Published in final edited form as:

Commun Stat Simul Comput. 2018 ; 48(2): . doi:10.1080/03610918.2018.1521974.

Monte Carlo studies of bootstrap variability in ROC analysis with data dependency

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

National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA

Abstract

ROC analysis involving two large datasets is an important method for analyzing statistics of interest for decision making of a classifier in many disciplines. And data dependency due to multiple use of the same subjects exists ubiquitously in order to generate more samples because of limited resources. Hence, a two-layer data structure is constructed and the nonparametric two-sample two-layer bootstrap is employed to estimate standard errors of statistics of interest derived from two sets of data, such as a weighted sum of two probabilities. In this article, to reduce the bootstrap variance and ensure the accuracy of computation, Monte Carlo studies of bootstrap variability were carried out to determine the appropriate number of bootstrap replications in ROC analysis with data dependency. It is suggested that with a tolerance 0.02 of the coefficient of variation, 2,000 bootstrap replications be appropriate under such circumstances.

Keywords

Bootstrap variability; Bootstrap replications; ROC analysis; Data dependency; Large datasets; Standard error

1. Introduction

The receiver operating characteristic (ROC) analysis involving two large datasets, i.e., genuine-score dataset and impostor-score dataset, is an important method to analyze statistics of interest for decision making of a classifier in many disciplines. A subject (e.g., face, speaker, etc.) may have different objects (e.g., images, speech segments, etc. correspondingly). To show the similarity of two objects decided by a classifier, comparing two different objects of the same subject creates a genuine score, whereas matching two objects of two different subjects generates an impostor score (Hanley and McNeil 1983; Fawcett 2006; Wu et al. 2017b; Wu, Martin, and Kacker 2011). The distributions of genuine scores and impostor scores characterize classifiers' matching abilities, and most importantly do not have well defined parametric forms (Wu et al. 2017b; Wu, Martin, and Kacker 2011; Wu and Wilson 2007).

CONTACT Jin Chu Wu jinchu.wu@nist.gov National Institute of Standards and Technology, Gaithersburg, MD 20899, USA.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp.

Different statistics are employed to measure the performance levels of classifiers depending on applications (Wu et al. 2017b; Wu, Martin, and Kacker 2011). One of the most complicated measures in ROC analysis is a weighted sum of the probabilities of type I error (miss) $\alpha(t)$ and type II error (false alarm) $\beta(t)$ at a given threshold t, which is called the detection cost function (DCF) (see Sec. 3) (Wu et al. 2017b). The former is the cumulative probability of genuine scores from the lowest score to a threshold t, while the latter is the one of impostor scores from the highest score to a threshold t, as schematically depicted in Figure 1a. As the threshold moves from the highest score down to the lowest score, an ROC curve is constructed as shown in Figure 1b.

A measure without an estimated standard error (SE) is incomplete, because the well-established statistical methods cannot be implemented while evaluating and comparing the performance levels for different matching systems to determine the statistical significance of their performance differences. It is difficult to compute the SE and confidence interval (CI) of the measure mentioned above analytically due to the covariance derived from the two negatively correlated probabilities of type I error and type II error (Wu et al. 2017b). The analytical approach cannot take account of how genuine scores and impostor scores are distributed, nor can it take data dependency into consideration (see below). It usually underestimates the SE of the measure (Wu et al. 2017a, 2017b). Thus, the SE and CI of such a measure is estimated using the bootstrap method (Wu et al. 2017b; Wu, Martin, and Kacker 2011; Wu et al. 2017a; Efron 1979; Efron and Tibshirani 1993; Hall 1986; Efron 1987). However, the bootstrap method assumes that the random samples drawn from a population are independent and identically distributed (i.i.d.).

Indeed, data dependency caused by multiple use of the same subjects in order to generate more samples because of limited resources exists ubiquitously, and many cases are related to ROC analysis on large datasets (Wu et al. 2017b; Efron and Tibshirani 1993; Liu and Singh 1992). Due to data dependency, the i.i.d. assumption is no longer valid, and the bootstrap samples cannot be randomly selected with replacement (WR) directly from the original data sample. Indeed, under such circumstances, the bootstrap method with i.i.d. assumption underestimates the SEs of statistics of interest, and the data dependency increases not only the estimated SEs but also the variations of SEs (Wu et al. 2017b).

As demonstrated in our research (Wu et al. 2017b), to preserve such data dependency while the bootstrap resampling takes place, and to make scores have equal probabilities of being selected in the random resampling, the data generated using the same subject may be grouped into a genuine set or an impostor set accordingly. All such genuine sets should have the same number of data points, and likewise for all impostor sets. Thus, a two-layer data structure is constructed: the first layer consists of the sets, and the second layer consists of scores within the sets. And later, because of such data structure, the same numbers of genuine (impostor) scores can be obtained at different iterations while the bootstrap algorithm is executed (see Sec. 4). All these reduce the variance of the computation.

Furthermore, as investigated in our prior work (Wu et al. 2017b), from the perspective of the multinomial probabilities of selecting bootstrap samples from the original scores and the distributions of the bootstrap replications of the statistic of interest, it was recommended that

the nonparametric two-sample two-layer bootstrap algorithm rather than the one-layer bootstrap and the conventional bootstrap with the i.i.d. assumption be adopted to estimate SEs and CIs of statistics of interest in ROC analysis on large datasets involving data dependency.

As described in our prior research (Wu et al. 2017b), the nonparametric two-sample two-layer bootstrap algorithm is referred to as without the assumption of any parametric model for score distributions, resampling randomly WR the genuine sets and the impostor sets first and subsequently scores within the sets. Scores within the same set are assumed to be conditionally independent, because they are generated by two sets of objects and objects in at least one of the two sets are different.

Then, the question arises: Under such circumstances as stated above, in order to reduce the bootstrap variance and ensure the accuracy of computation, how many bootstrap replications are sufficient, i.e., how many times should the raw data be resampled? This critical issue in ROC analysis with data dependency is the subject matter of this article.

This number is intrinsically related to bootstrap variability. As investigated in the literature (Efron and Tibshirani 1993; Hall 1986; Efron 1987), the substantial bootstrap variance is caused by the sampling variability and the bootstrap resampling variability. The former is because a finite number of samples usually form a subset of the entire population, and the latter is because a limited number of bootstrap replications of a statistic of interest created by resampling the raw datasets only constitute a subset of all possible bootstrap replications generated using multinomial probabilities of bootstrap selections.

Further, the bootstrap variance produces the variance of the SE and the variance of the two bounds of the CI of the bootstrap distribution formed by the bootstrap replications of the statistic. Hence, 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 samples employed in this article are tens of thousands of scores, and thus the sample size is large enough for ROC analysis in our applications (Wu and Wilson 2006). Therefore, only the number of bootstrap replications needs to be determined.

In the early analytical studies, the bootstrap variability was rigorously investigated, in which the normality assumption was made, the statistic of interest was a sample mean, the data were assumed to be i.i.d., and thus the coefficients of variation (CV) of SE and of two bounds of the CI were analytically derived (Efron and Tibshirani 1993; Hall 1986; Efron 1987). In ROC analysis on large datasets, as stated above, (1) no parametric model fits the distribution of score data samples; (2) the statistics of interest in ROC analysis are usually probabilities or even weighted sums of probabilities derived from two sets of data; and (3) data dependency is involved.

