statist.* 14 (4) (1986) 1453-1462.
10. B. Efron, Better bootstrap confidence intervals, *J. Amer. Statist. Assoc.* 82 (397) (1987) 171-185.
11. B. Efron, R.J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, New York, 1993.
12. J.C. Wu, C.L. Wilson, An empirical study of sample size in ROC-curve analysis of fingerprint data, in *Biometric Technology for Human Identification III*, Proceedings of SPIE Vol. 6202, 620207 (2006).
13. R.Y. Liu, K. Singh, Moving blocks jackknife and bootstrap capture weak dependence. *Exploring the limits of bootstrap*, ed. by LePage and Billard. John Wiley, New York, 1992.
14. R.M. Bolle, J.H. Connell, S. Pankanti, N.K. Ratha, A.W. Senior, *Guide to Biometrics*, Springer, New York, 2003 pp. 269-292.
15. B. Ostle, L.C. Malone, *Statistics in Research: Basic Concepts and Techniques for Research Workers*, fourth ed., Iowa State University Press, Ames, 1988.
16. R.J. Hyndman, Y. Fan, Sample quantiles in statistical packages, *American Statistician* 50 (1996) 361-365.
17. R. Cappelli, D. Maio, D. Maltoni, J.L. Wayman, A.K. Jain, Performance evaluation of fingerprint verification systems, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28 (1) 2006 3-18.

## Appendix 1 -- Tables

### 1 Scenario 1: TAR at a given FAR

#### 1.1 Algorithm 1

Num. of replications B		200	400	600	800	1000
CVSE	Min.	0.047524	0.034664	0.027754	0.023912	0.021570
	Max.	0.054346	0.039866	0.031685	0.026866	0.023686
	Range	0.006822	0.005202	0.003931	0.002954	0.002116
CVLB	Min.	0.000062	0.000044	0.000036	0.000030	0.000026
	Max.	0.000067	0.000047	0.000041	0.000037	0.000031
	Range	0.000005	0.000003	0.000005	0.000007	0.000005
CVUB	Min.	0.000054	0.000041	0.000032	0.000030	0.000026
	Max.	0.000062	0.000044	0.000036	0.000032	0.000030
	Range	0.000008	0.000003	0.000004	0.000002	0.000004

**Table 1 High-accuracy Algorithm 1's minimum, maximum, and range of 10 estimators of CVSEs, CVLBs, and CVUBs, as the number of iterations ran from 100 up to 1000 at intervals of 100 for each specified B. B ran from 200 up to 1000 at intervals of 200. The statistic of interest is TAR at a given FAR.**

Num. of replications B	1200	1400	1600	1800	2000
CVSE	0.021218	0.018613	0.017951	0.016331	0.016040
CVLB	0.000027	0.000024	0.000023	0.000023	0.000020
CVUB	0.000024	0.000023	0.000022	0.000020	0.000019

**Table 2 High-accuracy Algorithm 1's estimators of CVSEs, CVLBs, and CVUBs, while B ran from 1200 up to 2000 at intervals of 200 as the number of iterations was fixed at 500. The statistic of interest is TAR at a given FAR.**

#### 1.2 Algorithm 2

Num. of replications B		200	400	600	800	1000
CVSE	Min.	0.056895	0.037193	0.031792	0.026763	0.024033
	Max.	0.062609	0.043167	0.034696	0.030500	0.026695
	Range	0.005714	0.005974	0.002904	0.003737	0.002662
CVLB	Min.	0.000941	0.000677	0.000519	0.000473	0.000442
	Max.	0.001052	0.000734	0.000627	0.000526	0.000478
	Range	0.000111	0.000057	0.000108	0.000053	0.000036
CVUB	Min.	0.001068	0.000685	0.000637	0.000532	0.000488
	Max.	0.001171	0.000838	0.000738	0.000611	0.000544
	Range	0.000103	0.000153	0.000101	0.000079	0.000056

**Table 3 Low-accuracy Algorithm 2's minimum, maximum, and range of 10 estimators of CVSEs, CVLBs, and CVUBs, as the number of iterations ran from 100 up to 1000 at intervals of 100 for each specified B. B ran from 200 up to 1000 at intervals of 200. The statistic of interest is TAR at a given FAR.**

Num. of replications B	1200	1400	1600	1800	2000
CVSE	0.023673	0.022299	0.021272	0.018918	0.017705
CVLB	0.000457	0.000397	0.000354	0.000331	0.000318
CVUB	0.000445	0.000429	0.000420	0.000389	0.000389

**Table 4 Low-accuracy Algorithm 2's estimators of CVSEs, CVLBs, and CVUBs, while B ran from 1200 up to 2000 at intervals of 200 as the number of iterations was fixed at 500. The statistic of interest is TAR at a given FAR.**

## 2 Scenario 2: TAR at a given threshold value

### 2.1 Algorithm 1

Num. of replications B		200	400	600	800	1000
CVSE	Min.	0.045133	0.030526	0.025991	0.024365	0.020767
	Max.	0.051907	0.036352	0.031090	0.026430	0.023275
	Range	0.006774	0.005826	0.005099	0.002065	0.002508
CVLB	Min.	0.000055	0.000041	0.000035	0.000030	0.000027
	Max.	0.000068	0.000048	0.000039	0.000035	0.000030
	Range	0.000012	0.000006	0.000004	0.000005	0.000003
CVUB	Min.	0.000057	0.000040	0.000033	0.000028	0.000026
	Max.	0.000063	0.000044	0.000038	0.000033	0.000030
	Range	0.000006	0.000004	0.000004	0.000005	0.000004

**Table 5** High-accuracy Algorithm 1's minimum, maximum, and range of 10 estimators of CVSEs, CVLBs, and CVUBs, as the number of iterations ran from 100 up to 1000 at intervals of 100 for each specified B. B ran from 200 up to 1000 at intervals of 200. The statistic of interest is TAR at a given threshold value.

Num. of replications B	1200	1400	1600	1800	2000
CVSE	0.019498	0.018551	0.018105	0.016516	0.015611
CVLB	0.000026	0.000025	0.000022	0.000022	0.000021
CVUB	0.000025	0.000023	0.000023	0.000021	0.000019

**Table 6** High-accuracy Algorithm 1's estimators of CVSEs, CVLBs, and CVUBs, while B ran from 1200 up to 2000 at intervals of 200 as the number of iterations was fixed at 500. The statistic of interest is TAR at a given threshold value.

