University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2012 *Paper* 304

Computationally Efficient Confidence Intervals for Cross-validated Area Under the ROC Curve Estimates

Erin LeDell* Maya L. Petersen†

Mark J. van der Laan‡

*Division of Biostatistics, University of California, Berkeley, ledell@berkeley.edu

†Division of Biostatistics, University of California, Berkeley, mayaliv@berkeley.edu

‡Division of Biostatistics, University of California, Berkeley, laan@berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper304

Computationally Efficient Confidence Intervals for Cross-validated Area Under the ROC Curve Estimates

Erin LeDell, Maya L. Petersen, and Mark J. van der Laan

Abstract

In binary classification problems, the area under the ROC curve (AUC), is an effective means of measuring the performance of your model. Most often, cross-validation is also used, in order to assess how the results will generalize to an independent data set. In order to evaluate the quality of an estimate for cross-validated AUC, we must obtain an estimate for its variance. For massive data sets, the process of generating a single performance estimate can be computationally expensive. Additionally, when using a complex prediction method, calculating the cross-validated AUC on even a relatively small data set can still require a large amount of computation time. Thus, when the processes of obtaining a single estimate for cross-validated AUC is significant, the bootstrap, as a means of variance estimation, can be computationally intractable. As an alternative to the bootstrap, we demonstrate a computationally efficient influence curve based approach to obtaining a variance estimate for cross-validated AUC.

1 Introduction

The area under the ROC curve, or AUC, is a ranking-based measure of classification performance, which is a popular performance measure in binary classification problems. Its value can be interpreted as the probability that a randomly selected positive sample will rank higher than a randomly selected negative sample. AUC is a more discriminating performance measure than accuracy [Ling et al., 2003], and is invariant to relative class distributions [Bradley, 1997]. Due to its many strengths over other performance measures, AUC is widely used.

In practice, we are generally concerned with how well our results will generalize to new data. Cross-validation is a means of obtaining an estimate that is generalizable to data outside your training set, or can also be used to perform model selection. Common types of cross-validation procedures include V-fold [Geisser, 1975], leave-one-out [Stone, 1974, Allen, 1974, Geisser, 1975], and leave-p-out [Shao, 1993] cross-validation. Given the advantages of AUC as a performance measure, along with the desire to produce generalizable results, cross-validated AUC is a frequently used estimate in binary classification problems.

An important task in any estimation procedure is evaluating the quality of your estimates. In many cases, specification of a parametric model known to contain the truth is not possible, and approaches to inference which are robust to model misspecification are therefore needed. Two approaches to robust inference include inference based on resampling methods and inference based on influence curves. In practice, the use of resampling methods such as the nonparametric bootstrap [Efron, 1979, Efron and Tibshirani, 1993], is quite common due to their generic nature and simplicity. However, when data sets are large or when prediction methods are complex, bootstrapping can quickly become a computationally prohibitive procedure.

In machine learning, ensemble methods are prediction methods that make use of, or combine, several or many candidate learning algorithms to obtain better predictive performance than could be obtained from any of the constituent algorithms alone. This boost in performance often comes with a computational cost. Although cross-validation lends itself well to parallelization, it can still take hours, days or even weeks to generate a cross-validated performance measure, such as cross-validated AUC, depending on the complexity of the algorithm. Alternatively, given massive data sets, even simple prediction methods can be computationally expensive. In cases where obtaining a single estimate of cross-validated AUC requires a significant amount of time and/or resources, the bootstrap is either not an option, or at the very least, a undesirable option for obtaining variance estimates.

As a response to the computational costs of the bootstrap, variations of the bootstrap have been developed that achieve a more desirable computational footprint, such as the "m out of n bootstrap" [Bickel et al., 1997] and subsampling [Politis et al., 1999]. Another recent advancement that has been made in this area is the "Bag of Little Bootstraps" (BLB) method [Kleiner et al., 2011]. Unlike previous variations, BLB simultaneously addresses computational costs, statistical correctness and automation, which appears to be a promising generalized method for variance estimation on massive data sets.

Regardless of the reduction in computation that different variations of the bootstrap offer, all bootstrapping variants require repeated estimation on some subset of the data. Using influence curves for variance estimation, we avoid the need to fit additional models. In order to estimate variance using influence curves, you must first, unsurprisingly, calculate the influence curve for your estimator. For complex estimators, it can be a difficult task to derive the influence curve. However, once the derivation is complete, variance estimation is reduced to a simple and computationally negligible calculation. This is the main motivation for our use of influence curves as a means of variance estimation.

The main goal of this paper is to establish an influence curve based approach for estimating the asymptotic variance of the cross-validated area under the ROC curve estimator. We first provide a brief review of influence curve based variance estimation. We then demonstrate how to construct confidence intervals for the risk of an estimator using this method. Our target parameter, true

cross-validated AUC, is then defined, along with a corresponding estimator. We derive the influence curve for the AUC estimate for both i.i.d. data and pooled repeated measures data (multiple observations per independent sampling unit, such as a patient), and demonstrate the construction of 95% confidence intervals for these estimators. This procedure has been implemented as an R package called `cvAUC`, which we describe and provide a code example for. We conclude with a simulation that evaluates the coverage probability of the confidence intervals over data sets of varying dimension.

2 Influence curves for variance estimation

We provide a brief overview of influence curves and their relation to variance estimation. We outline the general procedure for obtaining confidence intervals using the influence curve of an estimator. This section serves as a gentle introduction to concepts and notation used throughout the paper.

Suppose that $O \equiv O_1, ..., O_n$ are i.i.d. samples from a probability distribution, P_0, that is known to be an element of a statistical model, $\mathcal{M}$. Let $\mathcal{F}$ be some class of functions of O. Throughout this paper, we will use the notation Pf, where P is a probability distribution, to denote $\int f(x)dP(x)$. We consider the empirical process, $(P_0 f : f \in \mathcal{F})$, which is a "vector" of true means. Let $\Psi : \mathcal{M} \to \mathbb{R}^d$ be a parameter of interest, and let $\psi_0 = \Psi(P_0) \equiv \Psi(P_0 f : f \in \mathcal{F})$ be the true parameter value; ψ_0 is a function of true means. In order to assume that asymptotically linear estimators of ψ_0 exist, we must assume that the parameter Ψ is pathwise differentiable [Bickel et al., 1993].

Let $\mathcal{M}_{NP}$ denote a nonparametric model that includes the empirical distribution, P_n, of $O_1, ..., O_n$. We consider the empirical process, $(P_n f : f \in \mathcal{F})$, which is a "vector" of empirical means. Let $\hat{\Psi} : \mathcal{M}_{NP} \to \mathbb{R}^d$ be an estimator of ψ_0 that maps the empirical distribution, P_n, or rather, a "vector" of empirical means, into an estimate $\hat{\Psi}(P_n) \equiv \hat{\Psi}(P_n f : f \in \mathcal{F})$. We assume that $\hat{\Psi}(P_0) = \psi_0$, so that the estimator targets the desired target parameter, ψ_0. This estimate is *asymptotically linear* at P_0 if

$$\hat{\Psi}(P_n) - \hat{\Psi}(P_0) = (P_n - P_0)IC(P_0) + o_P(1/\sqrt{n}),$$

for some mean zero function $IC(P_0)$ of O: i.e., $P_0 IC(P_0) = 0$. This function $IC(P_0)$ of O is called the *influence curve* of the estimator $\hat{\Psi}$.

Since $IC(P_0)$ is a zero mean function of O, we observe that $(P_n - P_0)IC(P_0) = \frac{1}{n}\sum_{i=1}^n IC(P_0)(O_i) - P_0 IC(P_0) = \frac{1}{n}\sum_{i=1}^n IC(P_0)(O_i)$, which is an empirical mean of mean zero i.i.d. random variables. So we have,

$$\hat{\Psi}(P_n) - \hat{\Psi}(P_0) = \frac{1}{n}\sum_{i=1}^n IC(P_0)(O_i) + o_P(1/\sqrt{n}).$$

By the Central Limit Theorem, we find that

$$\sqrt{n}\ \left(\hat{\Psi}(P_n) - \hat{\Psi}(P_0)\right) \xrightarrow{d} \mathcal{N}(0, \Sigma_0),$$

where $\Sigma_0 = P_0 IC(P_0) IC(P_0)^T$. This covariance matrix can be estimated with the empirical covariance matrix $\widehat{IC}(O_i)$, $i = 1, ..., n$ where $\widehat{IC}$ is an estimate of $IC(P_0)$. This method for establishing the asymptotic linearity and normality of the estimator is called the functional delta method [van der Vaart and Wellner, 1996, Gill, 1989]. The functional delta method is a generalization of the classical delta method for finite dimensional functions of a finite set of estimators.