Under such circumstances, therefore, (1) it is imperative that bootstrap variability be re-studied in order to determine the appropriate number of bootstrap replications; and (2) it is worth noting that it is hard to conduct the bootstrap variability studies via deriving analytical formulas of CVs and thus the Monte Carlo method is employed in this article.

In our prior research, the bootstrap variability was studied in ROC analysis on large datasets under the conditions that the data were assumed to be i.i.d., and the appropriate number of bootstrap replications was determined to be 2,000 (Wu, Martin, and Kacker 2014). With i.i.d. data, there is no need to regroup the datasets into a two-layer data structure and the nonparametric two-sample bootstrap algorithm can be applied directly on the two sets of scores.

In our current research, the bootstrap variability in ROC analysis on large datasets is studied in a completely different and much more complicated scenario, where (1) data dependency is involved, (2) a two-layer data structure is constructed, and (3) the non-parametric two-sample two-layer bootstrap algorithm is employed to estimate the SEs and 95% CIs of statistics of interest.

The Monte Carlo method is employed to generate the CVs of the SE and of the two bounds of the CI of the bootstrap distribution of the statistic for bootstrap variability studies. These Monte Carlo studies required executions of a quarter billion times of resampling and sorting 12,672 genuine scores, and a quarter billion times of resampling and sorting 31,720 impostor scores (see Sec. 2). It took months of CPU time. With a tolerance 0.02 of the CV, it was found that 2,000 bootstrap replications were also appropriate for ROC analysis on large datasets with data dependency in order to reduce the bootstrap variance and ensure the accuracy of the computation.

In our Monte Carlo studies, 12 matching systems[1] with different performance accuracies are employed, which are real datasets with dependencies (Wu et al. 2017b; The NIST Speaker Recognition Evaluation 2012). And the statistic of interest involved is a weighted sum of two probabilities, which is one of the most complicated statistics encountered in ROC analysis. Thus, the conclusion drawn from our studies is stable, reliable, and effective to other cases of ROC analysis.

How to transform a dataset with data dependency into a two-layer structure based on probability theory is described in Sec. 2. A statistic of interest in ROC analysis is presented in Sec. 3. The nonparametric two-sample two-layer bootstrap algorithm is provided in Sec. 4. The Monte Carlo algorithm for the bootstrap variability study in ROC analysis with data dependency is presented in Sec. 5. The number of bootstrap replications needed in ROC analysis involving data dependency is derived in Sec. 6. Finally, conclusions and discussion can be found in Sec. 7.

2. The two-layer data structure for datasets involving data dependency

As described in Sec. 1, due to data dependency, a two-layer data structure is constructed. The first layer consists of genuine sets and impostor sets, and the second layer consists of genuine scores and impostor scores within sets. This structure preserves the data dependency

[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.

while the bootstrap resampling takes place (Wu et al. 2017b; Efron and Tibshirani 1993; Liu and Singh 1992).

In the following, the first subscript index indicates genuine or impostor, and the second and third subscript indices numerate sets and scores, respectively. The set $\boldsymbol{S}_{\mathrm{G}}$ of all genuine sets and the set $\mathbf{S}_{\mathrm{I}}$ of all impostor sets are expressed by

$$\boldsymbol{S}_{\mathrm{i}} = \left\{ \boldsymbol{S}_{\mathrm{ij}} \middle| \mathrm{j} = 1, \ldots, \mathrm{m_i} \right\}, \mathrm{i} \in \{\mathrm{G,I}\}, \tag{1}$$

where $\mathbf{S}_{\mathrm{Gj}}$ are genuine sets and $\mathbf{S}_{\mathrm{I\,j}}$ are impostor sets, and $\mathrm{m_G}$ and $\mathrm{m_I}$ are the numbers of the genuine sets and impostor sets, respectively. Each set is expressed in terms of its scores by

$$\boldsymbol{S}_{\mathrm{ij}} = \left\{ \alpha_{\mathrm{ijk}} \middle| \mathrm{k} = 1, \ldots, \mu_{\mathrm{ij}} \right\}, \mathrm{j} = 1, \ldots, \mathrm{m_i} \text{ and } \mathrm{i} \in \{\mathrm{G,I}\}, \tag{2}$$

where $\alpha_{\mathrm{G\,j\,k}}$ are genuine scores, $\alpha_{\mathrm{I\,j\,k}}$ are impostor scores, and $\mu_{\mathrm{i\,j}}$ stands for the number of scores in the corresponding set. The sets $\mathrm{S_{ij}}$ are all in the sense of multiset, in which members are allowed to appear more than once. Indeed, a score can occur multiple times within a set.

The total number of genuine scores $\mathrm{N_G}$ and the total number of impostor scores $\mathrm{N_I}$ are, respectively,

$$\mathrm{N_i} = \sum_{\mathrm{j}=1}^{\mathrm{m_i}} \mu_{\mathrm{ij}}, \text{ where } \mathrm{i} \in \{\mathrm{G,I}\}. \tag{3}$$

It was recommended that the nonparametric two-sample two-layer bootstrap algorithm, rather than the one-layer bootstrap or the conventional bootstrap with the i.i.d. assumption, be adopted to compute the SEs of statistics of interest in ROC analysis when data dependency is involved (Wu et al. 2017b). Thus, only the probability associated with the two-layer random resampling needs to be discussed here. The two-sample two-layer resampling takes place randomly WR not only on the first layer of the data, i.e., on the genuine and impostor sets, but also subsequently on the second layer of the data, i.e., the genuine and impostor scores in the sets, which are assumed to be conditionally independent as noted in Sec. 1.

Then, the probability for a score α_{ijk} in a set $\mathbf{S}_{\mathrm{ij}}$ being selected is

$$\begin{gathered} \mathrm{P}_{2-\text{ layer}}\left(\alpha_{\mathrm{ijk}}\right) = \mathrm{P}\left(\boldsymbol{S}_{ij}\right) \times P\left(\alpha_{ijk} \middle| \boldsymbol{S}_{ij}\right) = \frac{1}{\mathrm{m_i}} \times \frac{1}{\mu_{\mathrm{ij}}}, \\ \text{where } \mathrm{k} = 1, \ldots, \mu_{\mathrm{ij}}, \mathrm{j} = 1, \ldots, \mathrm{m_i} \text{ and } \mathrm{i} \in \{\mathrm{G,I}\}. \end{gathered} \tag{4}$$

These probabilities are the same for all scores within a set, regardless of whether it is a genuine set or an impostor set. Note, however, that the probabilities for scores being selected are different from set to set due to different numbers of scores in different sets indicated by $\mu_{\mathrm{i\,j}}$.