### 2.2 Algorithm 2

Num. of replications B		200	400	600	800	1000
CVSE	Min.	0.046884	0.033602	0.027518	0.023721	0.020796
	Max.	0.053526	0.036123	0.030318	0.025797	0.022964
	Range	0.006642	0.002521	0.002800	0.002076	0.002168
CVLB	Min.	0.000361	0.000259	0.000215	0.000185	0.000162
	Max.	0.000409	0.000277	0.000230	0.000205	0.000175
	Range	0.000048	0.000018	0.000015	0.000020	0.000013
CVUB	Min.	0.000350	0.000253	0.000206	0.000183	0.000159
	Max.	0.000397	0.000278	0.000229	0.000192	0.000182
	Range	0.000046	0.000025	0.000023	0.000009	0.000023

**Table 7** Low-accuracy Algorithm 2's minimum, maximum, and range of 10 estimators of CVSEs, CVLBs, and CVUBs, as the number of iterations ran from 100 up to 1000 at intervals of 100 for each specified B. B ran from 200 up to 1000 at intervals of 200. The statistic of interest is TAR at a given threshold value.

Num. of replications B	1200	1400	1600	1800	2000
CVSE	0.020317	0.018649	0.018001	0.016775	0.015417
CVLB	0.000155	0.000138	0.000142	0.000128	0.000119
CVUB	0.000150	0.000146	0.000132	0.000129	0.000119

**Table 8** Low-accuracy Algorithm 2's estimators of CVSEs, CVLBs, and CVUBs, while B ran from 1200 up to 2000 at intervals of 200 as the number of iterations was fixed at 500. The statistic of interest is TAR at a given threshold value.



### 3 Scenario 3: FAR at a given threshold value

#### 3.1 Algorithm 1

Num. of replications B		200	400	600	800	1000
CVSE	Min.	0.046920	0.033853	0.028190	0.023259	0.021197
	Max.	0.052033	0.036638	0.029539	0.026110	0.024945
	Range	0.005112	0.002786	0.001349	0.002850	0.003748
CVLB	Min.	0.018391	0.012871	0.010412	0.009149	0.008345
	Max.	0.020984	0.013912	0.011661	0.009910	0.009570
	Range	0.002593	0.001041	0.001250	0.000761	0.001225
CVUB	Min.	0.013832	0.009846	0.008105	0.007568	0.006658
	Max.	0.015573	0.011428	0.008911	0.008174	0.007806
	Range	0.001741	0.001582	0.000806	0.000606	0.001147

**Table 9 High-accuracy Algorithm 1's minimum, maximum, and range of 10 estimators of CVSEs, CVLBs, and CVUBs, as the number of iterations ran from 100 up to 1000 at intervals of 100 for each specified B. B ran from 200 up to 1000 at intervals of 200. The statistic of interest is FAR at a given threshold value.**

Num. of replications B	1200	1400	1600	1800	2000
CVSE	0.021625	0.019059	0.017675	0.016809	0.015373
CVLB	0.008368	0.007647	0.007375	0.007022	0.006433
CVUB	0.006691	0.005849	0.005742	0.005388	0.005315

**Table 10 High-accuracy Algorithm 1's estimators of CVSEs, CVLBs, and CVUBs, while B ran from 1200 up to 2000 at intervals of 200 as the number of iterations was fixed at 500. The statistic of interest is FAR at a given threshold value.**

#### 3.2 Algorithm 2

Num. of replications B		200	400	600	800	1000
CVSE	Min.	0.047950	0.034903	0.026481	0.022305	0.021546
	Max.	0.053789	0.039294	0.031482	0.026376	0.023059
	Range	0.005839	0.004391	0.005000	0.004071	0.001514
CVLB	Min.	0.017901	0.011890	0.010299	0.009629	0.008471
	Max.	0.019640	0.015084	0.012322	0.010173	0.009521
	Range	0.001739	0.003194	0.002023	0.000544	0.001050
CVUB	Min.	0.013287	0.010158	0.008025	0.007195	0.006535
	Max.	0.015618	0.011602	0.009204	0.008099	0.008224
	Range	0.002330	0.001444	0.001179	0.000904	0.001689

**Table 11 Low-accuracy Algorithm 2's minimum, maximum, and range of 10 estimators of CVSEs, CVLBs, and CVUBs, as the number of iterations ran from 100 up to 1000 at intervals of 100 for each specified B. B ran from 200 up to 1000 at intervals of 200. The statistic of interest is FAR at a given threshold value.**

Num. of replications B	1200	1400	1600	1800	2000
CVSE	0.020671	0.019016	0.018532	0.016877	0.015095
CVLB	0.008657	0.007888	0.007592	0.006696	0.006458
CVUB	0.006265	0.005850	0.005729	0.005452	0.005164

**Table 12 Low-accuracy Algorithm 2's estimators of CVSEs, CVLBs, and CVUBs, while B ran from 1200 up to 2000 at intervals of 200 as the number of iterations was fixed at 500. The statistic of interest is FAR at a given threshold value.**

## 4 Scenario 4: EER

### 4.1 Algorithm 1

Num. of replications B		200	400	600	800	1000
CVSE	Min.	0.046155	0.034597	0.027452	0.024199	0.022360
	Max.	0.050898	0.037824	0.031983	0.026339	0.023736
	Range	0.004743	0.003227	0.004532	0.002139	0.001376
CVLB	Min.	0.008676	0.006376	0.005198	0.004332	0.003818
	Max.	0.009634	0.006946	0.006061	0.004997	0.004364
	Range	0.000958	0.000570	0.000864	0.000665	0.000546
CVUB	Min.	0.008086	0.005970	0.004702	0.004117	0.003758
	Max.	0.008772	0.006529	0.005376	0.004747	0.004054
	Range	0.000686	0.000559	0.000674	0.000630	0.000297

**Table 13 High-accuracy Algorithm 1's minimum, maximum, and range of 10 estimators of CVSEs, CVLBs, and CVUBs, as the number of iterations ran from 100 up to 1000 at intervals of 100 for each specified B. B ran from 200 up to 1000 at intervals of 200. The statistic of interest is EER.**

Num. of replications B	1200	1400	1600	1800	2000
CVSE	0.020348	0.019152	0.017907	0.017155	0.015610
CVLB	0.003768	0.003800	0.003270	0.003240	0.003033
CVUB	0.003520	0.003299	0.003122	0.002907	0.002673

**Table 14 High-accuracy Algorithm 1's estimators of CVSEs, CVLBs, and CVUBs, while B ran from 1200 up to 2000 at intervals of 200 as the number of iterations was fixed at 500. The statistic of interest is EER.**

### 4.2 Algorithm 2

Num. of replications B		200	400	600	800	1000
CVSE	Min.	0.047634	0.034402	0.027218	0.021963	0.017186
	Max.	0.051995	0.036459	0.029568	0.027059	0.023337
	Range	0.004361	0.002056	0.002350	0.005095	0.006150
CVLB	Min.	0.001918	0.001334	0.001161	0.000972	0.000929
	Max.	0.002091	0.001548	0.001254	0.001089	0.000998
	Range	0.000173	0.000213	0.000093	0.000117	0.000069
CVUB	Min.	0.001954	0.001386	0.001101	0.000920	0.000799
	Max.	0.002206	0.001522	0.001240	0.001046	0.000949
	Range	0.000253	0.000136	0.000139	0.000126	0.000150

