Communications in Statistics - Simulation and Computation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lssp20

Bootstrap Variability Studies in ROC Analysis on Large Datasets

Jin Chu Wu [a] , Alvin F. Martin [a] & Raghu N. Kacker [a]

[a] Information Technology Laboratory, National Institute of Standards and Technology , Gaithersburg , MD , USA

Accepted author version posted online: 16 May 2013.Published online: 13 Aug 2013.

To cite this article: Jin Chu Wu , Alvin F. Martin & Raghu N. Kacker (2014) Bootstrap Variability Studies in ROC Analysis on Large Datasets, Communications in Statistics - Simulation and Computation, 43:1, 225-236, DOI: 10.1080/03610918.2012.700362

To link to this article: http://dx.doi.org/10.1080/03610918.2012.700362

Communications in Statistics—Simulation and Computation®, 43: 225–236, 2014
ISSN: 0361-0918 print / 1532-4141 online
DOI: 10.1080/03610918.2012.700362

Taylor & Francis
Taylor & Francis Group

Bootstrap Variability Studies in ROC Analysis on Large Datasets

JIN CHU WU, ALVIN F. MARTIN, AND RAGHU N. KACKER

Information Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA

The nonparametric two-sample bootstrap is employed to compute uncertainties of measures in receiver operating characteristic (ROC) analysis on large datasets in areas such as biometrics, and so on. In this framework, the bootstrap variability was empirically studied without a normality assumption, exhaustively in five scenarios involving both high- and low-accuracy matching algorithms. With a tolerance 0.02 of the coefficient of variation, it was found that 2000 bootstrap replications were appropriate for ROC analysis on large datasets in order to reduce the bootstrap variance and ensure the accuracy of the computation.

Keywords Bootstrap variability; Bootstrap replications; ROC analysis; Large datasets; Uncertainty; Biometrics.

Mathematical Subject Classification

1. Introduction

The receiver operating characteristic (ROC) analysis is an important statistical technique in a wide variety of disciplines. Without loss of generality, in this article, biometric applications on large datasets will be taken as examples. 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. The two distributions of continuous scores are schematically depicted in Figure 1 (a). The cumulative probabilities of genuine and impostor scores from the highest score to a threshold are defined as the true accept rate (TAR) and the false accept rate (FAR), respectively. Thus, in the FAR-TAR coordinate system, as the threshold moves from the highest score down to the lowest score, an ROC curve is constructed as drawn in Figure 1 (b).

As extensively investigated (Wu et al., 2011; Wu and Wilson, 2007), genuine and impostor scores have no underlying parametric distribution functions; 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 matching accuracy. This suggests that parametric data modeling, including a normality assumption, is not appropriate for evaluating matching algorithms on large

Received September 19, 2011; Accepted May 30, 2012
Address correspondence to Dr. Jin Chu Wu. Ph.D., Information Technology Laboratory, National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899 USA; E-mail: jinchu.wu@nist.gov

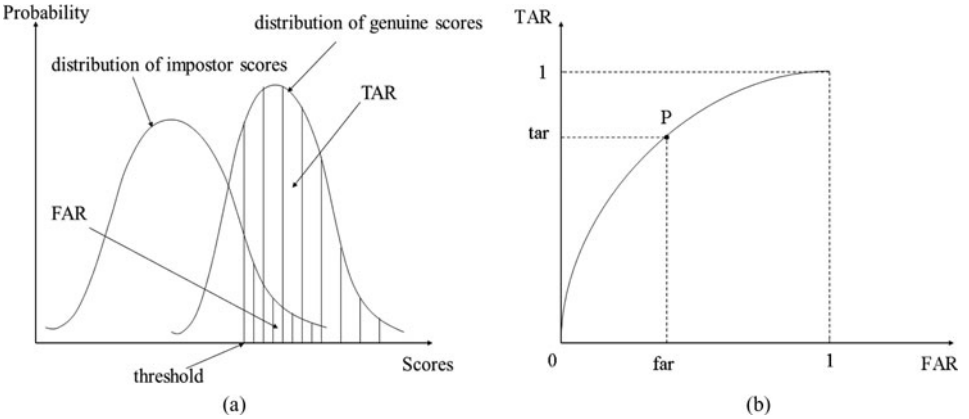

Figure 1. (a): A schematic diagram of two distributions of continuous genuine scores and impostor scores, showing three related variables: TAR, FAR, and threshold. (b): A schematic drawing of an ROC curve.

datasets. Moreover, the above two distributions are interrelated by the algorithm that generates them, and all statistics of interest are influenced by the combined impact of these two distributions.