When our target parameter is one-dimensional, as in cross-validated AUC, we can write the following:

$$\sqrt{n}\ \left(\hat{\Psi}(P_n) - \hat{\Psi}(P_0)\right) \xrightarrow{d} \mathcal{N}(0, \Phi^2(P_0)),$$

where $\Phi^2(P_0) = \int IC(P_0)(x)^2 dP_0(x)$. Let $\Phi^2(P_n)$ be an estimate of the asymptotic variance, $\Phi^2(P_0)$, where P_n is the empirical distribution. For example, we could estimate $\Phi^2(P_0)$ by

$$\Phi_n^2 = \Phi^2(P_n) = \frac{1}{n}\sum_{i=1}^{n} IC(P_n)(O_i)^2,$$

however, other estimators of the variance of the influence curve can be considered. Let z_r denote the r^{th} quantile of the standard normal distribution. It follows that for any estimate $\Phi_n^2 = \Phi^2(P_n)$ of $\Phi^2(P_0)$, we have that

$$\left(\hat{\Psi}(P_n) - z_{1-\alpha/2}\frac{\Phi_n}{\sqrt{n}}\ ,\ \ \hat{\Psi}(P_n) + z_{1-\alpha/2}\frac{\Phi_n}{\sqrt{n}}\right)$$

forms an approximate $100 \times (1-\alpha)\%$ confidence interval for $\psi_0 = \hat{\Psi}(P_0)$.

3 Cross-validated AUC as a target parameter

In this section, we formally introduce AUC. We then define the estimator for cross-validated AUC, as well as the target that it is estimating, the true cross-validated AUC.

Consider some probability distribution, P_0, that is known to be an element of a statistical model, $\mathcal{M}$. Let $O = (W, Y) \sim P_0 \in \mathcal{M}$, where Y is a binary outcome variable, and W represents one or more covariates or predictor variables. Without loss of generality, we will denote $Y = 1$ as the positive class and $Y = 0$ as the negative class, and ψ as a function that maps W into $(0, 1)$. The quantity, $\psi(W)$, is the predicted value or score of a sample. The Area Under the ROC curve can be defined as the following:

$$AUC(P_0, \psi) = \int_0^1 P_0\left(\psi(W) > c \mid Y = 1\right) P_0\left(\psi(W) = c \mid Y = 0\right) dc.$$

Alternatively, we can define AUC as

$$AUC(P_0, \psi) = P_0\left(\psi(W_1) > \psi(W_2) \mid Y_1 = 1, Y_2 = 0\right),$$

where (W_1, Y_1) and (W_2, Y_2) are i.i.d. samples from P_0. The quantity, $AUC(P_0, \psi)$, the true AUC, equals the probability, conditional on sampling two independent observations where one is positive $(Y_1 = 1)$ and the other is negative $(Y_2 = 0)$, that the predicted value (or rank) of the positive sample, $\psi(W_1)$, is higher than the predicted value (or rank) of the negative sample, $\psi(W_2)$.

Consider $O_1, ..., O_n$, i.i.d. samples from P_0, such that $O_i = (W_i, Y_i)$ for each i, and let P_n denote the empirical distribution. Let n_0 be the number of observations with $Y = 0$ and let n_1 be the number of observations with $Y = 1$. In machine learning, the ψ function is what is learned by a binary prediction algorithm using the training data. The AUC of the empirical distribution can be written as follows:

$$\begin{aligned} AUC(P_n, \psi) &= \frac{1}{n_0 n_1}\sum_{i=1}^{n}\sum_{j=1}^{n} I(\psi(W_j) > \psi(W_i)) I(Y_i = 0, Y_j = 1) \\ &= \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} I(\psi(W_j) > \psi(W_i)), \end{aligned}$$

where I is the indicator function.

The parameter we targeting is true cross-validated AUC. We do not require that the cross-validation be any particular type of cross-validation, such as V-fold, however, in practice, V-fold is common.

We will use a generalized notation to encode the data splitting procedure, where a binary indicator vector is used to specify which observations belong to the validation sample at each iteration of the cross-validation process.

Let $B_n \in \{0,1\}^n$ be a random split and let P^1_{n,B_n} and P^0_{n,B_n} be the empirical distributions of the validation $\{i : B_n(i) = 1\}$ and training sample $\{i : B_n(i) = 0\}$, respectively. We assume that B_n has only a finite number of values uniformly in n, as in V-fold cross-validation.

Recall that $O_1, ..., O_n \sim P_0 \in \mathcal{M}$ and let $\Psi : \mathcal{M} \to \mathbf{\Psi}$. We denote the target parameter as $\psi_0 = \Psi(P_0)$. Let $\mathcal{M}_{NP}$ denote a nonparametric model that includes the empirical distribution, P_n, of $O_1, ..., O_n$ and let $\hat{\Psi} : \mathcal{M}_{NP} \to \mathbb{R}$ be an estimator of ψ_0. We assume that $\hat{\Psi}(P_0) = \psi_0$. Given a random split, B_n, we define $\psi_{B_n} = \hat{\Psi}(P^0_{n,B_n})$, which is the estimator applied to the empirical distribution of the observations contained in the training sample, $\{i : B_n(i) = 0\}$.

Let $B^1_n, ..., B^V_n$ be the collection of random splits that define our cross-validation procedure. We will walk through the case of V-fold cross-validation as an example. In the case of V-fold cross-validation, each of the B^v_n encodes a single fold; the v^{th} validation fold is $\{i : B^v_n(i) = 1\}$, and the remaining samples belong to the v^{th} training sample, $\{i : B^v_n(i) = 0\}$. For each B^v_n, we define $\psi_{B^v_n} = \hat{\Psi}(P^0_{n,B^v_n})$, where P^0_{n,B^v_n} is the empirical distribution of the observations contained in the v^{th} training sample. The function $\psi_{B^v_n}$, which is learned from the v^{th} training sample, will be used to generate predicted values for the observations in the v^{th} validation fold. We define n^v_1 and n^v_0 to be the number of positive and negative samples in the v^{th} validation fold. We note that n^v_1 and n^v_0 are random variables that depend on the value of both B^v_n and $\{Y_i : B^v_n(i) = 1\}$. Formally,

$$n^v_1 = \sum_{i=1}^{n} I(Y_i = 1) I(B^v_n(i) = 1)$$
$$n^v_0 = \sum_{i=1}^{n} I(Y_i = 0) I(B^v_n(i) = 1)$$

The AUC for a single validation fold, $\{i : B^v_n(i) = 1\}$, is

$$AUC(P^1_{n,B^v_n}, \psi_{B^v_n}) = \frac{1}{n^v_0 n^v_1} \sum_{i=1}^{n} \sum_{j=1}^{n} I(\psi_{B^v_n}(W_j) > \psi_{B^v_n}(W_i)) I(Y_i = 0, Y_j = 1) I(B^v_n(i) = B^v_n(j) = 1).$$

Then the V-fold cross-validated AUC estimator is defined as

$$\begin{aligned} E_{B_n} AUC(P^1_{n,B_n}, \psi_{B_n}) &= \frac{1}{V} \sum_{v=1}^{V} AUC(P^1_{n,B_n}, \psi_{B^v_n}) \\ &= \frac{1}{V} \sum_{v=1}^{V} \left\{ \frac{1}{n^v_0 n^v_1} \sum_{i=1}^{n} \sum_{j=1}^{n} I(\psi_{B^v_n}(W_j) > \psi_{B^v_n}(W_i)) I(Y_i = 0, Y_j = 1) I(B^v_n(i) = B^v_n(j) = 1) \right\}. \end{aligned}$$

The target, ψ_0, of the V-fold cross-validated AUC estimator is defined as

$$\begin{aligned} E_{B_n} AUC(P_0, \psi_{B_n}) &= \frac{1}{V} \sum_{v=1}^{V} AUC(P_0, \psi_{B^v_n}) \\ &= \frac{1}{V} \sum_{v=1}^{V} P_0 \left(\psi_{B^v_n}(W_1) > \psi_{B^v_n}(W_2) \mid Y_1 = 1, Y_2 = 0 \right), \end{aligned}$$

where (W_1, Y_1) and (W_2, Y_2) are i.i.d. samples from P_0.