**Table 15 Low-accuracy Algorithm 2's minimum, maximum, and range of 10 estimators of CVSEs, CVLBs, and CVUBs, as the number of iterations ran from 100 up to 1000 at intervals of 100 for each specified B. B ran from 200 up to 1000 at intervals of 200. The statistic of interest is EER.**

Num. of replications B	1200	1400	1600	1800	2000
CVSE	0.020697	0.018524	0.016898	0.016887	0.016330
CVLB	0.000875	0.000789	0.000768	0.000676	0.000649
CVUB	0.000828	0.000796	0.000678	0.000675	0.000637

**Table 16 Low-accuracy Algorithm 2's estimators of CVSEs, CVLBs, and CVUBs, while B ran from 1200 up to 2000 at intervals of 200 as the number of iterations was fixed at 500. The statistic of interest is EER.**

## 5 Scenario 5: AURC

### 5.1 Algorithm 1

Num. of replications B		200	400	600	800	1000
CVSE	Min.	0.047289	0.033978	0.027837	0.023290	0.021658
	Max.	0.051731	0.036868	0.029854	0.026749	0.025690
	Range	0.004442	0.002890	0.002017	0.003459	0.004032
CVLB	Min.	0.000023	0.000018	0.000013	0.000012	0.000011
	Max.	0.000027	0.000021	0.000016	0.000014	0.000013
	Range	0.000004	0.000003	0.000003	0.000002	0.000001
CVUB	Min.	0.000022	0.000015	0.000013	0.000011	0.000010
	Max.	0.000025	0.000016	0.000014	0.000013	0.000011
	Range	0.000003	0.000002	0.000002	0.000001	0.000001

**Table 17 High-accuracy Algorithm 1's minimum, maximum, and range of 10 estimators of CVSEs, CVLBs, and CVUBs, as the number of iterations ran from 100 up to 1000 at intervals of 100 for each specified B. B ran from 200 up to 1000 at intervals of 200. The statistic of interest is AURC.**

Num. of replications B	1200	1400	1600	1800	2000
CVSE	0.020065	0.019558	0.017755	0.015916	0.015397
CVLB	0.000011	0.000011	0.000010	0.000008	0.000008
CVUB	0.000009	0.000009	0.000008	0.000008	0.000007

**Table 18 High-accuracy Algorithm 1's estimators of CVSEs, CVLBs, and CVUBs, while B ran from 1200 up to 2000 at intervals of 200 as the number of iterations was fixed at 500. The statistic of interest is AURC.**

### 5.2 Algorithm 2

Num. of replications B		200	400	600	800	1000
CVSE	Min.	0.048170	0.034661	0.028056	0.023298	0.021355
	Max.	0.052180	0.036136	0.030047	0.027780	0.024504
	Range	0.004010	0.001475	0.001991	0.004482	0.003148
CVLB	Min.	0.000089	0.000063	0.000051	0.000044	0.000041
	Max.	0.000100	0.000068	0.000055	0.000049	0.000044
	Range	0.000011	0.000005	0.000004	0.000005	0.000003
CVUB	Min.	0.000085	0.000060	0.000051	0.000044	0.000039
	Max.	0.000092	0.000069	0.000054	0.000048	0.000043
	Range	0.000007	0.000008	0.000003	0.000004	0.000004

**Table 19 Low-accuracy Algorithm 2's minimum, maximum, and range of 10 estimators of CVSEs, CVLBs, and CVUBs, as the number of iterations ran from 100 up to 1000 at intervals of 100 for each specified B. B ran from 200 up to 1000 at intervals of 200. The statistic of interest is AURC.**

Num. of replications B	1200	1400	1600	1800	2000
CVSE	0.020828	0.017814	0.017853	0.016148	0.015287
CVLB	0.000038	0.000036	0.000033	0.000030	0.000030
CVUB	0.000039	0.000035	0.000033	0.000029	0.000029

**Table 20 Low-accuracy Algorithm 2's estimators of CVSEs, CVLBs, and CVUBs, while B ran from 1200 up to 2000 at intervals of 200 as the number of iterations was fixed at 500. The statistic of interest is AURC.**

## Appendix 2 -- Figures

### 1 CVSE of Algorithms 1 and 2

#### 1.1 Scenario 1: TAR at a given FAR

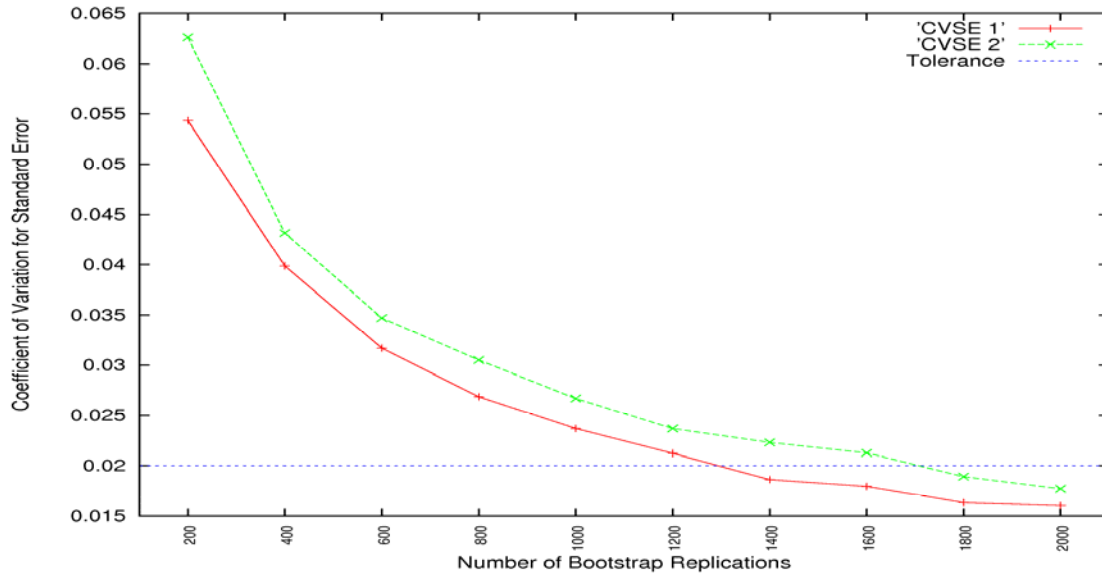


Figure 2 The estimators of CVSEs for high-accuracy Algorithm 1 and low-accuracy Algorithm 2 as a function of the number of bootstrap replications. The tolerance is set to be 0.02. The statistic of interest is TAR at a given FAR.