In order to reduce the variance of the computation, unequal selection probabilities for two-layer resampling must be eliminated; moreover, the numbers of scores obtained from iteration to iteration must be the same (see Sec. 4) while using the bootstrap to compute SEs and CIs of statistics of interest in ROC analysis with data dependency. Hence, the numbers of scores in genuine sets, i.e., $\mu_{G\,j}$, j = 1, … , m_G, are all set to a common μ_G so that each genuine score can have equal probability $1/N_G$ to be selected as per Eq. (3); and all $\mu_{I\,j}$, j = 1, … , m_I, are all set to a common μ_I so that the probability for each impostor score being selected is $1/N_I$.

Equal-number-of-score sets can be easily chosen from not-equal-number-of-score sets by randomly selecting without replacement scores from those sets, in which the numbers of scores are larger than a specified number. The specified numbers for genuine sets and for impostor sets are determined, respectively, by the trial-and-error optimization so that the total numbers of genuine scores and of impostor scores are as large as possible. Hence, the new datasets have 132 genuine sets, each of which contains 96 genuine scores; and the total number of genuine scores is 12,672. The new datasets contain 130 impostor sets, each of which has 244 impostor scores; and the total number of impostor scores is 31,720. The new data sizes still have tens of thousands of similarity scores (Wu and Wilson 2006).

3. A statistic of interest in ROC analysis

As stated in Sec. 1, one of the most complicated statistics in ROC analysis is a weighted sum of two probabilities. Hence, in our Monte Carlo studies of bootstrap variability with data dependency, the DCF at a threshold t (see Sec. 1) is chosen to be the statistic of interest. Just for the convenience of computing cumulative probabilities for type I error and type II error, the scores for a classifier are all converted to integer values if they are not, and expressed inclusively using the set $\mathbf{s} = \{s_{min}, s_{min}+1, \ldots, s_{max}\}$. While converting, as many decimal places of scores as possible are kept so that such a conversion does not result in loss of precision.

Let $f_i(s)$, $s \in \mathbf{s}$ and $i \in \{G, I\}$, denote the continuous probability density functions of scores and of genuine scores and impostor scores. The two-corresponding discrete probability distribution functions, denoted by $P_i(s)$, $s \in \mathbf{s}$ and $i \in \{G, I\}$, are expressed as

$$\boldsymbol{P}_i = \left\{ P_i(s) \,\middle|\, \forall s \in \mathbf{s} \text{ and } \sum_{s=Smin}^{Smax} P_i(s) = 1 \right\}, i \in \{G,I\}. \tag{5}$$

Then, the probabilities of type I error $\alpha(t)$ and type II error $\beta(t)$ at a threshold $t \in \mathbf{s}$ are expressed,

$$\alpha(t) = \int_{-\infty}^{t} f_G(s)ds = \sum_{s=Smin}^{t} P_G(s) = 1 - \sum_{s=t+1}^{Smax} P_G(s), \tag{6}$$

and

$$\beta(t) = \int_{t}^{+\infty} f_I(s)ds = \sum_{s=t}^{Smax} P_I(s), \tag{7}$$

where the boundary condition $P_G(s_{max} + 1) = 0$ is assumed and the normalization in Eq. (5) is employed (Wu, Martin, and Kacker 2011; Wu and Wilson 2007). For discrete probability distribution, while computing $\alpha(t)$ and $\beta(t)$ at a threshold t, the probabilities of genuine scores and impostor scores at this threshold t must be taken into account (Ostle and Malone 1988). Hence, in practice these two error rates can be obtained by moving one integer score at a time from the highest score s_{max} down to the threshold t to cumulate the probabilities of genuine scores and impostor scores, respectively.

The statistic of interest, i.e., the DCF, is a weighted sum of the probabilities of type I error $\alpha(t)$ and type II error $\beta(t)$ at a given threshold t (The NIST Speaker Recognition Evaluation 2012) and can be expressed thereafter in terms of the cumulative discrete probability distribution functions of genuine scores and impostor scores using Eqs. (6) and (7),

$$C(t) = C_{Miss} \times P_{Genuine} \times \left[1 - \sum_{s=t+1}^{Smax} P_G(s)\right] + C_{FalseAlarm} \times (1 - P_{Genuine}) \times \left[\sum_{s=t}^{Smax} P_I(s)\right]. \tag{8}$$

The thresholds were provided by the tested classifiers in order to make an explicit detection decision for each trial, and the parameters C_{Miss}, $C_{FalseAlarm}$, and $P_{Genuine}$ were set to be 10, 1, and 0.01 (The NIST Speaker Recognition Evaluation 2012).

How to design the DCF, how to choose the threshold, how to set these parameters, and how to generate genuine scores and impostor scores are all out of the scope of this article (The NIST Speaker Recognition Evaluation 2012; Doddington et al. 2000). However, these issues have no impact on how to estimate the SE of the DCF using the bootstrap algorithms on datasets with dependencies described in this article.

4. The nonparametric two-sample two-layer bootstrap algorithm due to data dependency

In Sec. 1 it was suggested that the nonparametric two-sample two-layer bootstrap algorithm be employed to estimate SEs and CIs of statistics of interest in ROC analysis when the two score distributions have no parametric model to fit and data involve dependency (Wu et al. 2017b). In the following, the superscript indices are used to numerate the resampling iterations. Here is a function Random_WR_Resampling_Set that will be frequently employed later on,

1. **function** Random_WR_Resampling_Set (N, $\boldsymbol{\Gamma}$, $\boldsymbol{\Theta}$)
2. **for** i = 1 **to** N **do**
3. select randomly WR an index j ∈ {1, … , N}
4. $\theta_i = \gamma_j$
5. **end for**
6. **end function**

where WR stands for "with replacement", $\boldsymbol{\Gamma}$ represents a set of sets or a set of scores, N is the cardinality of the set $\boldsymbol{\Gamma}$, $\boldsymbol{\Theta}$ represents a new set of sets or scores accordingly with the same cardinality, $\gamma_j \in \boldsymbol{\Gamma}$, and $\theta_i \in \boldsymbol{\Theta}$. Notice that this function can be applied to either a set of sets or a set of scores. It runs N iterations as shown from Step 2 to Step5. In the i-th iteration, a member of the set $\boldsymbol{\Gamma}$ is randomly selected WR to become a member of a new set $\boldsymbol{\Theta}$, as indicated by Steps 3 and 4. Hence, N members (sets or scores) are randomly selected WR from the set $\boldsymbol{\Gamma}$ to form a new set $\boldsymbol{\Theta}$.

The nonparametric two-sample two-layer bootstrap algorithm for data dependency is carried out by resampling randomly WR the genuine sets and the impostor sets first, and subsequently on scores within the sets where the scores are assumed to be conditionally independent as pointed out in Sec. 1. Thus, Algorithm I is shown as follows.