The nonparametric two-sample bootstrap method is employed to compute the uncertainties of measures in ROC analysis on large datasets, for instance the standard error (SE) of the TAR at a given FAR, the SE of the equal error rate (EER), and so on (Wu et al., 2011). In these cases, it is hard to compute the measurement uncertainties without using the bootstrap method. In this framework, one challenging issue is: How many bootstrap replications are needed to reduce the bootstrap variance and ensure the accuracy of computation?

The number of bootstrap replications is intrinsically related to bootstrap variability. As investigated in the literature (Efron, 1979; Efron, 1987; Efron and Tibshirani, 1993; Hall, 1986), 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; the latter is because the number of bootstrap replications is not infinite. In the meantime, the bootstrap variance produces variances of the SE and the two bounds of the confidence interval (CI) of the distribution that is formed by the bootstrap replications of the statistics. As a result, these variances are functions of the sample size as well as of the number of bootstrap replications. Inversely, the sample size and the number of bootstrap replications can be determined from these variances.

The data used in this article are all operational data, i.e., real-life data. Regarding the sample sizes, the total number of genuine scores is about 60,000 and the total number of impostor scores is about 120,000. They are fixed based on our previous study, which was carried out using Chebyshev's inequality (Wu and Wilson, 2006). It was found that if the numbers of scores increased from what were used here, the measurement accuracy would improve little. Therefore, in our applications, only the number of bootstrap replications needs to be determined.

The bootstrap variability was rigorously studied previously (Efron, 1987; Efron and Tibshirani, 1993; Hall, 1986). In these analytical studies, the normality assumption was made, and the statistics of interest was a sample mean. However, in our applications, as stated above, it is inappropriate to assume normality for score distributions, and the statistics

of interest are probabilities such as TAR, FAR, EER, etc., which are all related to ROC analysis. Under such circumstances, it is hard to derive formulas to conduct bootstrap variability studies analytically. Moreover, the sizes of the datasets in our cases are much larger than those in other applications, such as medical ones.

Hence, in this article, the bootstrap variability is empirically studied without a normality assumption, exhaustively in five scenarios for ROC analysis on large datasets involving both high- and low-accuracy matching algorithms[1]. Any point P on an ROC curve has two coordinates, FAR and TAR, and is associated with a threshold for the distributions of genuine and impostor scores, as illustrated in Figure 1. Thus there are only three variables, FAR, TAR, and threshold, and any one of them determines the other two. In practice, it is seldom required that TAR be specified in advance. Thus, the first three scenarios are measuring the TAR at a specified FAR, measuring the TAR at a given threshold value, and measuring the FAR at a given threshold value.

The fourth scenario is measuring the EER that is defined as the FAR at a threshold where 1 – TAR and FAR are equal. These two error rates are traded-off against each other. The smaller the EER is, the more accurate the matching algorithm is. The fifth scenario is measuring the area under ROC curve (AURC), which is equivalent to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Fawcett, 2006).

All related formulas for computing statistics of interest in the five scenarios can be found in Refs. Wu et al. (2011) and Wu and Wilson (2007). In ROC analysis, there may be other measures of interest, but they are generally variations or combinations of the measures discussed in the first three scenarios (see Section 4). These are very popular statistics in ROC analysis on large datasets. Our bootstrap variability studies took weeks of CPU time due to large-scale computations involving large datasets.

The discrete distribution functions of genuine scores and impostor scores and an algorithm for the nonparametric two-sample bootstrap are shown in Section 2. The bootstrap variability study is presented in Section 3, which determines the number of bootstrap replications for ROC analysis on large datasets. Finally, conclusions and discussion can be found in Section 4.