#### 1.2 Scenario 2: TAR at a given threshold value

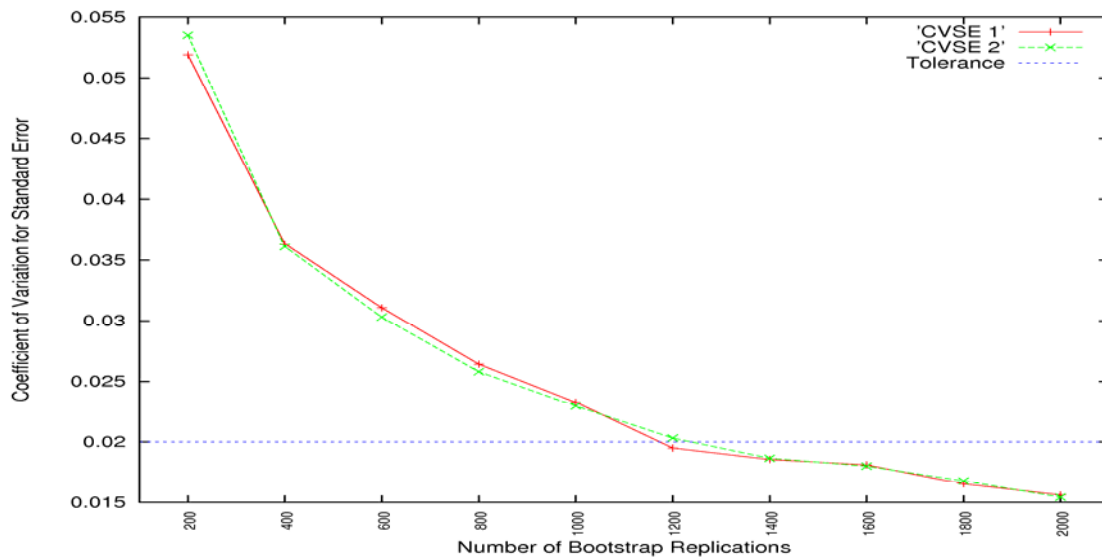


Figure 3 The estimators of CVSEs for high-accuracy Algorithm 1 and low-accuracy Algorithm 2 as a function of the number of bootstrap replications. The tolerance is set to be 0.02. The statistic of interest is TAR at a given threshold value.

### 1.3 Scenario 3: FAR at a given threshold value

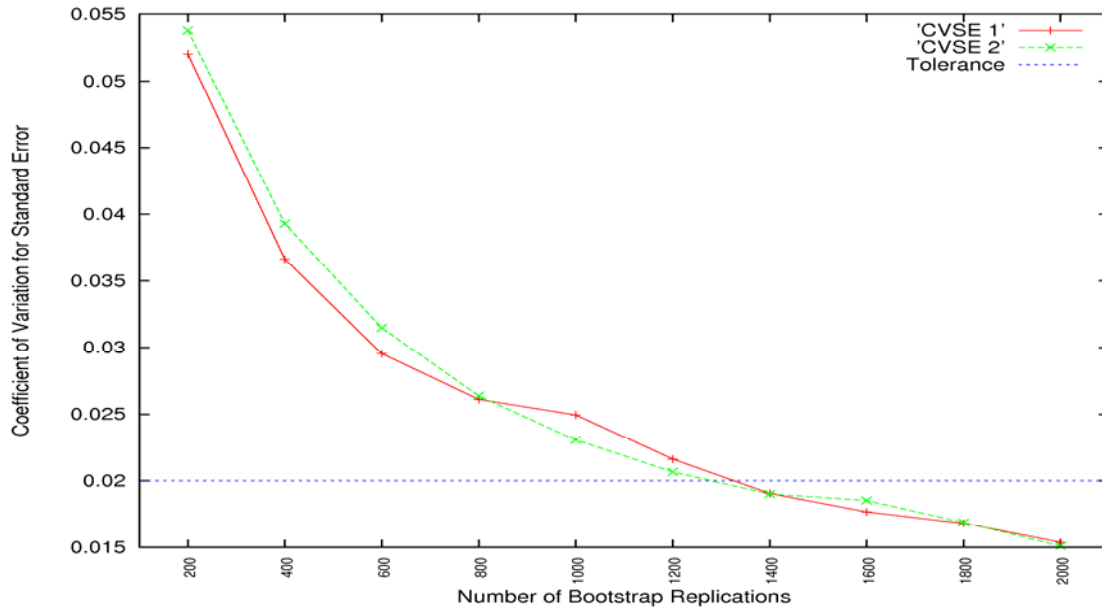


Figure 4 The estimators of CVSEs for high-accuracy Algorithm 1 and low-accuracy Algorithm 2 as a function of the number of bootstrap replications. The tolerance is set to be 0.02. The statistic of interest is FAR at a given threshold value.

### 1.4 Scenario 4: EER

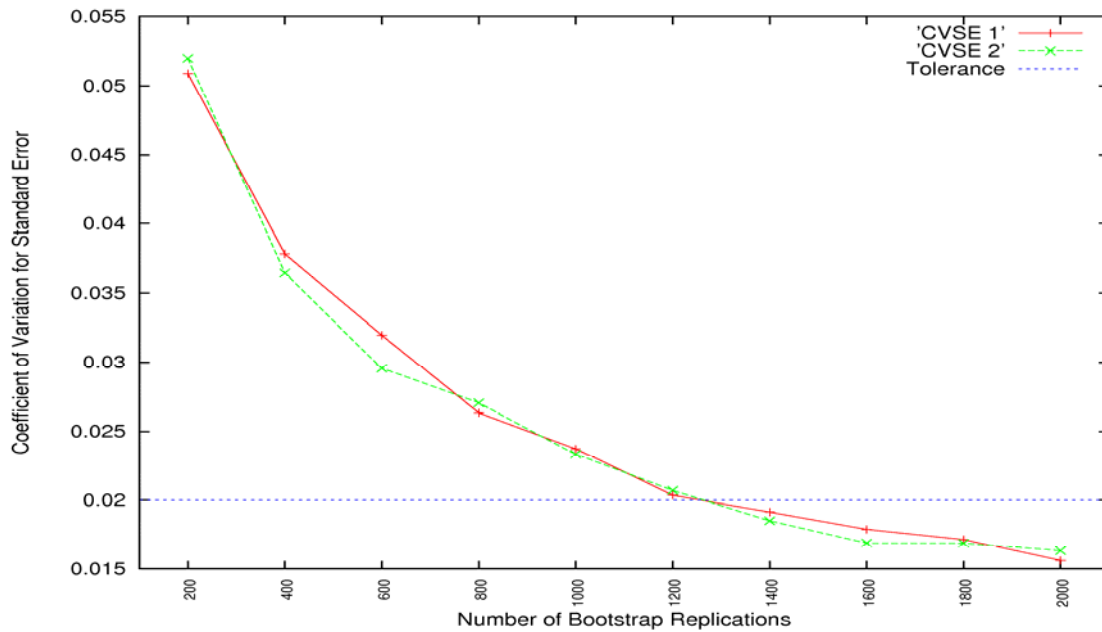


Figure 5 The estimators of CVSEs for high-accuracy Algorithm 1 and low-accuracy Algorithm 2 as a function of the number of bootstrap replications. The tolerance is set to be 0.02. The statistic of interest is EER.