Algorithm I (Nonparametric two-sample two-layer bootstrap for data dependency)

1. **for** i = 1 **to** B **do**
2. Random_WR_Resampling_Set (m_G, $\boldsymbol{S}_G$, $\boldsymbol{S'}_G{}^i = \{\boldsymbol{S'}_{G\,j}{}^i \mid j = 1, \ldots, m_G\}$)
3. **for** k = 1 **to** m_G **do**
4. Random_WR_Resampling_Set (μ_G, $\boldsymbol{S'}_{G\,k}{}^i$, $\boldsymbol{S''}_{G\,k}{}^i$)
5. **end for**
6. Random_WR_Resampling_Set (m_I, $\boldsymbol{S}_I$, $\boldsymbol{S'}_I{}^i = \{\boldsymbol{S'}_{I\,j}{}^i \mid j = 1, \ldots, m_I\}$)
7. **for** k = 1 **to** m_I **do**
8. Random_WR_Resampling_Set (μ_I, $\boldsymbol{S'}_{I\,k}{}^i$, $\boldsymbol{S''}_{I\,k}{}^i$)
9. **end for**
10. $\boldsymbol{S''}_G{}^i = \{\boldsymbol{S}''_{G\,j}{}^i \mid j = 1, \ldots, m_G\}$ and $\boldsymbol{S}''_I{}^i = \{\boldsymbol{S}''_{I\,j}{}^i \mid j = 1, \ldots, m_I\} \Rightarrow$ statistic $\hat{C}^i$
11. **end for**
12. $\{\hat{C}^i | i = 1, ..., B\} \Rightarrow S\hat{E}_B$ *and* $(\hat{Q}_B(\alpha/2), \hat{Q}_B(1-\alpha/2))$
13. **end**

where B is the number of bootstrap replications. The set $\boldsymbol{S}_G$ of all genuine sets and the set $\boldsymbol{S}_I$ of all impostor sets are expressed in Eq. (1), and m_G and m_I are the cardinalities of the sets $\boldsymbol{S}_G$ and $\boldsymbol{S}_I$, respectively.

This algorithm runs B times, as shown from Steps 1 to 11. In the i-th iteration, as indicated by Steps 2 and 6, the function Random_WR_Resampling_Set is applied twice to the first layer of datasets, i.e., to sets rather than scores. That is, m_G genuine sets are randomly selected WR from the set $\boldsymbol{S}_G$ to constitute a new set $\boldsymbol{S'}_G{}^i = \{\boldsymbol{S'}_{G\,j}{}^i \mid j=1,\ldots, m_G\}$, and m_I impostor sets are randomly selected WR from the set $\boldsymbol{S}_I$ to form a new set $\boldsymbol{S'}_I{}^i = \{\boldsymbol{S'}_{I\,j}{}^i \mid j = 1, \ldots, m_I\}$.

Subsequently, the same function Random_WR_Resampling_Set is applied to the second layer of datasets, i.e., the scores in sets (see Eq. (2)). As indicated by Steps 3 to 5, m_G iterations take place. In the k-th iteration, μ_G genuine scores are randomly selected WR from the genuine set $\boldsymbol{S}'_{G\,K}{}^{i}$, which is the k-th new genuine set from the first-layer resampling, to form the k-th new genuine set $\boldsymbol{S}''_{G\,k}{}^{i}$ of the second-layer resampling. The analogous interpretation can be applied to impostor scores in the impostor set $\boldsymbol{S}'_{I\,K}{}^{i}$ as shown from Steps 7 to 9.

As indicated by Step 10, in the i-th iteration, all genuine scores in the new set $\boldsymbol{S}''_{G}{}^{i} = \{\boldsymbol{S}''_{G\,j}{}^{i} \mid j = 1, \ldots, m_G\}$ and all impostor scores in the new set $S''_{I}{}^{i} = \{S''_{I\,j}{}^{i} \mid j = 1, \ldots, m_I\}$ are employed to generate the i-th bootstrap replication of the statistic of interest $\hat{C}^i$ using Eq. (8).

Finally, after B iterations, as indicated by Step 12, from the set $\left\{\hat{C}^i \mid i = 1\ldots, B\right\}$, the standard error $\widehat{SE}_B$ of the statistic of interest is estimated by the sample standard deviation of the B replications, and the $(1-\alpha)$ 100% $\hat{CI}\left(\widehat{Q}_B(\alpha/2), \widehat{Q}_B(1-\alpha/2)\right)$ at the significance level α is estimated by the $\alpha/2$ 100% and $(1-\alpha/2)$ 100% quantiles of the bootstrap distribution (Efron and Tibshirani 1993). Definition 2 of quantile in Ref. Hyndman and Fan (1996) is adopted. Thus, the sample quantile is obtained by inverting the empirical distribution function with averaging at discontinuities. If 95% $\hat{CI}$ is of interest, then α is set to be 0.05.

With the new data structure shown in Sec. 2, the same number of genuine scores, i.e., $N_G = m_G \times \mu_G$, and the same number of impostor scores, i.e., $N_I = m_I \times \mu_I$, are obtained in Step 10 at different iterations of the nonparametric two-sample two-layer bootstrap algorithm to compute the bootstrap replication of the statistic of interest. This can reduce the variance of computation.

5. The Monte Carlo algorithm of bootstrap variability study with data dependency

As pointed out in Sec. 1, it is crucial to study the variances of the SE and of the two bounds of the CI of the bootstrap distribution of the statistic of interest in ROC analysis on large datasets where data dependency is involved. To take account of the impact of the mean value, the CV rather than variance is employed (Efron and Tibshirani 1993). Here is Algorithm II for bootstrap variability study for data dependency, in which the Monte Carlo method is employed to generate all those CVs.

Algorithm II (Bootstrap variability study for data dependency)

1. **for** i = 1 **to** L **do**
2. **for** j = 1 **to** B **do**
3. (Algorithm I: Step 2 through Step 9)^{ij}
4. $\boldsymbol{S}''_{G}{}^{ij} = \{\boldsymbol{S}''_{G,k}{}^{ij} \mid k = 1, \ldots, m_G\}$ and $\boldsymbol{S}''_{I}{}^{ij} = \{\boldsymbol{S}''_{I\,k}{}^{ij} \mid k = 1, \ldots, m_I\}$ $\Rightarrow$ statistic $\hat{C}^{ij}$
5. **end for**
6. $\{\hat{C}^{ij} | j = 1, ..., B\} \Rightarrow \hat{SE}_B^i$ *and* $(\hat{Q}_B^i(\alpha/2), \hat{Q}_B^i(1-\alpha/2))$
7. **end for**
8. $\{\hat{SE}_B^i, \hat{Q}_B^i(\alpha/2), \hat{Q}_B^i(1-\alpha/2) | i = 1, ..., L\} \Rightarrow \hat{CV}_{B,L}(\kappa), \kappa = \mathbf{SE}_{B,L}, \mathbf{Q}_{B,L}(\alpha/2), \mathbf{Q}_{B,L}(1-\alpha/2)$
9. **end**

where L is the number of Monte Carlo iterations and B is the number of bootstrap replications. As indicated by Steps 1 to 7, Algorithm II runs L iterations for a specified B. The part from Steps 2 to 6 is equivalent to the nonparametric two-sample two-layer bootstrap Algorithm I, which generates the i-th $\widehat{SE}_B^i$,$\widehat{Q}_B^i(\alpha/2)$ and $\widehat{Q}_B^i(1-\alpha/2)$ of statistic interest in the i-th iteration for a given B.