2. The Discrete Distribution Functions and an Algorithm of Nonparametric Two-Sample Bootstrap

2.1. *The Discrete Distribution Functions of Scores*

All scores were converted into integers if they were not. Without loss of generality, the scores are expressed inclusively using the integer score set $\{s\} = \{s_{\min}, s_{\min}+1, \ldots, s_{\max}\}$, running consecutively from the lowest score $s_{\min}$ to the highest score $s_{\max}$. The genuine score set and the impostor score set in the sense of multiset are denoted as

$$\mathbf{G} = \{m_i | m_i \in \{s\} \quad \text{and} \quad i = 1, \ldots, N_G\}, \tag{1}$$

[1]Specific hardware and software products identified in this paper 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.

and

$$\mathbf{I} = \{n_i | n_i \in \{s\} \quad \text{and} \quad i = 1, \ldots, N_I\}, \tag{2}$$

where N_G and N_I are the total numbers of genuine scores and impostor scores, respectively. These two sets, **G** and **I**, constitute 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 (Efron, 1979; Efron and Tibshirani, 1993) is employed to compute the estimates of measurement uncertainties for ROC analysis in all five scenarios. The algorithm is as follows.

***Algorithm I* (Nonparametric two-sample bootstrap)**

1: **for** $i = 1$ **to** B **do**
2: select N_G scores randomly WR from **G** to form a set $\{\text{new } N_G \text{ genuine scores}\}_i$
3: select N_I scores randomly WR from **I** to form a set $\{\text{new } N_I \text{ impostor scores}\}_i$
4: $\{\text{new } N_G \text{ genuine scores}\}_i$ and $\{\text{new } N_I \text{ impostor scores}\}_i$ => statistic $\hat{T}_\text{i}$
5: **end for**
6: $\{\hat{T}_\text{i} | i = 1, \ldots, B\} \Rightarrow \hat{\text{SE}}_B$ 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 **G** in Equation (1) and impostor score set **I** in Equation (2) are generated by a matching algorithm. As shown from Step 1 to 5, this algorithm runs B times. In the ith iteration, N_G scores are randomly selected WR from the original genuine score set **G** to form a new set of N_G genuine scores, N_I scores are randomly selected WR from the original impostor score set **I** to form a new set of N_I impostor scores, and then from these two new sets of scores the ith bootstrap replication of the estimated statistic of interest, $\hat{T}_\text{i}$, is generated.

The $\hat{T}_\text{i}$ are different in the five scenarios discussed in Section 1. In Scenario 1, $\hat{T}_\text{i} = \hat{\text{TAR}}_i(f)$ at a specified $f = \text{FAR}$. In Scenario 2, $\hat{T}_\text{i} = \hat{\text{TAR}}_i(t)$ at a given threshold t. In Scenario 3, $\hat{T}_\text{i} = \hat{\text{FAR}}_i(t)$ at a given threshold t. In Scenario 4, $\hat{T}_\text{i} = \hat{\text{EER}}_i$. And in Scenario 5, $\hat{T}_\text{i} = \hat{\text{AURC}}_i$. The formulas for computing all these five statistics of interest can be found in Refs. Wu et al. (2011) and Wu and Wilson (2007).

Finally, as indicated in Step 6, from the set $\{\hat{T}_\text{i} | i = 1, \ldots, B\}$, the estimator of the SE, denoted by $\hat{\text{SE}}_B$, 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 distributions, denoted by $\hat{Q}_B(\alpha/2)$ and $\hat{Q}_B(1-\alpha/2)$, at the significance level α can be calculated (Efron and Tibshirani, 1993). Definition 2 of quantile in Ref. Hyndman and Fan (1996) 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% $\hat{\text{CI}}$. If 95% $\hat{\text{CI}}$ is of interest, then α is set to be 0.05.

3. The Bootstrap Variability Study and the Number of Bootstrap Replications

3.1. *An Algorithm for Empirical Study 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 the bootstrap distribution of the statistic of interest in ROC analysis on large datasets. To take into account the impact of the mean value, the coefficient of variation (CV) rather than just variance is employed (Efron and Tibshirani, 1993). The empirical study of bootstrap variability will be carried out in all five scenarios. Here is an algorithm for bootstrap variability study.