## 1.5 Scenario 5: AURC

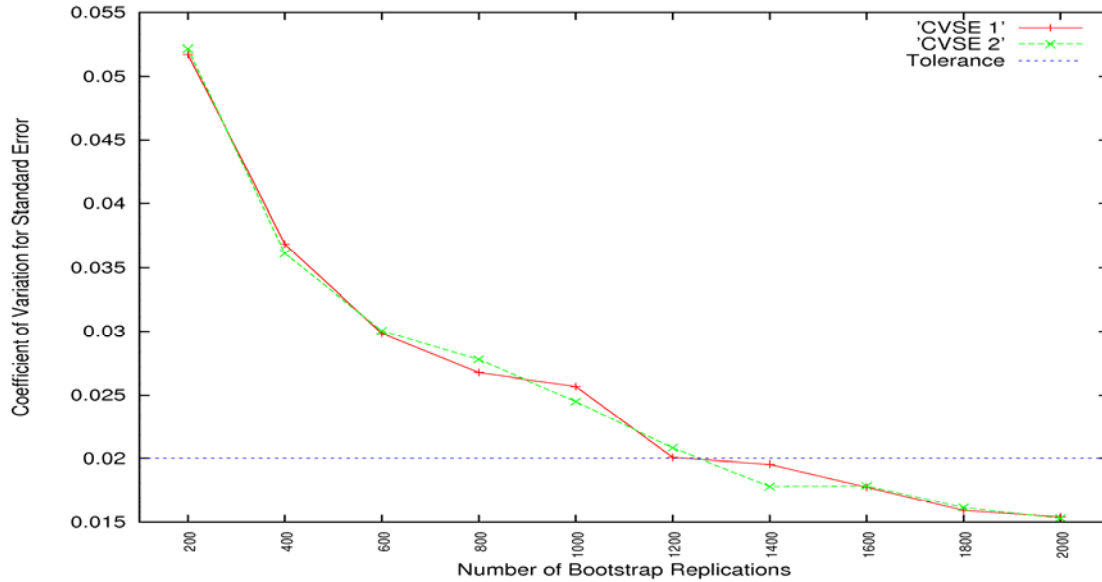


Figure 6 The estimators of CVSEs for high-accuracy Algorithm 1 and low-accuracy Algorithm 2 as a function of the number of bootstrap replications. The tolerance is set to be 0.02. The statistic of interest is AURC.

## 2 CVLB and CVUB of Algorithms 1 and 2

### 2.1 Scenario 1: TAR at a given FAR

#### 2.1.1 Algorithm 1

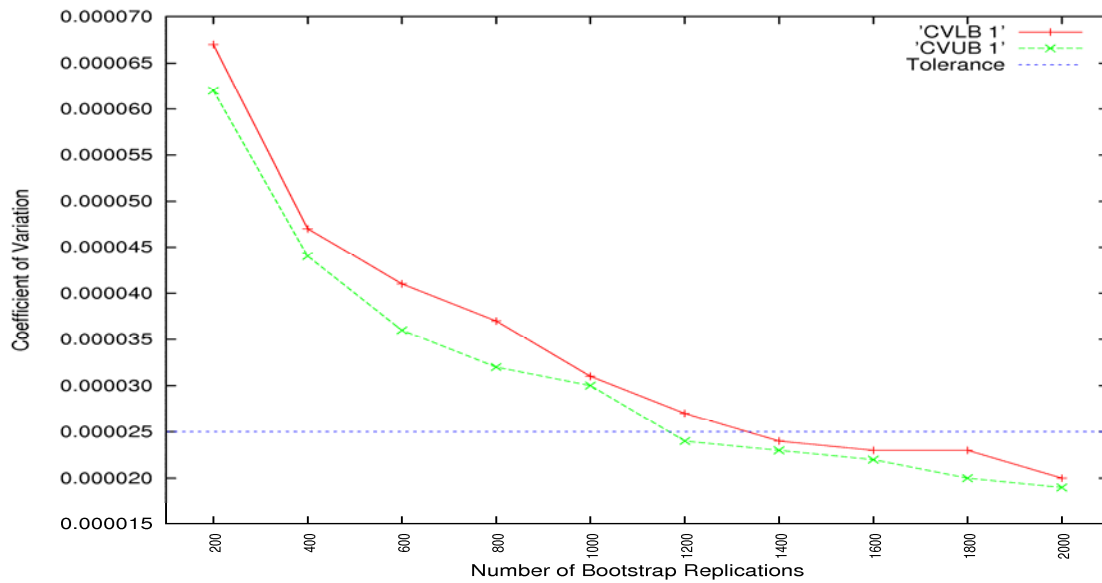


Figure 7 The estimators of CVLBs and CVUBs for high-accuracy Algorithm 1 as a function of the number of bootstrap replications. The tolerance is set to be 0.000025. The statistic of interest is TAR at a given FAR.

## 2.1.2 Algorithm 2

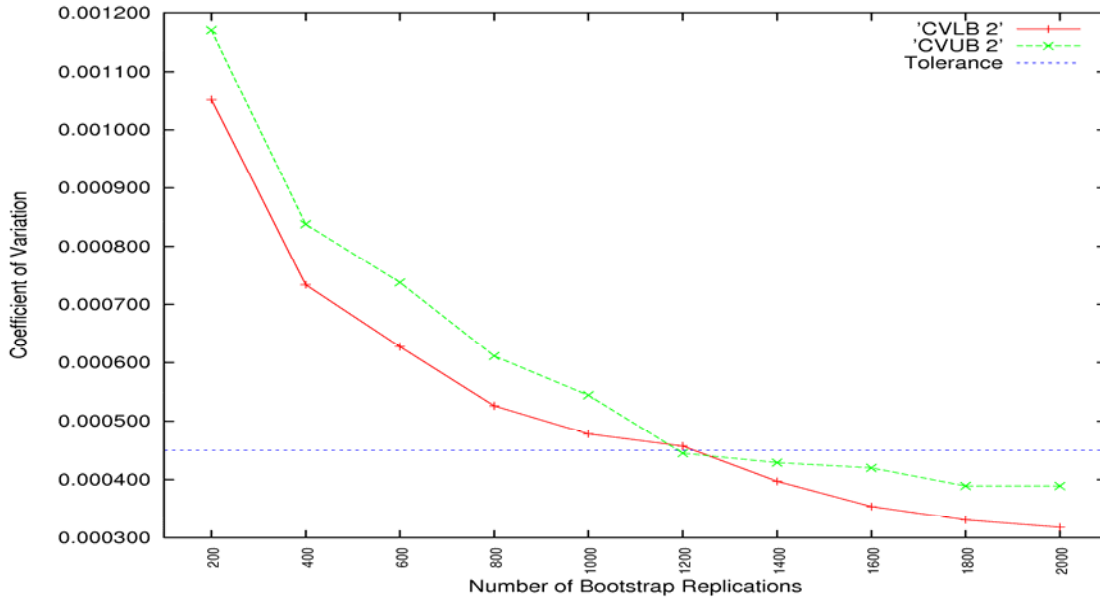


Figure 8 The estimators of CVLBs and CVUBs for low-accuracy Algorithm 2 as a function of the number of bootstrap replications. The tolerance is set to be 0.000450. The statistic of interest is TAR at a given FAR.