As shown in Step 8, for a specified B, after L iterations of executing the nonparametric two-sample two-layer bootstrap algorithm, the following three sets are generated,

$$\begin{aligned}
\mathbf{SE}_{B,L} &= \left\{ \widehat{SE}_B^i \middle| i = 1, \cdots, L \right\}, \\
\mathbf{Q}_{B,L}(\alpha/2) &= \left\{ \widehat{Q}_B^i(\alpha/2) \middle| i = 1, \cdots, L \right\}, \\
\mathbf{Q}_{B,L}(1-\alpha/2) &= \left\{ \widehat{Q}_B^i(1-\alpha/2) \middle| i = 1, \cdots, L \right\}.
\end{aligned} \tag{9}$$

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

$$\widehat{\mathrm{CV}}_{B,L}(\kappa) = \frac{\sqrt{\widehat{\mathrm{VAR}}_{B,L}(\kappa)}}{\widehat{\mathrm{E}}_{B,L}(\kappa)}, \text{ where } \kappa = \mathbf{SE}_{B,L}, \mathbf{Q}_{B,L}(\alpha/2), \mathbf{Q}_{B,L}(1-\alpha/2). \tag{10}$$

It is clear that the three estimated $\widehat{CV}$s are functions of the number of bootstrap replications B, the number of Monte Carlo iterations L, and the significance level α. In this study, the 95% $\widehat{\mathrm{CI}}$ is of interest and thus α is set to be 0.05, as stated in Sec. 4. Hence, the number of bootstrap replications B can be determined by the tolerable CVs as long as the sufficient number of iterations L can be determined in advance so that the accuracy of the Monte Carlo computation is guaranteed.

6. Results of bootstrap variability study in ROC analysis involving data dependency

To study bootstrap variability in ROC analysis where the datasets involve data dependency, 12 datasets generated by speaker recognition matching systems with different performance accuracies labeled with A through L were employed. Their DCFs are0.022199, 0.028996, 0.031588, 0.040098, 0.040880, 0.073500, 0.096988, 0.098744,0.161254, 0.223263, 0.236771, and 0.455384, respectively, at thresholds provided by the tested systems as pointed out in Sec. 3. The smaller the DCFs are, the more accurate the matching systems are. Thus, these systems represent a wide range of performance quality. As discussed in Sec. 5, the appropriate number of Monte Carlo iterations L needs to be determined prior to determining the number of bootstrap replications B, which is our primary objective. In the following, the three estimated $\widehat{\text{CV}}$s for the SE, the lower bound and upper bound of 95% CI are denoted by $\widehat{CVSE}$, $\widehat{CVLB}$, and $\widehat{CVUB}$ respectively.

6.1. The number of Monte Carlo iterations L

For all 12 matching systems, the number of bootstrap replications B first ran from 200 up to 1,000 at intervals of 200. For each specified B, the number of Monte Carlo iterations L ran from 100 up to 1,000 at intervals of 100; and thus 10 estimates of the CVSEs, CVLBs, and CVUBs were generated, from which the minimum, maximum, and range (i.e., max. − min.) were obtained for each of CVSE, CVLB, and CVUB.

All results of System A through System L when B ran from 200 up to 1,000 at intervals of 200 for CVSE, CVLB, and CVUB are depicted in Figure 2a–c, respectively. In Figure 2, each bar connects the minimum and the maximum of the 10 estimated CVs of a system at a specified B. To illustrate clearly, the 12 color bars at each B are drawn separately to represent the 12 systems alphabetically from left to right. To demonstrate the results of $\widehat{\text{CV}}$s clearly, the numerical results of the most accurate System A and the least accurate System L are presented in Tables 1 and 3.

In Figure 2b–c unlike Figure 2a, at any specified number of bootstrap replication B, the minimum and the maximum of CVLB and CVUB of a system are quite different from those of another system. This is because fewer samples occur in the tails of the distribution (Efron and Tibshirani 1993). Thus, the variations of CVLB and CVUB are quite large from system to system.

Nonetheless, as indicated by Figure 2, as well as by Tables 1 and 3, for a system, the tendency of changes of CVs is that minimal $\widehat{\text{CV}}$s and maximal $\widehat{\text{CV}}$s decrease, as the number of bootstrap replications B increases. Although the ranges fluctuate for some systems as B increases, the 12 systems' averages of CVSE ranges for each B running from 200 up to 1,000 at intervals of 200 are 0.006286, 0.004316, 0.003183, 0.003133, and 0.002739; the 12 systems' averages of CVLB ranges are 0.001014, 0.000605, 0.000646, 0.000707, and 0.000441; and the 12 systems' averages of CVUB ranges are 0.000991, 0.000526, 0.000544, 0.000461, and 0.000427. Hence, these 12 systems' averages of the CV ranges generally get smaller. In other words, the ranges become narrower as B increases.

As a result, to obtain the estimates of CVs while the number of bootstrap replications B varied from 1,200 up to 2,000 at intervals of 200, the number of Monte Carlo iterations L did not need to run from 100 up to 1,000 at intervals of 100 but was fixed at 500. This can save tremendous computing time without sacrificing computational accuracy. Under these circumstances, all estimates of such $\widehat{CVSE}$, $\widehat{CVLB}$, and $\widehat{CVUB}$ for the 12 matching systems are shown in Figures 3 and 4 (see Sec. 6.2). In the meantime, to show results clearly, the numerical results of Systems A and L are presented in Tables 2 and 4.

6.2. The number of nonparametric two-sample two-layer bootstrap replications B

Tables 1 through 4 show clearly that the estimated $\widehat{CVSE}$s are larger than the estimated $\widehat{CVLB}$s and $\widehat{CVUB}$s at any specified number of bootstrap replications B for a matching system. Indeed, this is a trend. The rationale is as follows.

According to Eq. (10), the CV is determined by both SE and mean. The SE and the mean associated with each of the three CVs are derived correspondingly by the distribution of SEs, or the distributions of lower bounds or upper bounds of the 95% CIs. All these distributions are created by the Monte Carlo iterations specified by Eq. (9). In the following, without loss of generality, the cases of B = 2,000 and L 500 regarding the most accurate System A and the least accurate System L are taken as examples.

The estimated $\widehat{\text{SE}}$s of the distributions of the SEs, and the lower bounds and upper bounds of the 95% CIs are 0.0000315, 0.0001076, and 0.0001329, respectively, for System A; and 0.0001495, 0.0005400, and 0.0005761, respectively, for System L. This shows that the distribution of the SEs has less dispersion than the distributions of the lower bounds and upper bounds of the 95% CIs. Again, this is because fewer samples occur in the tails of the distribution (Efron and Tibshirani 1993).

However, the estimated means of the corresponding distributions are 0.0019580,0.0186303, 0.0262923 for System A; and 0.0095160, 0.4356036, and 0.4728976 for System L. Among them, those associated with the two bounds of the 95% CIs are related to the matching accuracies of the systems, i.e., the DCF 0.022199 for System A and 0.455384 for System L as shown above. As a result, the corresponding estimated CVs are 0.016068, 0.005773, and 0.005055 for System A as shown in the last column of Table 2; and 0.015707, 0.001240, and 0.001218 for System L as presented in the last column of Table 4.