In other words, our target parameter, the true cross-validated AUC, corresponds to fitting the prediction function on each training set, evaluating its true risk (or true probability of correctly ranking two randomly selected observations, where one is a positive sample and the other a negative sample), and taking the average of these true risks across the validation sets. The target parameter thus describes the true classification performance of a predictor fit using the training data. Our estimator of this quantity is based on fitting the prediction function using observations in each training set, estimating its risk using observations in the corresponding the validation sets, and taking the average of these estimates across validation sets. We now wish to estimate the variance of this estimator, and in particular, to construct confidence intervals.

4 Confidence intervals for the risk of an estimator

In order to construct valid confidence intervals for our cross-validated AUC estimator, we must first establish its asymptotic normality. In this section, we present a general theorem that establishes the asymptotic normality of, and provides the influence curve for, the cross-validated risk of an estimator. This provides a general template for the construction of confidence intervals for the cross-validated risk of an estimator. In the following section, we can then apply these results using AUC as a loss function to derive an influence curve based estimate of the variance of our cross-validated AUC estimator.

Let $O \sim P_0 \in \mathcal{M}$ and let $\Psi : \mathcal{M} \to \mathbf{\Psi}$ be an infinite dimensional target parameter. Let $L(\psi)(O)$ be a loss function such that $\psi_0 = \text{argmin}_\psi P_0 L(\psi)$. Let $\hat{\Psi} : \mathcal{M}_{NP} \to \mathbf{\Psi}$ be an estimator, and $\psi_n = \hat{\Psi}(P_n) \in \mathbf{\Psi}$ is the estimate obtained by applying the estimator to the empirical distribution P_n of the i.i.d. sample $O_1, ..., O_n$. The following theorem establishes asymptotic linearity of the cross-validated risk of an estimator under specific conditions and provides a consistent estimator of the asymptotic variance of this estimator. Once an estimate for asymptotic variance has been derived, we construct a 95% confidence interval for the cross-validated risk estimate.

Theorem 1. *Let $B_n \in \{0,1\}^n$ be a random split and let P^1_{n,B_n} and P^0_{n,B_n} be the empirical distributions of the validation $\{i : B_n(i) = 1\}$ and training sample $\{i : B_n(i) = 0\}$, respectively. We assume that B_n has only a finite number of values uniformly in n, as in V-fold cross-validation. We assume that $p = \sum_i B_n(i)/n$ is bounded away from a $\delta > 0$, with probability 1. Define*

$$\hat{R}(\hat{\Psi}, P_n) = E_{B_n} P^1_{n,B_n} L(\hat{\Psi}(P^0_{n,B_n})),$$

where $P^1_{n,B_n} f \equiv E_{P^1_{n,B_n}} f$.

We also define a target of this cross-validated risk as

$$\tilde{R}(\hat{\Psi}, P_n) = E_{B_n} P_0 L(\hat{\Psi}(P^0_{n,B_n})),$$

where $P_0 f \equiv E_{P_0} f$.

We assume that there exists a $\psi_1 \in \mathbf{\Psi}$ so that $P_0 \left\{ L(\hat{\Psi}(P_n)) - L(\psi_1) \right\}^2$ converges to zero in probability as $n \to \infty$. It is assumed that $\sup_{\psi \in \mathbf{\Psi}} \sup_O |L(\psi)(O)| < \infty$, where the supremum over O is over a support of P_0.

Then,

$$\hat{R}(\hat{\Psi}, P_n) - \tilde{R}(\hat{\Psi}, P_n) = \frac{1}{n} \sum_{i=1}^{n} \{L(\psi_1)(O_i) - P_0 L(\psi_1)\} + o_P(1/\sqrt{n}).$$

In particular, $\sqrt{n}\left(\hat{R}(\hat{\Psi}, P_n) - \tilde{R}(\hat{\Psi}, P_n)\right)$ *converges to a normal distribution with mean zero and variance*

$$\sigma^2 = P_0 \left\{ L(\psi_1)(O_i) - P_0 L(\psi_1) \right\}^2.$$

Thus, one can construct an asymptotically 0.95*-confidence interval for* $\tilde{R}(\hat{\Psi}, P_n)$ *given by*

$$\hat{R}(\hat{\Psi}, P_n) \pm 1.96 \frac{\sigma_n}{\sqrt{n}},$$

where σ_n^2 *is a consistent estimator of* σ^2*.*

A consistent estimator of σ^2 *is obtained as*

$$\sigma_n^2 = E_{B_n} P^1_{n,B_n} \left\{ L\left(\hat{\Psi}(P^0_{n,B_n})\right) - \hat{R}(\hat{\Psi}, P_n) \right\}^2.$$

Proof. First we note that:

$$\begin{aligned} \hat{R}(\hat{\Psi}, P_n) - \tilde{R}(\hat{\Psi}, P_n) &= E_{B_n}\left(P^1_{n,B_n} - P_0\right) L\left(\hat{\Psi}(P^0_{n,B_n})\right) \\ &= E_{B_n}\left(P^1_{n,B_n} - P_0\right) L(\psi_1) + E_{B_n}\left(P^1_{n,B_n} - P_0\right) \left\{ L\left(\hat{\Psi}(P^0_{n,B_n})\right) - L(\psi_1) \right\} \end{aligned}$$

The second term is shown to be $o_P(1/\sqrt{n})$ [van der Laan and Rose, 2011] (see Lemma 27.6 and 27.7) and corresponding technical report [Zheng and van der Laan, 2011], involving the application of empirical process theory [van der Vaart and Wellner, 1996] (see Lemma 2.14.1). The first term equals $\sqrt{n}\left(P_n - P_0\right) L(\psi_1)$. This proves the first statement.

By the same proof as in [van der Laan and Rose, 2011], mentioned above, it follows that $\hat{R}(\hat{\Psi}, P_n)$ converges to $P_0 L(\psi_1)$ as $n \to \infty$ and that $E_{B_n} P^1_{n,B_n} \left\{ L\left(\hat{\Psi}(P^0_{n,B_n})\right) - \hat{R}(\hat{\Psi}, P_n) \right\}^2$ converges to $P_0 \left\{ L(\psi_1) - P_0 L(\psi_1) \right\}^2$, which proves that σ_n^2 is a consistent estimator for σ^2.

□

5 Confidence intervals for the AUC of an estimator

Now we apply the results from the previous section, using AUC as the loss function. We derive the influence curve for the AUC estimator and derive influence curve based confidence intervals for the cross-validated AUC. Then we provide a description of the practical construction of the confidence intervals from an i.i.d. data sample.

We consider the identical scenario, where $O = (W, Y) \sim P_0 \in \mathcal{M}$, where Y is binary, and W represents one or more variables. In a binary classification problem, Y is the outcome and W represents the covariates or predictor variables. In the case where $Y \in \{0, 1\}$, we let $\Psi : \mathcal{M} \to \mathbf{\Psi}$ be an infinite dimensional target parameter that maps W into (0,1). We let $\hat{\Psi} : \mathcal{M}_{NP} \to \mathbf{\Psi}$ be an estimator, and $\psi_n = \hat{\Psi}(P_n) \in \mathbf{\Psi}$ is the estimate obtained by applying the estimator to the empirical distribution P_n of the i.i.d. sample $O_1, ..., O_n$.