***Algorithm II* (Bootstrap variability study)**

1: **for** i = 1 **to** L **do**
2: **for** j = 1 **to** B **do**
3: select N_G scores randomly WR from **G** to form a set $\{\text{new } N_G \text{ genuine scores}\}_{ij}$
4: select N_I scores randomly WR from **I** to form a set $\{\text{new } N_I \text{ impostor scores}\}_{ij}$
5: $\{\text{new } N_G \text{ genuine scores}\}_{ij}$ and $\{\text{new } N_I \text{ impostor scores}\}_{ij}$ => statistic $\hat{T}_{ij}$
6: **end for**
7: $\{\hat{T}_{ij} | j = 1, \ldots, B\} \Rightarrow \hat{\text{SE}}_{Bi}$ and $(\hat{Q}_{Bi}(\alpha/2), \hat{Q}_{Bi}(1-\alpha/2))$
8: **end for**
9: $\{\hat{\text{SE}}_{Bi}, \hat{Q}_{Bi}(\alpha/2), \hat{Q}_{Bi}(1-\alpha/2) | i = 1, \ldots, \text{L}\} \Rightarrow \hat{\text{CV}}_{\text{B,L}(\kappa)}, \kappa = \mathbf{SE_{B,L}}, \mathbf{Q_{B,L}}(\boldsymbol{\alpha/2}), \mathbf{Q_{B,L}}(\mathbf{1}-\boldsymbol{\alpha/2})$
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 $\hat{\text{SE}}_{Bi}$, $\hat{Q}_{Bi}(\alpha/2)$ and $\hat{Q}_{Bi}(1-\alpha/2)$ of a statistic of interest in the i th iteration for a given B. The statistics of interest in five scenarios are specified 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,

$$\begin{aligned} \mathbf{SE_{B,L}} &= \{\hat{\text{SE}}_{Bi} | i = 1, \ldots, L\}, \\ \mathbf{Q_{B,L}}(\boldsymbol{\alpha/2}) &= \{\hat{Q}_{Bi}(\alpha/2) | i = 1, \ldots, L\}, \\ \mathbf{Q_{B,L}}(\mathbf{1}-\boldsymbol{\alpha/2}) &= \{\hat{Q}_{Bi}(1-\alpha/2) | i = 1, \ldots, L\}. \end{aligned} \tag{3}$$

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

$$\hat{\text{CV}}_{B,L}(\kappa) = \frac{\sqrt{\hat{\text{VAR}}_{B,L}(\kappa)}}{\hat{E}_{B,L}(\kappa)}, \quad \textit{where } \kappa = \mathbf{SE_{B,L}}, \mathbf{Q_{B,L}}(\boldsymbol{\alpha/2}), \mathbf{Q_{B,L}}(\mathbf{1}-\boldsymbol{\alpha/2}). \tag{4}$$

It is clear that the three estimated $\hat{\text{CV}}$s are functions of the number of bootstrap replications B and the number of Monte Carlo iterations L, as well as the significance level α. Therefore, the number of bootstrap replications B can be determined by the tolerable

Table 1
High-accuracy Algorithm A's minimum, maximum, and range of 10 estimators of CVSEs, CVLBs, and CVUBs, as the number of iterations ran from 100 up to 1,000 at intervals of 100 for each specified B. B ran from 200 up to 1,000 at intervals of 200. The statistic of interest is TAR at a given FAR

No. of replications B		200	400	600	800	1,000
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

CVs. Then, the question is: How many iterations L are sufficient for a specified B to guarantee the accuracy of the Monte Carlo computation?

3.2. *Determine the Number of Monte Carlo Iterations L*

Two fingerprint-image matching algorithms are employed[2]. Algorithm **A** is of high accuracy and Algorithm **B** is of low accuracy. The significance level is set to be 5%. As discussed in Section 1, the total number of genuine scores is about 60,000 and the total number of impostor scores is about 120,000. With this number of impostor scores, in Scenario 1, the FAR is set to be 0.001 so that the number of false-accept instances would be about 120 that is reasonable large (Wu et al., 2011). In Scenarios 2 and 3, the system threshold yielding an FAR 0.001 is chosen to show the operational significance for each algorithm (Wu et al., 2011). In the following context, the estimates of the three CVs for the SE, the lower bound and upper bound of 95% CI are denoted by $\hat{\text{CV}}$SE, $\hat{\text{CV}}$LB, and $\hat{\text{CV}}$UB, respectively.