This shows that the estimated $\widehat{CVSE}$ is quite larger than the estimated $\widehat{CVLB}$ and $\widehat{CVUB}$ with B = 2,000 and L = 500 for Systems A and L. Indeed, all our results show that the estimated $\widehat{CVSE}$s are larger than the corresponding estimated $\widehat{CVLB}$s and $\widehat{CVUB}$s for any specified B and L and all matching systems.

This suggests that if the estimates of $\widehat{CVSE}$ satisfy a specified tolerance, then the corresponding estimates of $\widehat{CVLB}$ and $\widehat{CVUB}$ can meet the same tolerance as well. Hence, in order to determine the number of nonparametric two-sample two-layer bootstrap replications in ROC analysis on large datasets where data dependency is involved, only the estimates of $\widehat{CVSE}$ need to be investigated.

All estimated $\widehat{CVSE}$s of the 12 matching systems as a function of the number of bootstrap replications B running from 200 up to 2,000 at intervals of 200 are jointly depicted in Figure 3. In the cases where the number of bootstrap replications B is set to be from 200 up to 1,000 at intervals of 200, only the maximal $\widehat{CVSE}$s are employed. The same treatment is applied to $\widehat{CVLB}$ and $\widehat{CVUB}$. The corresponding 24 curves are jointly drawn in Figure 4.

The tolerance for all CVs is set to be 0.02, which is acceptable for our applications (Efron and Tibshirani 1993; Wu, Martin, and Kacker 2014). As shown in Figure 4, all estimated $\widehat{CVLB}$ and $\widehat{CVUB}$ are less than this tolerance. With this 0.02 tolerance, Figure 3 shows that 1,600 nonparametric two-sample two-layer bootstrap replications are sufficient. If the tolerance is set to be 0.0175, then Figure 3 shows that 2,000 nonparametric two-sample two-layer bootstrap replications are sufficient. Indeed, there is no significant difference between 0.0175 and 0.02 as far as the tolerance is concerned for our applications.

To be consistent with the number of bootstrap replications suggested in our prior bootstrap variability study where the data were assumed to be i.i.d. (Wu, Martin, and Kacker 2014), and indeed to be a bit more conservative, it is suggested that 2,000 nonparametric two-sample two-layer bootstrap replications be used in ROC analysis on large datasets where data dependency is involved in order to reduce the bootstrap variance and assure the statistical accuracy of the computation.

7. Conclusions and discussion

The nonparametric two-sample two-layer bootstrap method is very useful in estimating SEs and CIs of statistics of interest in ROC analysis on large datasets in which data dependency is involved. The number of bootstrap replications is a critical parameter for reducing the bootstrap variance and ensuring the accuracy of the computation. And this number can be determined from the bootstrap variability study.

In ROC analysis on large datasets, usually no parametric model fits score distributions and the statistics of interest are typically probabilities derived from two sets of data. In our prior bootstrap variability studies, the data were assumed to be i.i.d., and thus the nonparametric two-sample bootstrap resampling took place only on scores without a need to group them into sets. With a tolerance 0.02 for CVs, the appropriate number of the nonparametric two-sample bootstrap replications was determined to be 2,000 (Wu, Martin, and Kacker 2014).

However, data dependency caused by multiple use of the same subjects in order to generate more data due to limited resources exists ubiquitously in many disciplines involving ROC analysis. The data dependency requires that a two-layer data structure be constructed and the nonparametric two-sample two-layer bootstrap algorithm be used to estimate SEs and CIs of statistics of interest. In the light of these fundamental changes, the bootstrap variability was re-studied in order to determine the appropriate number of the bootstrap replications.

To conduct bootstrap variability studies, under the circumstances described above, it is difficult to derive analytical formulas of CVs. Instead, the Monte Carlo method is employed to generate the CVs of the SE and of the two bounds of the CI of the bootstrap distribution

formed by the bootstrap replications of the statistic of interest. In the Monte Carlo studies with data dependency, 12 matching systems with different performance accuracies were taken as examples, and a weighted sum of two probabilities derived from two sets of data, which is more complicated than others often employed in ROC analysis, was used as the statistic of interest.

Moreover, based on the numbers of Monte Carlo iterations and the numbers of bootstrap replications as stated in Sec. 6, and the total numbers of genuine scores and impostor scores as noted in Sec. 2 for each of the 12 matching systems, the Monte Carlo studies took months of CPU time. Thus, the conclusion of 2,000 bootstrap replications in ROC analysis on large datasets with data dependency is stable, reliable, and effective to other cases of ROC analysis.

To take account of the impact of the mean value, the CVs rather than the variances are employed (Efron and Tibshirani 1993). Further, as shown in Sec. 6.2, the means of the distributions of the lower bounds and upper bounds of the 95% CIs of a measure, and thus CVLB and CVUB are related to the matching accuracies of the systems. In the prior bootstrap variability study, the matching accuracies, such as the true accept rate at a fixed false accept rate or at a given threshold, etc., were very high close to 1 (Wu, Martin, and Kacker 2014). However, in this bootstrap variability study, they are between0.022199 and 0.455384 (see Sec. 6.2). This may cause that the $\widehat{CVLB}$s and $\widehat{CVUB}$s in the prior study were quite smaller than those in this study.

Nonetheless, all estimated $\widehat{CVLB}$s and $\widehat{CVUB}$s are smaller than the estimated $\widehat{CVSE}$s under the same assumptions. This implies that only the estimates of $\widehat{CVSE}$s need to be investigated while comparing with the tolerance of CVs. With a tolerance 0.02 of CVs, which is the same as the one adopted in our prior bootstrap variability study, and to be more conservative, it is suggested that the appropriate number of the nonparametric two-sample two-layer bootstrap replications be 2,000 in ROC analysis on large datasets where data dependency is involved, in order to reduce the bootstrap variance and ensure the accuracy of the computation.

As stated in Sec. 1, the bootstrap variance is also related to the sample sizes. The sample sizes employed in this article are appropriate for ROC analysis in our applications. This is because if the numbers of scores increased from those used in this article, the measurement accuracy would improve little in our applications (Wu and Wilson 2006). Nonetheless, should bootstrap variability need to be reinvestigated for some reason, the Monte Carlo methods for studying bootstrap variability developed in this article would remain the same.

References

Doddington GR, Przybocki MA, Martin AF, and Reynolds DA. 2000 The NIST speaker recognition evaluation: Overview, methodology, systems, results, perspective. Speech Communication 31 (2–3):225–54.

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

Efron B 1987 Better bootstrap confidence intervals. Journal of the American Statistical Association 82 (397):171–85.

Efron B, and Tibshirani RJ. 1993 An introduction to the bootstrap. New York: Chapman & Hall.

Fawcett T 2006 An introduction to ROC analysis. Pattern Recognition Letters 27 (8):861–74.

Hall P 1986 On the number of bootstrap simulations required to construct a confidence interval. The Annals of Statistics 14 (4):1453–62.