In order to derive influence curve based confidence intervals for cross-validated AUC, we must first show that $AUC(P_n, \psi)$ is an asymptotically linear estimator of $AUC(P_0, \psi)$, where $\psi \in \mathbf{\Psi}$. To show this, we must prove that

$$AUC(P_n, \psi) - AUC(P_0, \psi) = (P_n - P_0) IC_{AUC}(P_0, \psi) + o_P(1/\sqrt{n}),$$

where $IC_{AUC}(P_0, \psi)$ is the influence curve for the Area Under the Curve estimator. Then, as in the previous theorem, we use empirical process theory to analyze the cross-validated empirical process

terms as in [Zheng and van der Laan, 2011]. Using the notation that was defined in Section 3, it follows that

$$\begin{aligned}
&E_{B_n} AUC(P^1_{n,B_n}, \hat{\Psi}(P^0_{n,B_n})) - E_{B_n} AUC(P_0, \hat{\Psi}(P^0_{n,B_n})) \\
&\quad = E_{B_n}(P^1_{n,B_n} - P_0) IC_{AUC}(P_0, \hat{\Psi}(P^0_{n,B_n})) + o_P(1/\sqrt{n}) \\
&\quad = E_{B_n}(P^1_{n,B_n} - P_0) IC_{AUC}(P_0, \psi_1) \\
&\qquad + E_{B_n}(P^1_{n,B_n} - P_0) \left\{ IC_{AUC}(P_0, \hat{\Psi}(P^0_{n,B_n})) - IC_{AUC}(P_0, \psi_1) \right\} + o_P(1/\sqrt{n}) \\
&\quad = (P_n - P_0) IC_{AUC}(P_0, \psi_1) + o_P(1/\sqrt{n}),
\end{aligned}$$

where the influence curve is given by

$$\begin{aligned}
IC_{AUC}(P_0, \psi)(O) = &\frac{I(Y=1)}{P_0(Y=1)} P_0\left(\psi(W) < x \mid Y = 0\right)|_{x=\psi(W)} \\
&+ \frac{I(Y=0)}{P_0(Y=0)} P_0\left(\psi(W) > x \mid Y = 1\right)|_{x=\psi(W)} \\
&- \left\{ \frac{I(Y=0)}{P_0(Y=0)} + \frac{I(Y=1)}{P_0(Y=1)} \right\} AUC(P_0, \psi).
\end{aligned}$$

We have shown that $AUC(P_n, \psi)$ is indeed an asymptotically linear estimator of $AUC(P_0, \psi)$.

The following theorem is the analogue to Theorem 1 from the previous section, using AUC as the loss function. We begin by defining the influence curve for AUC, as given above. We define the cross-validated AUC estimator, along with the target of this estimator, true cross-validated AUC. As in Theorem 1, we derive an estimate for the asymptotic variance of cross-validated AUC and construct a 95% confidence interval.

Theorem 2. *Let $AUC(P_0, \psi) = \int_0^1 P_0\left(\psi(W) > c \mid Y = 1\right) P_0\left(\psi(W) = c \mid Y = 0\right) dc$. The efficient influence curve $AUC(P_0, \psi)$ for a nonparametric model for P_0 is given by*

$$\begin{aligned}
IC_{AUC}(P_0, \psi)(O) = &\frac{I(Y=1)}{P_0(Y=1)} P_0\left(\psi(W) < x \mid Y = 0\right)|_{x=\psi(W)} \\
&+ \frac{I(Y=0)}{P_0(Y=0)} P_0\left(\psi(W) > x \mid Y = 1\right)|_{x=\psi(W)} \\
&- \left\{ \frac{I(Y=0)}{P_0(Y=0)} + \frac{I(Y=1)}{P_0(Y=1)} \right\} AUC(P_0, \psi).
\end{aligned}$$

For each ψ, the empirical $AUC(P_n, \psi)$ is asymptotically linear with influence curve $IC_{AUC}(P_0, \psi)$.

Let $B_n \in \{0,1\}^n$ be a random split and let P^1_{n,B_n} and P^0_{n,B_n} be the empirical distributions of the validation $\{i : B_n(i) = 1\}$ and training sample $\{i : B_n(i) = 0\}$, respectively. We assume that B_n has only a finite number of values uniformly in n, as in V-fold cross-validation. We assume that $p = \sum_i B_n(i)/n$ is bounded away from a $\delta > 0$, with probability 1. Define the cross-validated area under the ROC curve as

$$\hat{R}(\hat{\Psi}, P_n) = E_{B_n} AUC(P^1_{n,B_n}, \hat{\Psi}(P^0_{n,B_n})).$$

We also define the target of this cross-validated area under the ROC curve as

$$\tilde{R}(\hat{\Psi}, P_n) = E_{B_n} AUC(P_0, \hat{\Psi}(P^0_{n,B_n})).$$

We assume that there exists a $\psi_1 \in \mathbf{\Psi}$ so that $P_0 \left\{ IC_{AUC}(P_0, \hat{\Psi}(P_n)) - IC_{AUC}(P_0, \psi_1) \right\}^2$ converges to zero in probability as $n \to \infty$. We also assume that $\sup_{\psi \in \mathbf{\Psi}} \sup_O |IC_{AUC}(P_0, \psi)(O)| < \infty$, where the supremum over O is over a support of P_0. Then,

$$\hat{R}(\hat{\Psi}, P_n) - \tilde{R}(\hat{\Psi}, P_n) = \frac{1}{n}\sum_{i=1}^{n} IC_{AUC}(O_i) + o_P(1/\sqrt{n}).$$

In particular, $\sqrt{n}\left(\hat{R}(\hat{\Psi}, P_n) - \tilde{R}(\hat{\Psi}, P_n)\right)$ *converges to a normal distribution with mean zero and variance*

$$\sigma^2 = P_0\left\{IC_{AUC}(P_0, \psi_1)\right\}^2.$$

Thus, one can construct an asymptotically 0.95*-confidence interval for* $\tilde{R}(\hat{\Psi}, P_n)$ *given by*

$$\hat{R}(\hat{\Psi}, P_n) \pm 1.96\frac{\sigma_n}{\sqrt{n}}$$

where σ_n^2 *is a consistent estimator of* σ^2*.*

A consistent estimator of σ^2 *is obtained as*

$$\sigma_n^2 = E_{B_n} P^1_{n,B_n}\left\{IC_{AUC}(P^1_{n,B_n}, \hat{\Psi}(P^0_{n,B_n}))\right\}^2.$$

In the estimate for σ^2, we estimate the unknown conditional probabilities of the influence curve IC_{AUC} with the empirical distribution of the validation sample, so that $P^1_{n,B_n}(\psi(W) > x \mid Y = 0)$ will be consistent at $\psi = \hat{\Psi}(P^0_{n,B_n})$ under no conditions on the estimator $\hat{\Psi}$. This is why we replaced P_0 in $IC_{AUC}(P_0, \psi)$ by the empirical distribution of the validation sample. However, the probabilities $P_0(Y = 1)$ and $P_0(Y = 0)$ can be estimated using the whole sample.

5.1 A practical implementation for i.i.d. data

For further clarity, we provide a description of the practical construction of the confidence intervals from an i.i.d. data set, as implemented in our software package. Consider an i.i.d. sample of size n with a binary outcome Y. For each observation, $O_i = (W_i, Y_i)$, we have a d-dimensional numeric vector W_i and a binary outcome, Y_i. Without loss of generality, let $Y_i \in \{0, 1\}$, for all $i = 1, ..., n$, however, Y can be any ordered two-class variable. In this example, we will use V-fold cross-validation and define the the splits as $B_n^1, ..., B_n^V$, as defined previously. Recall that P^1_{n,B_n^v} and P^0_{n,B_n^v} are the empirical distributions of the v^{th} validation and training sample, respectively and P_n is the empirical distribution of the whole data sample.