In all five scenarios for each algorithm, the number of replications B first ran from 200 up to 1,000 at intervals of 200. For each of such Bs, the number of Monte Carlo iterations L ran from 100 up to 1,000 at intervals of 100, and thus 10 estimates of CVSEs, CVLBs, and CVUBs were generated, respectively. From these 10 estimates in each case, the minimum, maximum, and range were obtained. The tendencies of changes of CVs with respect to the numbers B and L in different scenarios are the same. Therefore, in Table 1 and Table 3 are presented only the results of Scenario 1, where the statistic of interest is TAR at a given FAR, for high-accuracy Algorithm **A** and low-accuracy Algorithm **B**, respectively. The results regarding all other four scenarios of the two algorithms can be found in Ref. Wu et al. (2010).

As shown in these two tables, both minimal $\hat{\text{CV}}$s and maximal $\hat{\text{CV}}$s decrease as the number of bootstrap replications B increases. For Algorithm **A**, the maximal $\hat{\text{CV}}$LBs and $\hat{\text{CV}}$UBs are less than 0.00007; for Algorithm **B**, they are less than 0.0012. Regarding the

[2]All algorithms used in this article are proprietary. They cannot be disclosed.

Table 2
High-accuracy Algorithm A's estimators of CVSEs, CVLBs, and CVUBs, while B ran from 1,200 up to 2,000 at intervals of 200 as the number of iterations was fixed at 500. The statistics of interest is TAR at a given FAR

No. of replications B	1,200	1,400	1,600	1,800	2,000
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

ranges, for Algorithm **A**, the ranges of 10 estimated $\hat{\text{CV}}$SEs vary from about 0.007 down to 0.002 and the ranges for the two bounds of 95% CIs are not greater than 0.000008; for Algorithm **B**, the ranges of 10 estimated $\hat{\text{CV}}$SEs vary from about 0.006 down to 0.003 and the ranges for the two bounds of 95% CIs are less than 0.0002.

As a result, to obtain the estimates of CVs at a fixed number of replications B higher than 1,000, the number of Monte Carlo iterations L does not need to run from 100 up to 1,000 at intervals of 100. To save tremendous computing time, while the number of replications B varied from 1,200 up to 2,000 at intervals of 200, the number of Monte Carlo iterations L was fixed at 500. The corresponding estimates of CVs for Algorithms **A** and **B** in Scenario 1 are shown in Tables 2 and 4. The results for all other four scenarios of the two algorithms can also be found in Ref. Wu et al. (2010).

3.3. *Determine the Number of Nonparametric Two-Sample Bootstrap Replications*

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. Concerning the distribution of SEs versus the distributions of lower bounds and upper bounds of 95% CIs, which are all created by Monte Carlo iterations as

Table 3
Low-accuracy Algorithm B's minimum, maximum, and range of 10 estimators of CVSEs, CVLBs, and CVUBs, as the number of iterations ran from 100 up to 1,000 at intervals of 100 for each specified B. B ran from 200 up to 1,000 at intervals of 200. The statistic of interest is TAR at a given FAR

No. of replications B		200	400	600	800	1,000
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 4
Low-accuracy Algorithm B's estimators of CVSEs, CVLBs, and CVUBs, while B ran from 1,200 up to 2,000 at intervals of 200 as the number of iterations was fixed at 500. The statistic of interest is TAR at a given FAR

No. of replications B	1,200	1,400	1,600	1,800	2,000
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

shown in Eq. (3), the magnitudes of the estimated means are quite different, and thus the magnitudes of the estimated $\hat{\text{CV}}$s are also quite different.

For instance, in Scenario 1 where the statistic of interest is the TAR at a given FAR, for high-accuracy Algorithm **A**, generated by 500 Monte Carlo iterations while the number of bootstrap replications B was set to be 2,000, the estimated $\hat{\text{SE}}$s of the distributions of the SEs, and the lower bounds and upper bounds of the 95% CIs are 0.0000053, 0.0000198, and 0.0000192, respectively. It shows that the distribution of the SEs has less dispersion than the distributions of the lower bounds and upper bounds of 95% CIs. This is because in the tail of the distribution fewer samples occur (Efron and Tibshirani, 1993).