## 2.2 Scenario 2: TAR at a given threshold value

### 2.2.1 Algorithm 1

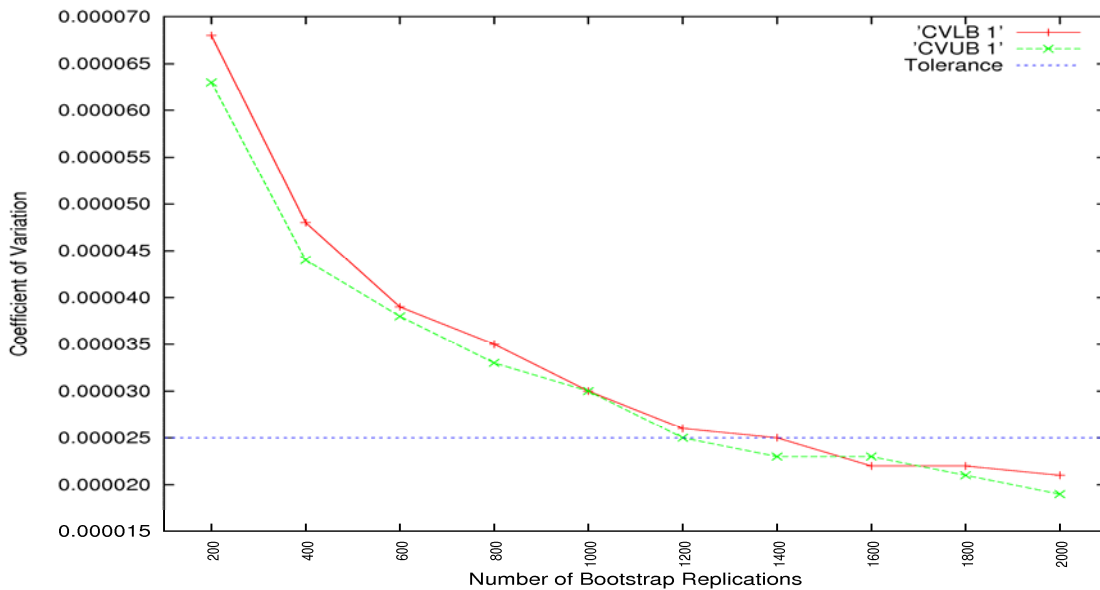


Figure 9 The estimators of CVLBs and CVUBs for high-accuracy Algorithm 1 as a function of the number of bootstrap replications. The tolerance is set to be 0.000025. The statistic of interest is TAR at a given threshold value.

## 2.2.2 Algorithm 2

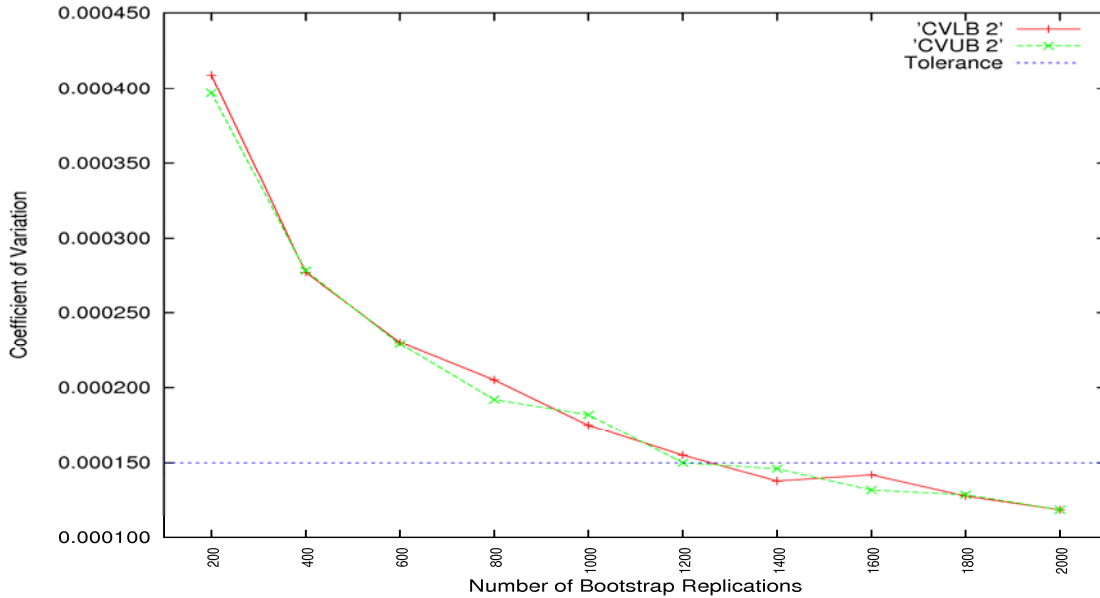


Figure 10 The estimators of CVLBs and CVUBs for low-accuracy Algorithm 2 as a function of the number of bootstrap replications. The tolerance is set to be 0.000150. The statistic of interest is TAR at a given threshold value.