As in Section 3, we calculate the V-fold cross-validated AUC estimator as

$$\begin{aligned}
\hat{R}(\hat{\Psi}, P_n) &= E_{B_n} AUC(P^1_{n,B_n}, \psi_{B_n}) \\
&= \frac{1}{V}\sum_{v=1}^{V} AUC(P^1_{n,B_n}, \psi_{B_n^v}) \\
&= \frac{1}{V}\sum_{v=1}^{V}\left\{\frac{1}{n_0^v n_1^v}\sum_{i=1}^{n}\sum_{j=1}^{n} I(\psi_{B_n^v}(W_j) > \psi_{B_n^v}(W_i)) I(Y_i = 0, Y_j = 1) I(B_n^v(i) = B_n^v(j) = 1)\right\}.
\end{aligned}$$

In order to construct influence curve based confidence intervals for our V-fold cross-validated AUC estimator, we estimate the asymptotic variance as:

$$\begin{aligned}
\sigma_n^2 &= E_{B_n} P^1_{n,B_n}\left\{IC_{AUC}(P^1_{n,B_n}, \hat{\Psi}(P^0_{n,B_n}))\right\}^2 \\
&= \frac{1}{V}\sum_{v=1}^{V}\left\{\frac{1}{n}\sum_{i=1}^{n}\left\{IC_{AUC}(P^1_{n,B_n^v}, \hat{\Psi}(P^0_{n,B_n^v}))(O_i)\right\}^2 I(B_n^v(i) = 1)\right\},
\end{aligned}$$

where $\psi_{B_n^v} = \hat{\Psi}(P^0_{n,B_n^v})$, and

$$\begin{aligned} IC_{AUC}(P^1_{n,B_n^v}, \hat{\Psi}(P^0_{n,B_n^v}))(O_i) = & \frac{I(Y_i = 1)}{P_n(Y = 1)} P^1_{n,B_n^v}\left(\psi_{B_n^v}(W) < x \mid Y = 0\right)|_{x=\psi_{B_n^v}(W_i)} \\ & + \frac{I(Y_i = 0)}{P_n(Y = 0)} P^1_{n,B_n^v}\left(\psi_{B_n^v}(W) > x \mid Y = 1\right)|_{x=\psi_{B_n^v}(W_i)} \\ & - \left\{\frac{I(Y_i = 0)}{P_n(Y = 0)} + \frac{I(Y_i = 1)}{P_n(Y = 1)}\right\} AUC(P^1_{n,B_n}, \psi_{B_n^v}). \end{aligned}$$

Despite the density of the notation above, each of the components in the influence curve can be calculated very easily. Fix $v \in \{1, ..., V\}$ and $i \in \{1, ..., n\}$, and we will demonstrate how to calculate the quantity, $IC_{AUC}(P^1_{n,B_n^v}, \hat{\Psi}(P^0_{n,B_n^v}))(O_i)$.

The terms, $P_n(Y = 1) \equiv \frac{1}{n}\sum_{j=1}^n I(Y_j = 1)$ and $P_n(Y = 0) \equiv \frac{1}{n}\sum_{j=1}^n I(Y_j = 0)$, are the proportions of positive and negative samples, respectively, in the empirical distribution.

Let $n_1^v = \sum_{j=1}^n I(Y_j = 1)I(B_n^v(j) = 1)$ be the number of positive samples in the v^{th} validation sample and let $n_0^v = \sum_{j=1}^n I(Y_j = 0)I(B_n^v(j) = 1)$ be the number of negative samples in the v^{th} validation sample. Also, recall that $\psi_{B_n^v}$ is the function learned by the v^{th} training sample, which maps a vector, W, of covariates, to a predicted value, $\psi_{B_n^v}(W) \in (0, 1)$. For a given sample, $O_i = (W_i, Y_i)$, we calculate the predicted value, $\psi_{B_n^v}(W_i)$, and note whether Y_i is labeled as positive ($Y_i = 1$) or negative ($Y_i = 0$). Above, each of the terms in the expression for the influence curve contains an indicator function, conditional on the value of Y_i. Therefore, given the value of Y_i, we need only to evaluate the active part of the expression.

When $Y_i = 1$, we need to evaluate:

$$P^1_{n,B_n^v}\left(\psi_{B_n^v}(W) < x \mid Y = 0\right)|_{x=\psi_{B_n^v}(W_i)} = \frac{1}{n_0^v}\sum_{j=1}^n I(W_j < \psi_{B_n^v}(W_i))I(Y_j = 0)I(B_n^v(j) = 1)$$

This sum counts the number of *negative* samples in the validation sample that have a predicted value *less than* $\psi_{B_n^v}(W_i)$, the predicted value for sample i. Then, we divide by the total number of negative samples in the validation sample.

Similarly, when $Y_i = 0$, we need to evaluate:

$$P^1_{n,B_n^v}\left(\psi_{B_n^v}(W) > x \mid Y = 1\right)|_{x=\psi_{B_n^v}(W_i)} = \frac{1}{n_0^v}\sum_{j=1}^n I(W_j > \psi_{B_n^v}(W_i))I(Y_j = 1)I(B_n^v(j) = 1)$$

This sum counts the number of *positive* samples in the validation sample that have a predicted value *greater than* $\psi_{B_n^v}(W_i)$, the predicted value for sample i. Then, we divide by the total number of positive samples in the validation sample.

The remaining term in the expression for the influence curve is $AUC(P^1_{n,B_n}, \psi_{B_n^v})$ multiplied by inverse probability of $P_n(Y = 1)$ or $P_n(Y = 0)$, depending on the value of the indicator function at Y_i. As shown in Section 3, the value of $AUC(P^1_{n,B_n}, \psi_{B_n^v})$ can be calculated directly as follows:

$$AUC(P^1_{n,B_n^v}, \psi_{B_n^v}) = \frac{1}{n_0^v n_1^v}\sum_{k=1}^n\sum_{j=1}^n I(\psi_{B_n^v}(W_j) > \psi_{B_n^v}(W_k))I(Y_k = 0, Y_j = 1)I(B_n^v(k) = B_n^v(j) = 1).$$

Thus, for fixed $v \in \{1, ..., V\}$ and $i \in \{1, ..., n\}$, we have demonstrated how to calculate the quantity, $IC_{AUC}(P^1_{n,B_n^v}, \hat{\Psi}(P^0_{n,B_n^v}))(O_i)$, from an i.i.d. data set. Then we square this term and sum over i.i.d.

samples, i, and cross-validation folds, v, to get

$$\sigma_n^2 = \frac{1}{V}\sum_{v=1}^{V}\left\{\frac{1}{n}\sum_{i=1}^{n}\left\{IC_{AUC}(P^1_{n,B_n^v}, \hat{\Psi}(P^0_{n,B_n^v}))(O_i)\right\}^2 I(B_n^v(i)=1)\right\},$$

an estimate for the asymptotic variance of $\hat{R}(\hat{\Psi}, P_n)$, our V-fold cross-validated AUC estimator. The target of this estimator is

$$\tilde{R}(\hat{\Psi}, P_n) = E_{B_n} AUC(P_0, \hat{\Psi}(P^0_{n,B_n})) = \frac{1}{V}\sum_{v=1}^{V} AUC(P_0, \hat{\Psi}(P^0_{n,B_n^v})),$$

the true V-fold cross-validated AUC. Then, as in Theorem 2, one can construct an asymptotically 0.95-confidence interval for $\tilde{R}(\hat{\Psi}, P_n)$ as

$$\hat{R}(\hat{\Psi}, P_n) \pm 1.96\frac{\sigma_n}{\sqrt{n}}.$$

6 Generalization to the cross-validated AUC for pooled repeated measures data

Above, we derived a consistent, influence curve based, estimator of the asymptotic variance of cross-validated AUC for the simple setting in which have n i.i.d. observations. Each of these observations, O_i has a predictor variable, W_i, coupled with a binary outcome variable, Y_i, that we wish to predict. Now we consider the common setting in which one has repeated measures for each observation. This data structure arises frequently in medical studies, where each patient is frequently measured at multiple time points. We focus on the case where the order of these measures is not meaningful, and one simply wishes to obtain a single summary of classifier performance pooled over all measures. We begin by providing a formal definition of the target parameter, the pooled cross-validated AUC, for such cases. We then extend the results presented in the previous sections to derive an influence curve based variance estimator for the cross-validated AUC of a pooled repeated measures data set.