The estimated means of the corresponding distributions are 0.000331, 0.992617, and 0.993913, where the estimated mean for the SEs is much less than 1 but those for the two bounds of the 95% CIs are close to 1 due to the matching accuracy of algorithm. Therefore, the corresponding estimated CVs are 0.016040, 0.000020, and 0.000019 as presented in the

Table 5
High-accuracy Algorithm A's minimum, maximum, and range of 10 estimators of CVSEs as the number of iterations ran from 100 up to 1,000 at intervals of 100 for each specified B in all five scenarios. B ran from 200 up to 1,000 at intervals of 200

No. of replications B		200	400	600	800	1,000
1. TAR (f)	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
2. TAR (t)	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
3. FAR (t)	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
4. EER	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
5. AURC	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

Table 6
High-accuracy Algorithm A's estimators of CVSEs while B ran from 1,200 up to 2,000 at intervals of 200 as the number of iterations was fixed at 500 in all five scenarios

No. of replications B	1,200	1,400	1,600	1,800	2,000
1. TAR (f)	0.021218	0.018613	0.017951	0.016331	0.016040
2. TAR (t)	0.019498	0.018551	0.018105	0.016516	0.015611
3. FAR (t)	0.021625	0.019059	0.017675	0.016809	0.015373
4. EER	0.020348	0.019152	0.017907	0.017155	0.015610
5. AURC	0.020065	0.019558	0.017755	0.015916	0.015397

last column of Table 2. It shows that the estimated $\hat{\text{CV}}$SE is much larger than the estimated $\hat{\text{CV}}$LB and $\hat{\text{CV}}$UB. This feature occurs in all other cases in Scenario 1 as presented in Table 1 through Table 4.

Generally speaking, the estimated $\hat{\text{CV}}$s for low-accuracy Algorithm **B** are greater than the corresponding $\hat{\text{CV}}$s for high-accuracy Algorithm **A**, except for CVLB and CVUB in Scenario 4 where the statistic of interest is EER (Wu et al., 2010). This is due to the combined impact of the magnitudes of the estimated means and $\hat{\text{S}}$Es in this scenario. Another thing worth pointing out is that the estimated $\hat{\text{CV}}$LBs and $\hat{\text{CV}}$UBs in Scenarios 3 and 4 are larger than those in other scenarios (Wu et al., 2010). This is because the magnitudes of the estimated means in these two scenarios are quite small.

Nonetheless, the crucial point is that the estimated $\hat{\text{CV}}$SEs are much larger than the corresponding estimated $\hat{\text{CV}}$LBs and $\hat{\text{CV}}$UBs with regard to the same number of bootstrap

Table 7
Low-accuracy Algorithm B's minimum, maximum, and range of 10 estimators of CVSEs as the number of iterations ran from 100 up to 1,000 at intervals of 100 for each specified B in all five scenarios. B ran from 200 up to 1,000 at intervals of 200

No. of replications B		200	400	600	800	1,000
1. TAR (f)	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
2. TAR (t)	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
3. FAR (t)	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
4. EER	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
5. AURC	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

Table 8
Low-accuracy Algorithm B's estimators of CVSEs while B ran from 1,200 up to 2,000 at intervals of 200 as the number of iterations was fixed at 500 in all five scenarios

No. of replications B	1,200	1,400	1,600	1,800	2,000
1. TAR (f)	0.023673	0.022299	0.021272	0.018918	0.017705
2. TAR (t)	0.020317	0.018649	0.018001	0.016775	0.015417
3. FAR (t)	0.020671	0.019016	0.018532	0.016877	0.015095
4. EER	0.020697	0.018524	0.016898	0.016887	0.016330
5. AURC	0.020828	0.017814	0.017853	0.016148	0.015287

replications B and the same number of Monte Carlo iterations L for both high- and low-accuracy algorithms in all five scenarios. This indicates that if the estimates of $\hat{\text{CV}}$SE satisfy a specified tolerance, then the corresponding estimates of $\hat{\text{CV}}$LB and $\hat{\text{CV}}$UB can meet the same tolerance as well. As a result, in order to determine the number of nonparametric two-sample bootstrap replications in ROC analysis on large datasets, only the estimates of $\hat{\text{CV}}$SE need to be investigated.