Hanley JA, and McNeil BJ. 1983 A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148 (3):839–43. [PubMed: 6878708]

Hyndman RJ, and Fan Y. 1996 Sample quantiles in statistical packages. American Statistician 50 (4):361–5.

Liu RY, and Singh K. 1992 Moving blocks jackknife and bootstrap capture weak dependence In Exploring the limits of bootstrap, LePage Rand Billard L, eds. New York: John Wiley.

Ostle B, and Malone LC. 1988 Statistics in research: Basic concepts and techniques for research workers, 4th ed Ames: Iowa State University Press.

The NIST Speaker Recognition Evaluation. 2012 https://www.nist.gov/sites/default/files/documents/itl/iad/mig/NIST_SRE12_evalplan-v17-r1.pdf.

Wu JC, Halter M, Kacker RN, Elliott JT, and Plant AL. 2017a A novel measure and significance testing in data analysis of cell image segmentation. BMC Bioinformatics 18:168. [PubMed: 28292256]

Wu JC, Martin AF, Greenberg CS, and Kacker RN. 2017b The impact of data dependence on speaker recognition evaluation. IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (1):5–18.

Wu JC, Martin AF, and Kacker RN. 2011 Measures, uncertainties, and significance test in operational ROC analysis. Journal of Research of the National Institute of Standards and Technology 116 (1):517–37. [PubMed: 26989582]

Wu JC, Martin AF, and Kacker RN. 2014 Bootstrap variability studies in ROC analysis on large datasets. Communications in Statistics: Simulation and Computation 43 (1):225–36.

Wu JC, and Wilson CL. 2006 An empirical study of sample size in ROC-curve analysis of fingerprint data. In Proceedings of SPIE biometric technology for human identification III 6202:620207.

Wu JC, and Wilson CL. 2007 Nonparametric analysis of fingerprint data on large data sets. Pattern Recognition 40 (9):2574–84.

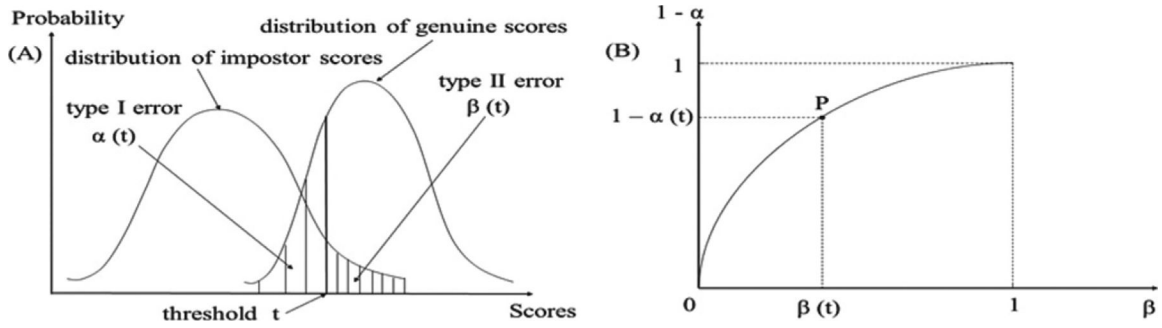

Figure 1.

(a): A schematic diagram of two continuous distributions of genuine scores and impostor scores, showing three related variables: type I error $\alpha(t)$, type II error $\beta(t)$, and threshold t. (b): A schematic drawing of an ROC curve.

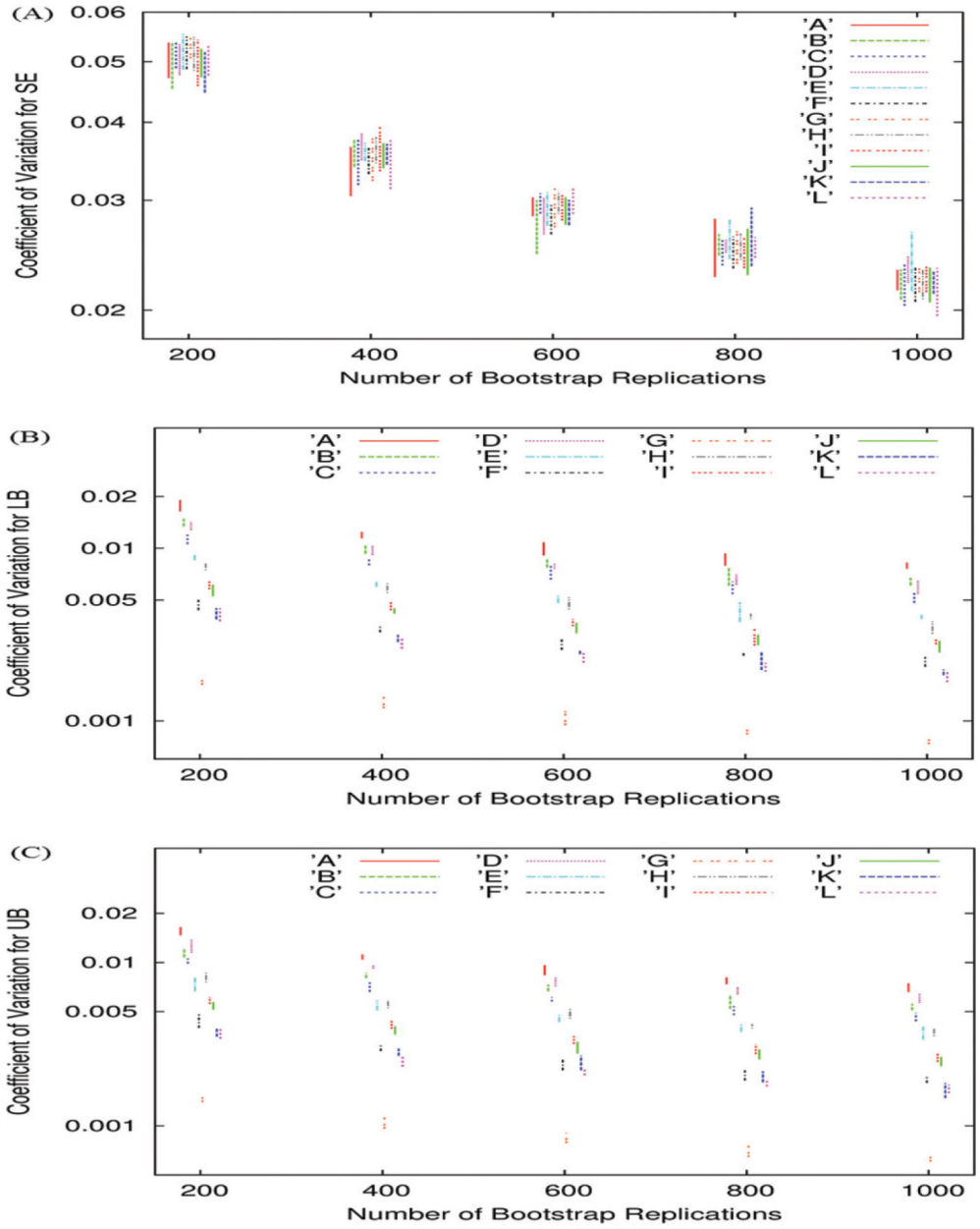

Figure 2.