As before, we let $P_0 \in \mathcal{M}$ and $\Psi : \mathcal{M} \to \mathbf{\Psi}$. We denote the target parameter as $\psi_0 = \Psi(P_0)$. Let $O = (W(t), Y(t) : t \in \tau) \sim P_0$ for a possibly random index set $\tau \subset \{1, ..., T\}$. Here $Y(t)$ is binary for each t. We observe n i.i.d. copies $O_i = (W_i(t), Y_i(t) : t \in \tau_i)$, $i = 1, ..., n$ of O. Let $\mathcal{M}_{NP}$ denote a nonparametric model that includes the empirical distribution, P_n, of $O_1, ..., O_n$ and let $\hat{\Psi} : \mathcal{M}_{NP} \to \mathbb{R}$ be an estimator of ψ_0. We assume that $\hat{\Psi}(P_0) = \psi_0$. We consider the case where t is not a meaningful index, and that either $\psi_0(t, w) = E_0\left(Y(t) \mid W(t) = w\right)$ does not depend on t, or that the investigator has no interest in understanding the dependence on t.

Consider the distribution,

$$\bar{P}_0(w, y) = \frac{1}{E_0|\tau|}\sum_{t=1}^{T} P_0(t \in \tau) P_0\left(W(t) = w, Y(t) = y \mid t \in \tau\right).$$

This represents the limit distribution of the empirical distribution $\bar{P}_n$ of the pooled sample:

$$\bar{P}_n(w, y) = \frac{1}{\sum_{i=1}^{n}|\tau_i|}\sum_{i=1}^{n}\sum_{t\in\tau_i} I\left(W_i(t) = w, Y_i(t) = y\right).$$

One could define as a measure of interest for evaluation a predictor ψ, the area under the ROC curve one would obtain if one treats the pooled sample as N i.i.d. observations. That is, we define

$$\overline{AUC}(\bar{P}_0, \psi) = \int_0^1 \bar{P}_0\left(\psi(W) > c \mid Y = 1\right) \bar{P}_0\left(\psi(W) = c \mid Y = 0\right) dc,$$

where, without loss of generality, we let the positive class be represented by $Y = 1$ and the negative class be represented by $Y = 0$.

The AUC for the empirical distribution of the pooled sample can be expressed explicitly as follows. Let $n_0 = \sum_{i=1}^{n} \sum_{t \in \tau_i} I(Y_i(t) = 0)$ and let $n_1 = \sum_{j=1}^{n} \sum_{s \in \tau_j} I(Y_j(s) = 1)$. Then we have

$$\overline{AUC}(\bar{P}_n, \psi) = \frac{1}{n_0 n_1} \sum_{i=1}^{n} \sum_{t \in \tau_i} \sum_{j=1}^{n} \sum_{s \in \tau_j} I(\psi(W_j(s)) > \psi(W_i(t))) I(Y_i(t) = 0, Y_j(s) = 1).$$

Now we consider the cross-validated AUC of a pooled repeated measures data set. Let $B_n \in \{0,1\}^n$ be a random split and let $\bar{P}^1_{n,B_n}$ and $\bar{P}^0_{n,B_n}$ be the empirical distributions of the pooled data within the validation $\{i : B_n(i) = 1\}$ and training sample $\{i : B_n(i) = 0\}$, respectively. We assume that B_n has only a finite number of values uniformly in n, as in V-fold cross-validation. Given a random split, B_n, we define $\psi_{B_n} = \hat{\Psi}(\bar{P}^0_{n,B_n})$.

As in the i.i.d. example in the previous section, we will walk through the case of V-fold cross-validation. Let $B^1_n, ..., B^V_n$ be the collection of random splits that define our cross-validation procedure. In the case of V-fold cross-validation, each of the B^v_n encodes a single fold; the v^{th} validation fold is $\{i : B^v_n(i) = 1\}$, and the remaining samples belong to the v^{th} training sample, $\{i : B^v_n(i) = 0\}$. Note that since our independent units are collections of pooled time points, $O_i = (W_i(t), Y_i(t) : t \in \tau_i)$, that all pooled samples from each i.i.d. sample, O_i will be contained within the same validation fold.

For each B^v_n, we define $\psi_{B^v_n} = \hat{\Psi}(\bar{P}^0_{n,B^v_n})$, where $\bar{P}^0_{n,B^v_n}$ is the empirical distribution of the pooled data contained in the v^{th} training sample. The function $\psi_{B^v_n}$, which is learned from the v^{th} training sample, will be used to generate predicted values for the observations in the v^{th} validation fold. We define n^v_1 and n^v_0 to be the number of positive and negative samples in the v^{th} validation fold. We note that n^v_1 and n^v_0 are random variables that depend on the value of both B^v_n and $\{Y_i : B^v_n(i) = 1\}$. Formally,

$$n^v_1 = \sum_{i=1}^{n} \sum_{t \in \tau_i} I(Y_i(t) = 1) I(B^v_n(i) = 1)$$

$$n^v_0 = \sum_{i=1}^{n} \sum_{t \in \tau_i} I(Y_i(t) = 0) I(B^v_n(i) = 1)$$

The AUC for a single validation fold, $\{i : B^v_n(i) = 1\}$, for pooled repeated measures data, is

$$\overline{AUC}(\bar{P}^1_{n,B^v_n}, \psi_{B^v_n}) = \frac{1}{n^v_0 n^v_1} \sum_{i=1}^{n} \sum_{t \in \tau_i} \sum_{j=1}^{n} \sum_{s \in \tau_j} I(\psi_{B^v_n}(W_j(s)) > \psi_{B^v_n}(W_i(t))) I(Y_i(t) = 0, Y_j(s) = 1) I(B^v_n(i) = B^v_n(j) = 1).$$

Then the V-fold cross-validated AUC estimator, for pooled repeated measures data, is defined as

$$E_{B_n} \overline{AUC}(\bar{P}^1_{n,B_n}, \psi_{B_n}) = \frac{1}{V} \sum_{v=1}^{V} \overline{AUC}(\bar{P}^1_{n,B^v_n}, \psi_{B^v_n})$$

$$= \frac{1}{V} \sum_{v=1}^{V} \left\{ \frac{1}{n^v_0 n^v_1} \sum_{i=1}^{n} \sum_{t \in \tau_i} \sum_{j=1}^{n} \sum_{s \in \tau_j} I(\psi_{B^v_n}(W_j(s)) > \psi_{B^v_n}(W_i(s))) I(Y_i(t) = 0, Y_j(s) = 1) I(B^v_n(i) = B^v_n(j) = 1) \right\}.$$

We also define the target, ψ_0, of the V-fold cross-validated AUC estimate as

$$E_{B_n}\overline{AUC}(\bar{P}_0,\psi_{B_n}) = \frac{1}{V}\sum_{v=1}^{V}\overline{AUC}(\bar{P}_0,\psi_{B_n^v})$$
$$= \frac{1}{V}\sum_{v=1}^{V}\bar{P}_0\left(\psi_{B_n^v}(W_1) > \psi_{B_n^v}(W_2) \mid Y_1=1, Y_2=0\right),$$

where $(W_1,Y_1)\equiv(W_1(t),Y_1(t))$ and $(W_2,Y_2)\equiv(W_2(t),Y_2(t))$ are single time-point observations.

The following theorem is the pooled repeated measures analogue to Theorem 2, where $O=(W(t),Y(t):t\in\tau)\sim P_0$ for a possibly random index set $\tau\subset\{1,...,T\}$. Below we let $(W,Y)\equiv(W(t),Y(t))$ denote a single time-point observation, for some $t\in\tau$.