## 2.3 Scenario 3: FAR at a given threshold value

### 2.3.1 Algorithm 1

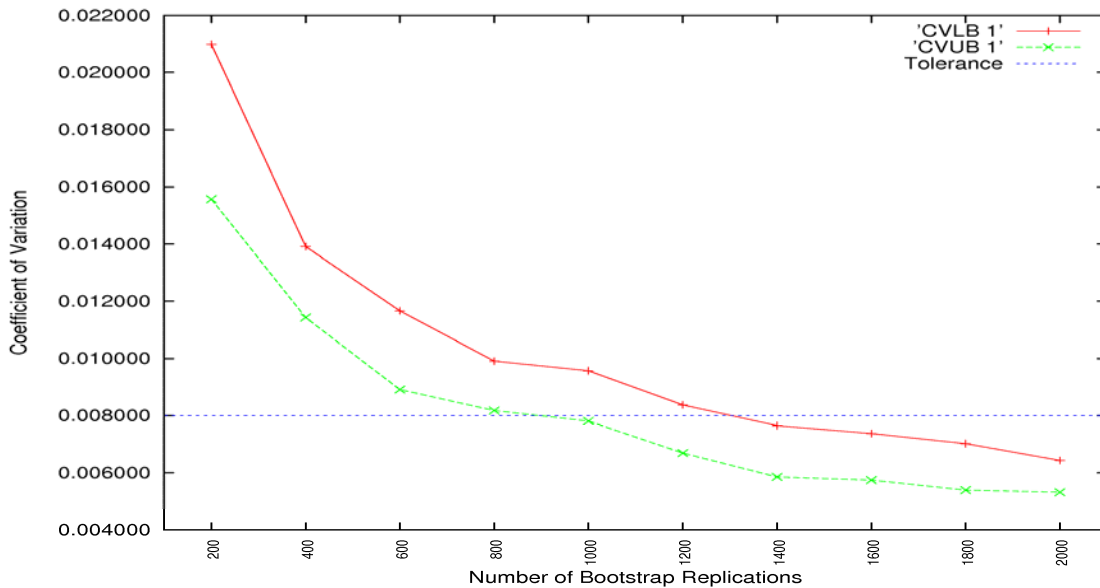


Figure 11 The estimators of CVLBs and CVUBs for high-accuracy Algorithm 1 as a function of the number of bootstrap replications. The tolerance is set to be 0.008. The statistic of interest is FAR at a given threshold value.



### 2.3.2 Algorithm 2

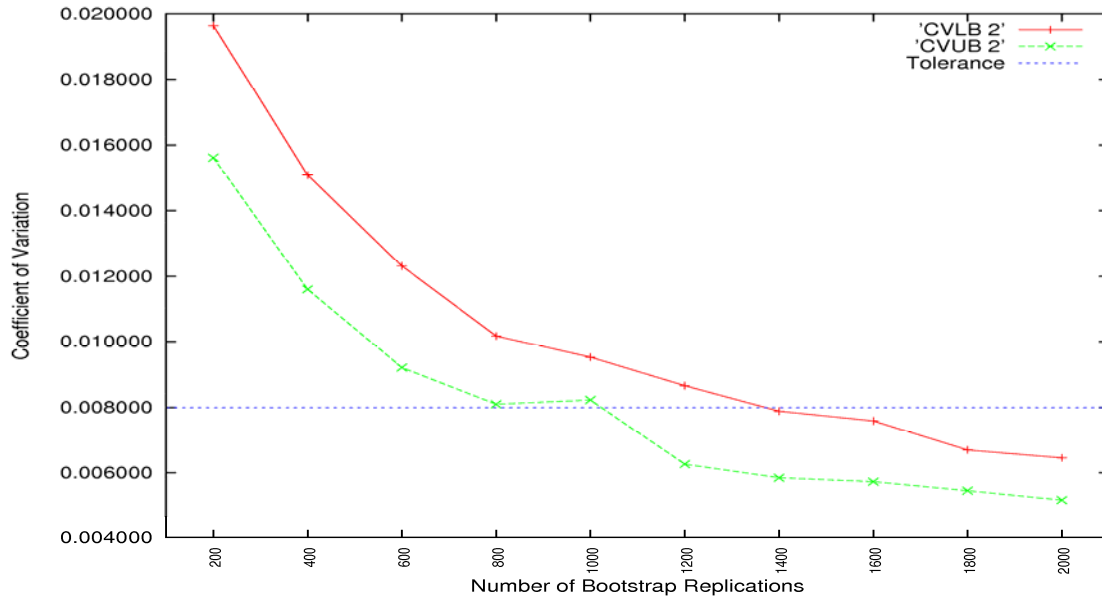


Figure 12 The estimators of CVLBs and CVUBs for low-accuracy Algorithm 2 as a function of the number of bootstrap replications. The tolerance is set to be 0.008. The statistic of interest is FAR at a given threshold value.

### 2.4 Scenario 4: EER

#### 2.4.1 Algorithm 1

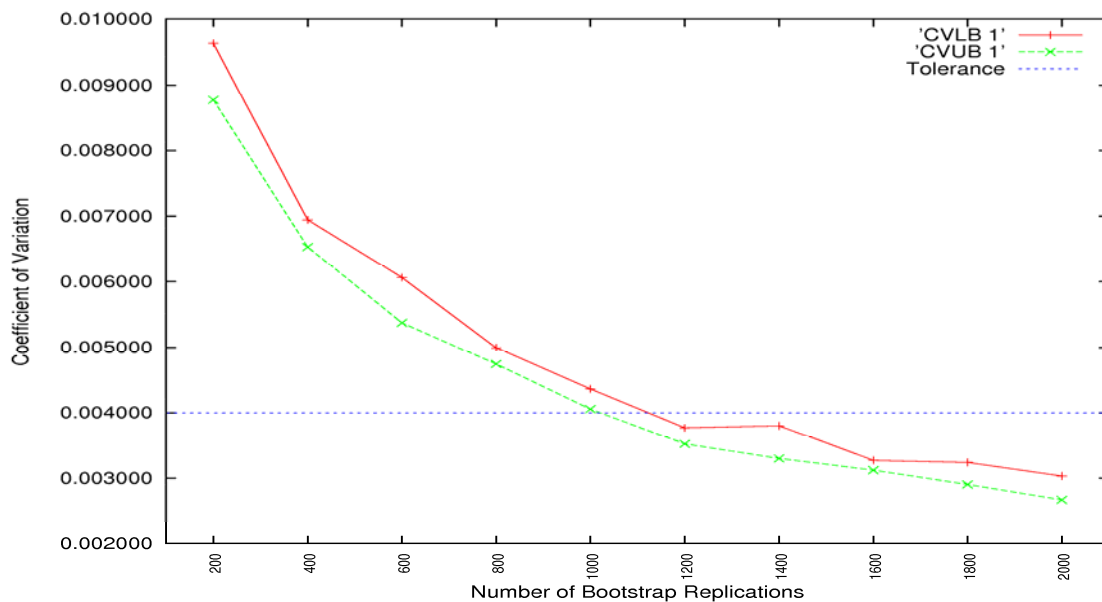


Figure 13 The estimators of CVLBs and CVUBs for high-accuracy Algorithm 1 as a function of the number of bootstrap replications. The tolerance is set to be 0.004. The statistic of interest is EER.