Bars between minimum and maximum of 10 estimates of CVSEs (a), CVLBs (b), and CVUBs(c), as the number of Monte Carlo iterations *L* ran from 100 up to 1,000 at intervals of 100 for each specified *B*, and the number of bootstrap replications B ran from 200 up to 1,000 at intervals of 200 for System A through System L. To illustrate clearly, the 12 color bars at each B are drawn separately to represent the 12 systems alphabetically from left to right. The vertical axis is on a log scale.

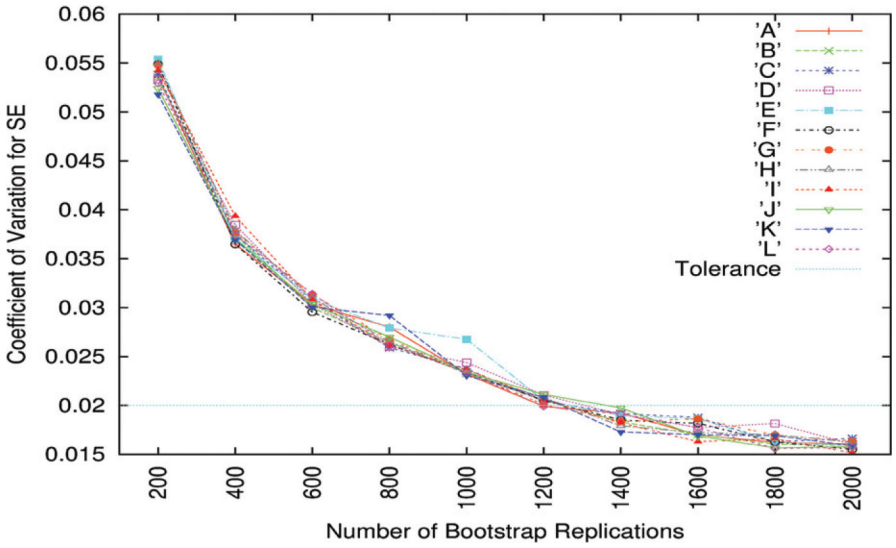

Figure 3.

The estimates of CVSEs for all 12 systems as a function of the number of bootstrap replications, where the statistic of interest is a DCF. The tolerance is set to be 0.02.

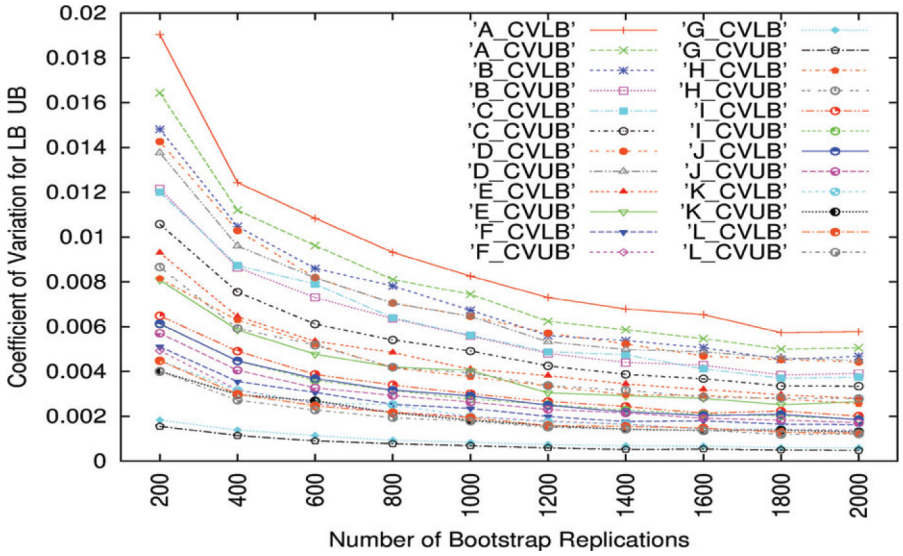

Figure 4.

The estimates of CVLBs and CVUBs for all 12 systems as a function of the number of bootstrap replications, where the statistic of interest is a DCF.

Table 1.

System A's minimum, maximum, and range of 10 estimates of CVSEs, CVLBs, and CVUBs, as the number of Monte Carlo iterations L 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.

Num. of replications *B*	**200**	**400**	**600**	**800**	**1,000**
CVSE					
Min.	0.047070	0.030467	0.028298	0.022614	0.021531
Max.	0.053639	0.036492	0.030321	0.028009	0.023215
Range	0.006569	0.006026	0.002023	0.005395	0.001683
CVLB					
Min.	0.016383	0.011433	0.009101	0.007927	0.007607
Max.	0.019033	0.012432	0.010839	0.009316	0.008256
Range	0.002650	0.000999	0.001738	0.001389	0.000649
CVUB					
Min.	0.014699	0.010422	0.008353	0.007367	0.006589
Max.	0.016435	0.011199	0.009617	0.008098	0.007444
Range	0.001736	0.000777	0.001263	0.000731	0.000855

Table 2.

System A's estimates of CVSEs, CVLBs, and CVUBs, while B ran from 1,200 up to 2,000 at intervals of 200 as the number of Monte Carlo iterations L was fixed at 500.

Num. of replications *B*	**1,200**	**1,400**	**1,600**	**1,800**	**2,000**
CVSE	0.019914	0.019226	0.016977	0.016226	0.016068
CVLB	0.007298	0.006788	0.006538	0.005732	0.005773
CVUB	0.006240	0.005862	0.005462	0.004999	0.005055

Table 3.

System L's minimum, maximum, and range of 10 estimates of CVSEs, CVLBs, and CVUBs, as the number of Monte Carlo iterations L 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.

Num. of replications *B*	**200**	**400**	**600**	**800**	**1,000**
CVSE					
Min.	0.047549	0.031298	0.028535	0.024287	0.019595
Max.	0.052935	0.037416	0.031404	0.026173	0.023385
Range	0.005385	0.006118	0.002869	0.001886	0.003790
CVLB					
Min.	0.003772	0.002631	0.002182	0.001928	0.001682
Max.	0.004479	0.002989	0.002468	0.002167	0.001941
Range	0.000707	0.000357	0.000286	0.000239	0.000259
CVUB					
Min.	0.003420	0.002308	0.002043	0.001741	0.001586
Max.	0.004017	0.002720	0.002276	0.001936	0.001784
Range	0.000598	0.000411	0.000234	0.000195	0.000197

Table 4.

System L's estimates of CVSEs, CVLBs, and CVUBs, while B ran from 1,200 up to 2,000 at intervals of 200 as the number of Monte Carlo iterations L was fixed at 500.

Num. of replications *B*	**1,200**	**1,400**	**1,600**	**1,800**	**2,000**
CVSE	0.019895	0.019107	0.017731	0.015606	0.015707
CVLB	0.001581	0.001547	0.001484	0.001289	0.001240
CVUB	0.001514	0.001429	0.001376	0.001190	0.001218