Theorem 3. *The efficient influence curve of $\overline{AUC}(\bar{P}_0,\psi)$ for a nonparametric model for P_0 is given by:*

$$IC_{\overline{AUC}}(\bar{P}_0,\psi)(O) = \frac{1}{E_0|\tau|}\sum_{t\in\tau}IC_{AUC}(\bar{P}_0,\psi)(W(t),Y(t)),$$

where

$$IC_{AUC}(\bar{P}_0,\psi)(W,Y) = \frac{I(Y=1)}{\bar{P}_0(Y=1)}\bar{P}_0\left(\psi(W)<x \mid Y=0\right)|_{x=\psi(W)}$$
$$+\frac{I(Y=0)}{\bar{P}_0(Y=0)}\bar{P}_0\left(\psi(W)>x \mid Y=1\right)|_{x=\psi(W)}$$
$$-\left\{\frac{I(Y=0)}{\bar{P}_0(Y=0)}+\frac{I(Y=1)}{\bar{P}_0(Y=1)}\right\}AUC(\bar{P}_0,\psi).$$

For each ψ, the estimator $\overline{AUC}(\bar{P}_n,\psi)$ obtained by plugging in the pooled empirical distribution $\bar{P}_0$ is asymptotically linear with influence curve $IC_{AUC}(\bar{P}_0,\psi)$.

Let $B_n\in\{0,1\}^n$ be a random split and let P^1_{n,B_n} and P^0_{n,B_n} be the empirical distributions of the validation $\{i:B_n(i)=1\}$ and training sample $\{i:B_n(i)=0\}$, respectively. Let $\bar{P}^1_{n,B_n}$ be the empirical distribution of the pooled data within the validation sample. We assume that B_n has only a finite number of values uniformly in n, as in V-fold cross-validation. We assume that $p=\sum_i B_n(i)/n$ is bounded away from a $\delta>0$, with probability 1. Define the cross-validated area under the ROC curve as

$$\hat{R}(\hat{\Psi},P_n) = E_{B_n}\overline{AUC}(\bar{P}^1_{n,B_n},\hat{\Psi}(P^0_{n,B_n})).$$

We also define the target of this cross-validated area under the ROC curve as

$$\tilde{R}(\hat{\Psi},P_n) = E_{B_n}\overline{AUC}(\bar{P}_0,\hat{\Psi}(P^0_{n,B_n})).$$

We assume that there exists a $\psi_1\in\mathbf{\Psi}$ so that $P_0\left\{IC_{AUC}(P_0,\hat{\Psi}(P_n))-IC_{AUC}(P_0,\psi_1)\right\}^2$ converges to zero in probability as $n\to\infty$. We also assume that $\sup_{\psi\in\mathbf{\Psi}}\sup_O|IC_{AUC}(P_0,\psi)(O)|<\infty$, where the supremum over O is over a support of P_0. Then,

$$\hat{R}(\hat{\Psi},P_n)-\tilde{R}(\hat{\Psi},P_n) = \frac{1}{n}\sum_{i=1}^{n}IC_{\overline{AUC}}(\bar{P}_0,\psi_1)(O_i)+o_P(1/\sqrt{n}).$$

In particular, $\sqrt{n}\left(\hat{R}(\hat{\Psi},P_n)-\tilde{R}(\hat{\Psi},P_n)\right)$ converges to a normal distribution with mean zero and variance

$$\sigma^2 = P_0\left\{IC_{\overline{AUC}}(\bar{P}_0,\psi_1)\right\}^2.$$

Thus, one can construct an asymptotically 0.95*-confidence interval for* $\tilde{R}(\hat{\Psi}, P_n)$ *given by*

$$\hat{R}(\hat{\Psi}, P_n) \pm 1.96 \frac{\sigma_n}{\sqrt{n}}$$

where σ_n^2 *is a consistent estimator of* σ^2*.*

A consistent estimator of σ^2 *is obtained as*

$$\sigma_n^2 = E_{B_n} P_{n,B_n}^1 \left\{ IC_{\overline{AUC}}(\bar{P}_{n,B_n}^1, \hat{\Psi}(P_{n,B_n}^0)) \right\}^2.$$

7 Software

We implemented the construction of influence curve based confidence intervals for cross-validated AUC on i.i.d. data as well as pooled repeated measures data, as an R package. The package, called `cvAUC`, depends on functionality from the `ROCR` package [Sing et al., 2005] to calculate the area under the ROC curve.

For each observation, the user provides a predicted value, as generated by a binary prediction algorithm, and a corresponding binary class label. Using the notation above, the user must provide the values $\psi(W)$ and Y for each observation. As in the `ROCR` package, the labels can be supplied as ordered factors as opposed to numeric values, if desired. The user must also indicate which observations belong to each cross-validation split, and there are multiple options for encoding this information. Since V-fold cross-validation is the most commonly used type of cross-validation, we will provide an example below using V-fold cross-validation. To avoid bias in the cross-validated AUC estimate in the pooled repeated measures setting, repeated measures from the independent sampling unit, such as a patient, must all belong to the same validation fold.

The main functions of the package are `ci.cvAUC` and `ci.pooled.cvAUC`, which report cross-validated AUC and calculate corresponding confidence intervals (confidence level supplied by the user) for i.i.d. and pooled repeated measures data. Below is an example of how one might use the package.

7.1 Example using i.i.d. data

The package is designed to be used after predicted values are generated for all observations in each fold. However, we will demonstrate a self-contained example, from start to finish, to provide context. We begin by creating the predicted values and folds object that will be passed as arguments to the `ci.cvAUC` function. The following steps outline the process of generating these data objects.

1. Load a data set with a binary outcome. For the i.i.d. case we use a simulated data set of 500 observations, included with the package, of graduate admissions data. There are five predictor variables and the outcome is admitted vs. not admitted.
2. Divide the indices randomly into 10 folds, stratifying by outcome. Stratification is not necessary, but is commonly performed in order to create validation folds with similar distributions. Store this information in a list called `folds`.
3. Define a function to fit a model on the training data and to generate predicted values for the observations in the validation fold, for a single iteration of the cross-validation procedure. We use a logistic regression fit.
4. Apply this function across all folds to generate predicted values for each validation fold. The concatenated version of these predicted values is stored in vector called `predictions`. The outcome vector, Y, is the `labels` argument.

Once we have created the `predictions`, `labels`, and `folds` objects, we can use the `ci.cvAUC(prediction, labels, folds, confidence=0.95)` function to generate a 10-fold cross-validated AUC estimate with a 95% confidence interval.

R code:

```
iid_example <- function(data, V=10){
  require(cvAUC)
  .cvFolds <- function(Y, V){  #Create CV folds (stratify by outcome)
    Y0 <- split(sample(which(Y==0)), rep(1:V, length=length(which(Y==0))))
    Y1 <- split(sample(which(Y==1)), rep(1:V, length=length(which(Y==1))))
    folds <- vector("list", length=V)
    for (v in seq(V)) {folds[[v]] <- c(Y0[[v]], Y1[[v]])}
    return(folds)
  }
  .doFit <- function(v, folds, data){  #Train/test glm for each fold
    fit <- glm(Y~., data=data[-folds[[v]],], family=binomial)
    pred <- predict(fit, newdata=data[folds[[v]],], type="response")
    return(pred)
  }
  folds <- .cvFolds(Y=data$Y, V=V)  #Create folds
  predictions <- unlist(sapply(seq(V), .doFit, folds=folds, data=data))  #CV train/predict
  predictions[unlist(folds)] <- predictions  #Re-order pred values
  # Get CV AUC and confidence interval
  out <- ci.cvAUC(predictions=predictions, labels=data$Y, folds=folds, confidence=0.95)
  return(out)
}


# Load data
library(cvAUC)
data(admissions)

# Get performance
set.seed(1)
out <- iid_example(data=admissions, V=10)
```

The output is given as follows:

```
> out
$cvAUC
[1] 0.9046473

$se
[1] 0.01620238

$ci
[1] 0.8728913 0.9364034

$confidence
[1] 0.95
```

Therefore, we have estimated cross-validated AUC as 0.901 with a 95% confidence interval approximately equal to $[0.873, 0.936]$. The system runtime for the `ci.cvAUC` step in the example above was less than 0.001 seconds on a on a machine with 8GB of RAM. Although this data set is relatively small, these results demonstrate the efficiency of influence curve based variance estimation. More information and code examples, including the example above, can be found in the user manual for the package. The package is available at: http://www.stat.berkeley.edu/ laan/Software/index.html, and will be available on CRAN.