All estimated $\hat{\text{CV}}$SEs of Algorithms **A** and **B** in the five scenarios are shown in Table 5 through Table 8, and depicted in Figure 2. In the cases where the number of replications

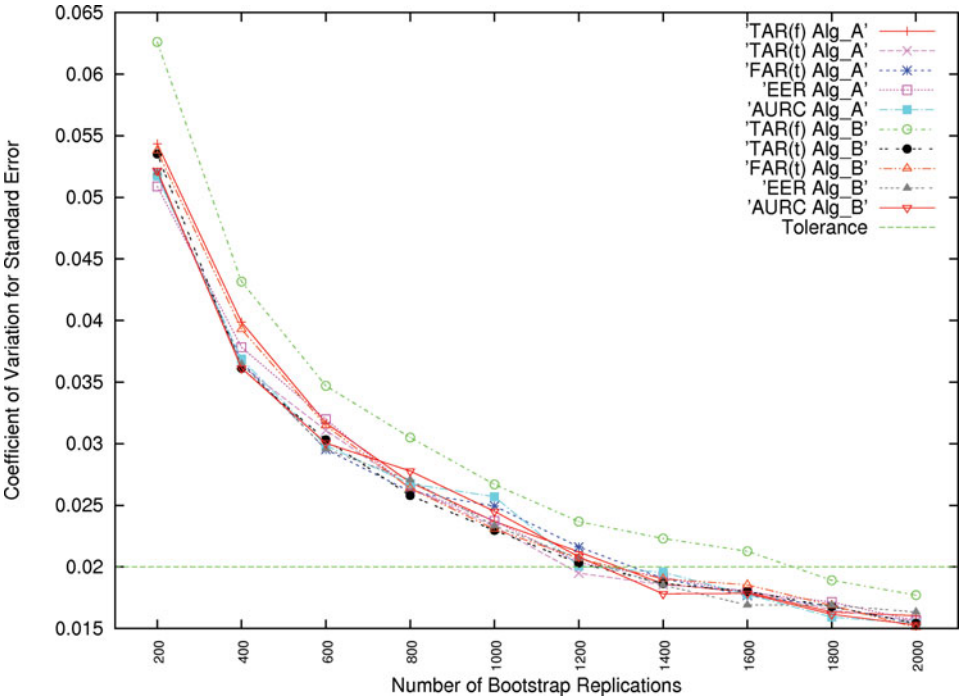

Figure 2. The estimators of CVSEs for high-accuracy Algorithm A and low-accuracy Algorithm B in all five scenarios as a function of the number of bootstrap replications. The tolerance is set to be 0.02. The statistics of interest are TAR at a specified FAR, TAR at a given threshold value, FAR at a given threshold value, EER, and AURC, respectively. (Color figure available online.)

B was set to be from 200 up to 1,000 at intervals of 200, only the maximal $\hat{\text{CV}}$SEs shown in Table 5 and Table 7 are employed. It shows that all estimated $\hat{\text{CV}}$SEs decrease as the number of replications *B* increases. All tables and figures regarding $\hat{\text{CV}}$LBs and $\hat{\text{CV}}$UBs can be found in Ref. Wu et al. (2010).

The tolerance for all CVSEs is set to be 0.02, which is acceptable for our applications (Efron and Tibshirani, 1993). As shown in Figure 2, with this 0.02 tolerance, for Scenario 1 where the statistic of interest is TAR at a given FAR, 1,400 two-sample bootstrap replications are sufficient for high-accuracy Algorithm **A**, and 1800 replications are sufficient for low-accuracy Algorithm **B**. In all other four scenarios for both Algorithms **A** and **B**, with the tolerance 0.02, 1,400 bootstrap replications are sufficient.

To reconcile numbers of bootstrap replications for all five scenarios where different statistics of interest are used as well as for different matching quality algorithms, and further to be more conservative, it is suggested that 2,000 nonparametric two-sample bootstrap replications be used in order to reduce the bootstrap variance and assure statistical accuracy of the computation in ROC analysis on large datasets.