## 2.4.2 Algorithm 2

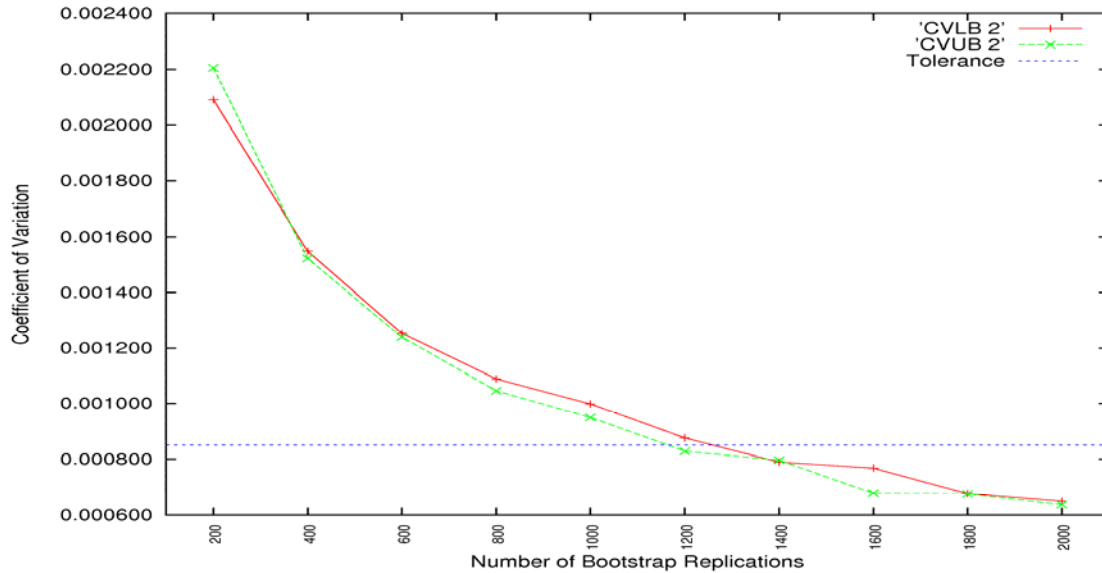


Figure 14 The estimators of CVLBs and CVUBs for low-accuracy Algorithm 2 as a function of the number of bootstrap replications. The tolerance is set to be 0.000850. The statistic of interest is EER.

## 2.5 Scenario 5: AURC

### 2.5.1 Algorithm 1

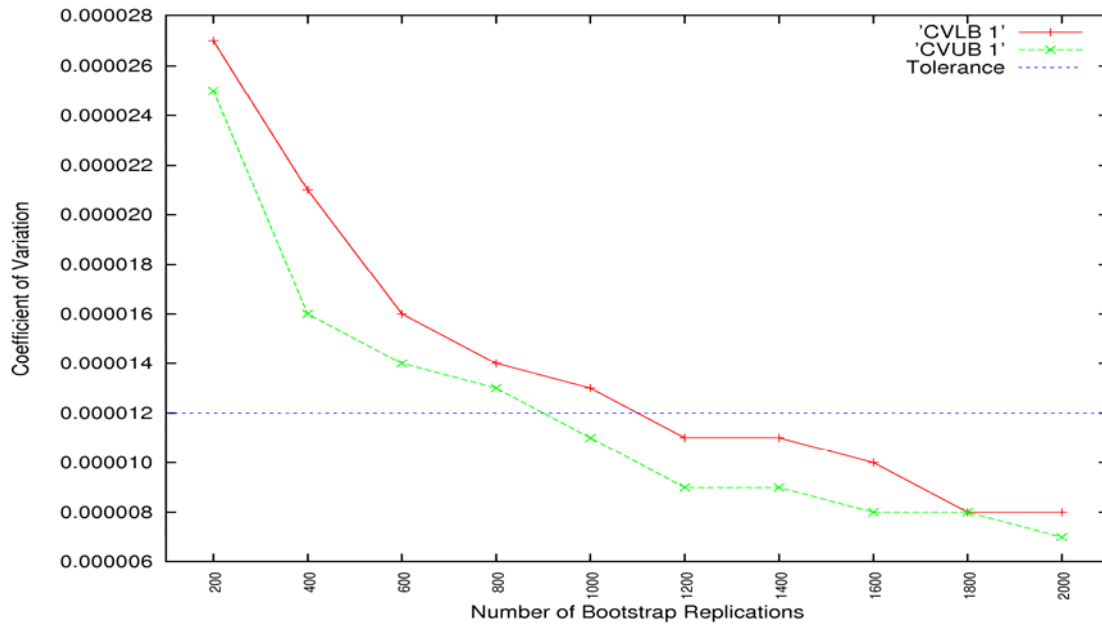
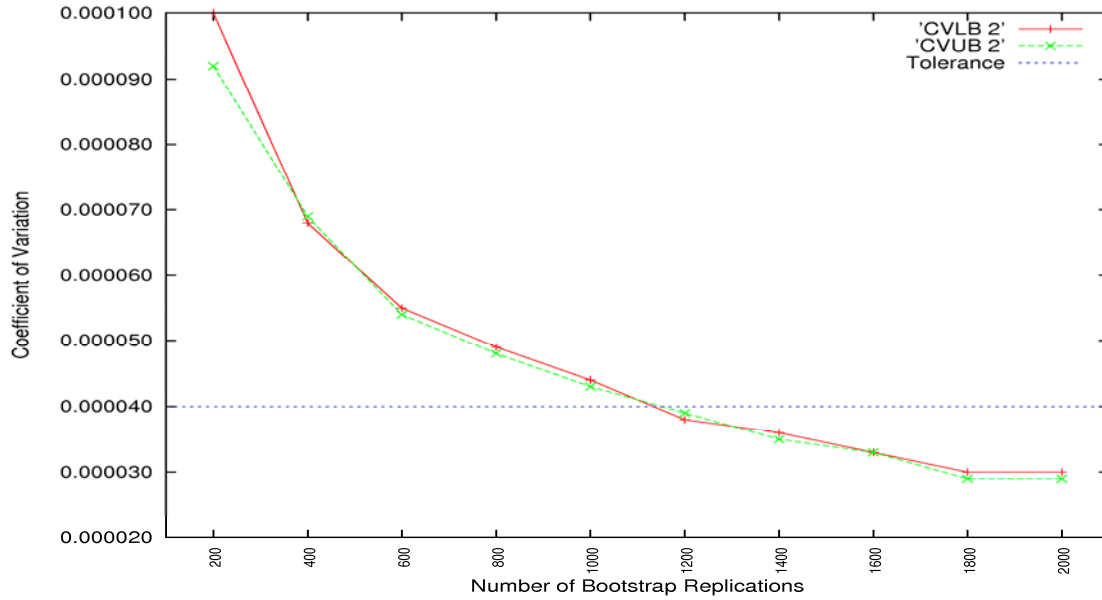


Figure 15 The estimators of CVLBs and CVUBs for high-accuracy Algorithm 1 as a function of the number of bootstrap replications. The tolerance is set to be 0.000012. The statistic of interest is AURC.

## 2.5.2 Algorithm 2



**Figure 16** The estimators of CVLBs and CVUBs for low-accuracy Algorithm 2 as a function of the number of bootstrap replications. The tolerance is set to be 0.000040. The statistic of interest is AURC.