8 Coverage Probability

In this section, we implement a simulation to demonstrate how the coverage probability of our influence curve based confidence intervals is affected by the adaptability of our estimator. The

coverage probability of a confidence interval is the proportion of the time, over repetitions of the identical experiment, that the interval contains the true value of interest. Our true value of interest is true cross-validated AUC. The coverage of influence curve based confidence intervals relies on the normal limit distribution, thus the larger the number of covariates, the larger the sample size required for the normal distribution to provide a good approximation of the true distribution of the estimator. In the simulation below, the number of observations, n, is fixed, however we experiment with an increasing number of covariates, k. As we increase the number of covariates, the number of main terms in our linear model increases, thus making our estimator more adaptive. This can result in overfitting and so coverage will suffer accordingly. The simulation is included as a function within the `cvAUC` package and is flexible, so that the user can specify different parameters from the ones that we use here.

8.1 Simulation

Let k represent the dimension of a multivariate normal distribution. Let μ be a k-dimensional vector of zeros, let ν be a k-dimensional vector of ones, and let Σ be the k-dimensional identity matrix. For each value of k, we generated 100,000 observations from $\mathcal{N}_k(\mu, \Sigma)$, and for each these observations, we let $Y = 0$. We then generated 100,000 observations from $\mathcal{N}_k(\nu, \Sigma)$ and let $Y = 1$ for each these observations. We consider these 200,000 k-dimensional points with binary outcome Y to represent our true data distribution, P_0. We note that our target parameter, true cross-validated AUC, is itself random, but that it represents a true target. We are interested in the confidence interval that contains this random target 95% of the time. The samples were generated using the `mvrnorm` function of the R package, `MASS` [Venables and Ripley, 2002].

To calculate the coverage probability of our influence curve based confidence intervals, we generate 1,000 confidence intervals and report the proportion of times that the confidence interval contained the true CV AUC. For each iteration, we sample $n = 1000$ points from the same distribution as our population data (500 points from $\mathcal{N}_k(\mu, \Sigma)$ and 500 points from $\mathcal{N}_k(\nu, \Sigma)$ to create a binary labeled sample of size $n = 1000$).

We perform 10-fold cross-validation by splitting these n observations into 10 validation folds, stratifying by outcome, Y, as is common with a binary outcome. For each of the 10 validation folds, we define a corresponding training sample, which is the remainder of the observations not contained within the validation sample. As we have mentioned previously, the cross-validation procedure is not required to be V-fold, however it is a common choice in practice and is convenient for demonstration purposes. For each validation fold, we train a logistic regression fit using the observations from the remaining 9 folds. Using the fit model, we then generate predictions for each of the samples in the validation fold and calculate the empirical AUC. We will call this the *fold AUC*. We also calculate the *true AUC* by generating predicted values for all of the 200,000 data points in our population data and calculating the empirical AUC among this distribution.

This process is repeated for each of the 10 validation folds, at which point we average the fold AUCs to get the estimate for cross-validated AUC. We also average the 10 true AUCs to get the true cross-validated AUC. We use the `ci.cvAUC` function from our `cvAUC` R package to calculate a 95% confidence interval for our CV AUC estimate. We note whether the true CV AUC falls within the confidence interval.

For each value of $k \in \{5, 10, 20, 50, 60, 70, 80, 90, 100\}$, this process is repeated 1,000 times to obtain an estimate of the coverage probability of our confidence intervals, indexed by k. The coverage probability is the proportion times that the true CV AUC fell within our confidence interval. For 95% confidence intervals, we expect the coverage probability to be close to 0.95. The coverage probabilities for each of k is shown Figure 1.

The results of the simulation indicate that for a sample size of $n = 1000$, when $k \leq 70$, the influence curve derived confidence intervals achieve close to a 0.95 coverage probability. However, for $k > 70$, we see a reduction in coverage.

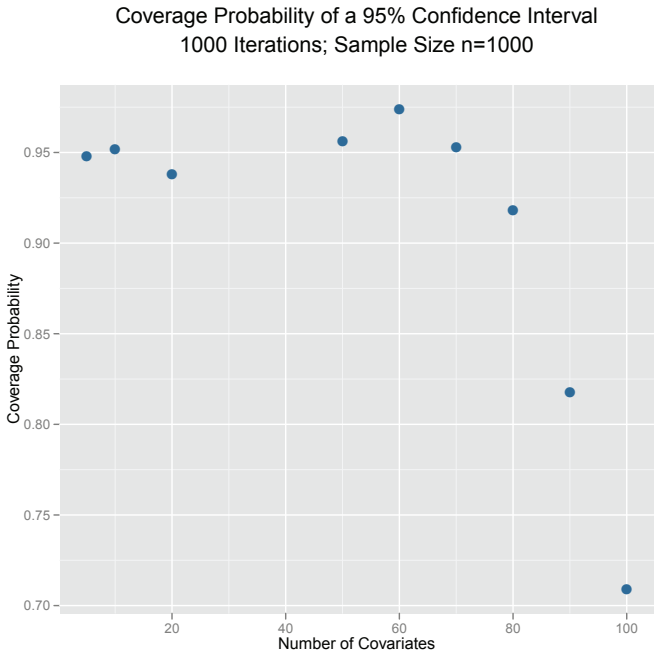

Figure 1: Coverage Probabilities, indexed by k, the number of covariates.

9 Summary

The cross-validated AUC represents an attractive and commonly used measure of performance in binary classification problems. However, resampling based approaches to constructing confidence intervals for this quantity are often computationally infeasible in real data sets. In this paper, we established the asymptotical linearity of the cross-validated AUC estimator and derived its influence curve for both the i.i.d. and pooled repeated measures cases. We then suggested an computationally efficient approach to constructing confidence intervals based on estimating this influence curve. We implemented our approach as a publicly available R package called `cvAUC`. As demonstrated in our simulation, for a fixed sample size n, as the number of variables in the data increases, the adaptability of our estimator increases, which causes overfitting. This results in the coverage probabilities decreasing below the desired coverage rate. Thus, as the number of variables increases, more data is required in order to achieve the desired 0.95 coverage probability for a 95% confidence interval. The simulation showed that for a sample size of $n = 1000$ with 70 or fewer covariates, influence curve based confidence intervals for cross-validated AUC achieve accurate coverage rates. We have demonstrated a computationally efficient alternative to bootstrapping for estimating the variance of cross-validated AUC estimates.

Acknowledgements

Maya Petersen is a recipient of a Doris Duke Clinical Scientist Development Award. This work was supported by the Doris Duke Charitable Foundation Grant number: 2011042. Mark van der Laan is supported by NIH grant number: R01 AI074345. We also want to thank the developers of the `ROCR` R package [Sing et al., 2005] for their contribution to our area under the ROC curve calculations.

References

D. M. Allen. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125–127, 1974.

P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, 1993.

P. J. Bickel, F. Götze, and W. R. Van Zwet. Resampling fewer than n observations: gains, losses, and remedies for losses. *STATIST. SINICA*, 7:1–32, 1997.

A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159, 1997.

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

B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

S. Geisser. The predictive sample reuse method with applications. *Amer. Statist. Assoc.*, 70:320–328, 1975.

R. D. Gill. Non- and semi-parametric maximum likelihood estimators and the von mises method (part 1). *Scandinavian Journal of Statistics*, 16(2):97–128, 1989.

Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I. Jordan. A scalable bootstrap for massive data. *eprint arXiv:1112.5016*, 2011.

C.X. Ling, J. Huang, and H. Zhang. Auc: a statistically consistent and more discriminating measure than accuracy. *Proceedings of IJCAI 2003*, 2003.

D.N. Politis, J.P. Romano, and M. Wolf. Subsampling. *Springer, New York*, 1999.

J. Shao. Linear model selection by cross-validation. *Amer. Statist. Assoc.*, 88(422):486–494, 1993.

Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. ROCR: Visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941, 2005.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147, 1974.

M. J. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Series in Statistics. Springer, first edition, 2011.

A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002.

W. Zheng and M. J. van der Laan. Targeted maximum likelihood estimation of natural direct effect. Technical Report 288, U.C. Berkeley Division of Biostatistics Working Paper Series, 2011.