4. Conclusions and Discussion

In our applications, the normality assumption for score distributions cannot be made, the statistics of interest are all probabilities related to ROC analysis 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 nonparametric two-sample bootstrap replications needed to reduce the bootstrap variance and ensure the accuracy of the computation.

The nonparametric two-sample bootstrap variability related to the SE and the two bounds of the 95% CI of bootstrap distributions was empirically studied without a normality assumption in five scenarios for ROC analysis on large datasets, 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. In addition, the bootstrap variability study was conducted on both high- and low-accuracy matching algorithms.

With a tolerance 0.02 for CVs, which is acceptable in our applications, 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 2,000.

As pointed out in Section 1, the bootstrap variance is also caused by the sample size. If the numbers of scores increased from what were used in this article, the measurement accuracy would improve little in our applications (Wu and Wilson, 2006). On the other hand, the recommended 2,000 bootstrap replications could certainly be applied to cases where smaller datasets are encountered, for instance in medical applications. If for some reason the number of bootstrap replications should need to be reinvestigated, the empirical methods for studying 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 (Cappelli et al., 2006), the false nonmatch rate (FNMR), which is equal to 1 – TAR, at a given FAR or threshold value is employed. It is trivial to show that in the Algorithm II (bootstrap variability study) shown in Section 3.1, with respect to the same two new sets of scores randomly selected WR from the two original sets of scores, the SE of FNMR is equal to the SE of TAR, while the lower bound and upper bound of the 95% CI for FNMR can be obtained by interchanging the two bounds for TAR

and subtracting them from 1. Thus, the two bounds of the 95% CIs of FNMR are quite close to 0 as opposed to 1 for TAR. Hence, if switching from TAR to FNMR, the CVSE will remain the same but the CVLB and CVUB will be larger.

For instance, in Scenario 1 for both Algorithms **A** and **B**, when the number of bootstrap replications B was set to be 2,000 as shown in the last columns of Table 2 and Table 4, the estimated $\hat{\text{CVLB}}$ and $\hat{\text{CVUB}}$ of high-accuracy Algorithm **A** changed from 0.000020 and 0.000019 to 0.003152 and 0.002687, respectively; those of low-accuracy Algorithm **B** changed from 0.000318 and 0.000389 to 0.001595 and 0.001196, respectively. However, they are all less than the 0.02 tolerance. Hence, the assertion that the number of nonparametric two-sample bootstrap replications be 2,000 is still valid if FNMR is employed. But 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, the statistics of interest is a detection cost function defined as a weighted sum of probabilities of type I and type II errors at a given threshold value (Wu et al., 2012). Such a metric is not dealt with in this article. However, the probability of type I error (i.e., FNMR) and the probability of type II error (i.e., FAR) have been examined here.

References

Cappelli, R., Maio, D., Maltoni, D., Wayman, J. L., Jain, A. K. (2006). Performance evaluation of fingerprint verification systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(1):3–18.

Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. *Annals of Statistics* 7:1–26.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association* 82(397):171–185.

Efron, B., Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters* 27:861–874.

Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics* 14(4):1453–1462.

Hyndman, R. J., Fan, Y. (1996). Sample quantiles in statistical packages. *American Statistician* 50:361–365.

Wu, J. C., Martin, A. F., Greenberg, C. S., Kacker, R. N. (2012). Data dependency on measurement uncertainties in speaker recognition evaluation. In: *Active and Passive Signatures III*, Proceedings of SPIE Vol. 8382, 83820D.

Wu, J. C., Martin, A. F., Kacker, R. N. (2010). *Further Studies of Bootstrap Variability for ROC Analysis on Large Datasets*. Gaithersberg, MD: National Institute of Standards and Technology.

Wu, J. C., Martin, A. F., Kacker, R. N. (2011). Measures, uncertainties, and significance test in operational ROC analysis. *Journal of Research of the National Institute of Standards and Technology* 116(1):517–537.

Wu, J. C., Wilson, C. L. (2006). An empirical study of sample size in ROC-curve analysis of fingerprint data. *Proceedings of SPIE* 6202: p. 620207. DOI:10.1117/12.665601

Wu, J. C., Wilson, C. L. (2007). Nonparametric analysis of fingerprint data on large data sets. *Pattern Recognition* 40(9):2574–2584.