aucvar: a **R** package for variance estimation of AUC

Francisca Moya Jimenez*
Qing Wang†

Summary

AUC, area under an ROC curve, is one of the most commonly used measures to evaluate the performance of a binary classifier across various discrimination thresholds between 0 and 1 (Bradley 1997). The developed **R** package "aucvar" realizes several variance estimation methods for AUC based on its two-sample U-statistic expression. It complements existing **R** packages that mainly focus on the calculation of AUC or visualization of the ROC curve alone. In particular, it aims at facilitating statistical inference of AUC in practical applications. In addition to computing the unbiased variance estimator of AUC proposed by Wang and Guo (2020), "aucvar" also offers tools that can be used to realize the naive nonparametric bootstrap (Efron 1979) and the delete-d jackknife (Efron and Stein 1981) variance estimators. Moreover, there is a ready-to-use function to construct an asymptotic confidence interval of AUC. As an extension to the AUC-specific application, the package also includes a function that computes the unbiased variance estimator of a general two-sample U-statistic.

Statement of Need

Binary classification is one of the most important problems in statistics and machine learning, where one is interested in building a model or fitting a classification algorithm that correctly predicts whether an experimental unit, such as a patient, belongs to the category of interest or not. Given a binary classification method, there are many existing measures to evaluate its performance. For instance, in medical sciences researchers are often interested in computing sensitivity (i.e. true positive rate) and specificity (i.e. true negative rate). However, these performance metrics depend on a discrimination threshold, a fixed constant in (0,1), that is used to dichotomize the predicted binary outcomes. To account for performance discrepancies due to the use of different discrimination

*Corresponding author
†ORCID:0000-0002-1853-4022

thresholds, the receiver operating characteristic (ROC) curve stands out as a valuable graphical tool that effectively visualizes the trade-off between sensitivity and specificity across all possible discrimination thresholds between 0 and 1. In practice, one often considers the area under a ROC curve, called AUC, as a metric for assessing the overall performance of a binary classifier. Mathematically, AUC represents the probability that a randomly chosen observation with label 1 (i.e. category of interest) is ranked "higher" than a randomly chosen observation with label 0 (i.e. category not of interest). Here, "ranked higher" means having a larger predicted probability of falling into the category of interest in the context of generalized linear regression. In practice, a model with a larger value of AUC is considered better.

Given a real data set and a fitted binary classification method, the unbiased sample estimate of AUC can be written as a two-sample U-statistic, akin to the Mann-Whitney test statistic (Mann and Whitney 1947). When comparing different binary classifiers, a common practice is to select the model with the largest estimated AUC score, possibly computed on a validation data set. However, due to sampling variation, the model with the largest estimated AUC score for a given data set is not necessarily truly optimal. Therefore, it is of great practical interest to perform statistical inference on the true AUC. To do so, one needs to understand the sampling variation of an AUC estimate. With the help of an AUC variance estimator, one can easily conduct statistical inference, such as constructing a confidence interval for the true AUC value, or performing a hypothesis test to compare the true AUC values between two competing classification models. The main purpose of the developed "aucvar" **R** package is to fill this gap and enable the realization of statistical inferential tools for AUC in practice.

Numerical Illustrations

In the following, we highlight and showcase some of the main functionalities of "aucvar" using the breast cancer data set which was originally posted in the UCI Machine Learning Repository (Repository 1995) and is also available in the "aucvar" package. To load the data set, one first needs to require the "aucvar" package, and then can directly call the data set using the name "breastcancer". For simplicity, we omit all data entries with missing values and only consider a complete case data analysis. The primary goal of analyzing this data set is to predict whether or not a patient has breast cancer based on a number of attributes.

```
# library(aucvar)
# mydata <- na.omit(breastcancer)
```

The original data set contains 9 predictor variables and 1 binary categorical response, where the label "malignant" (i.e. level "1") indicates the presence of breast cancer, and the label "benign" (i.e. level "0") suggests otherwise. We use

the full logistic regression model, including all 9 predictors, as an example to explain the functions in "aucvar". After fitting this logistic regression model, one can apply the developed "auc()" function to compute its sample AUC score by simply supplying a vector of the predicted probabilities from the fitted logistic regression model and a vector of the true response labels as shown below.

```
# > full.model <- glm(Class~.,family=binomial(link="logit"),data=mydata)
# > prob <- predict(full.model, type="response")
# > labels <- mydata$Class
# > auc(p_pred=prob, label_true=labels)
# [1] 0.9963248
```

The reported AUC score is computed based on a complete two-sample U-statistic estimator for the true AUC. It is a sample statistic that may suffer from sampling variation. That is, when one fits the model on a different data set sampled from the same population and recomputes the AUC score, the reported value would become different. Therefore, we next illustrate how to evaluate the variability of the AUC estimate.

To estimate the variance of the sample AUC score, one can compute an unbiased variance estimator suggested by Wang and Guo (2020), using the developed "varAUC()" function. The function only requires inputting a vector of the predicted probabilities, a vector of the corresponding true labels, and a parameter B that determines the number of partitions utilized in the efficient realization of the unbiased variance estimator through a partition-resampling scheme. When B is not provided, the exact unbiased variance estimator is realized without the use of the partition-resampling scheme.

```
# > V.unbiased <- varAUC(p_pred=prob, label_true=labels, B=1000)
# > sqrt(V.unbiased)
# [1] 0.001609924
```

In addition, one can also apply the naive nonparametric bootstrap technique (Efron 1979) or the delete-d jackknife method (Efron and Stein 1981) to compute the variance of the sample AUC, using functions "var_boot()" and "var_jack()" respectively. In each function, the input argument B determines the number of random bootstrap (or jackknife) samples used in the calculation. Both functions support generalized linear regression (i.e. a glm object (R Core Team 2022)) as the binary classification model.

```
# > model_formula <- "Class~`Clump Thickness`+`Uniformity of Cell Size`+
# + `Uniformity of Cell Shape`+`Marginal Adhesion`+
# + `Single Epithelial Cell Size`+`Bare Nuclei` +
# + `Bland Chromatin`+`Normal Nucleoli`+`Mitoses`"
# > V.boot <- var_boot(model_formula, label_true=mydata$Class,
# + data=mydata, B = 1000, link="logit")
# > sqrt(V.boot)
# [1] 0.001422026
```

```
# > V.jack <- var_jack(model_formula, label_true=mydata$Class,
# + data=mydata, B = 1000, d = 20, link="logit")
# > sqrt(V.jack)
# [1] 0.001562701
```

With the help of an AUC variance estimator, one can easily perform statistical inference to better understand the true value of AUC. For instance, the "CI_AUC()" function returns an asymptotic confidence interval, using either the unbiased variance estimator (default), the bootstrap variance estimator, or the delete-d jackknife variance estimator of AUC. It is constructed using the asymptotic normality of a general two-sample U-statistic (Lee 1990).

```
## default setting that applies the unbiased variance estimator
# > model_formula <- "Class~`Clump Thickness`+ `Uniformity of Cell Size`+
# + `Uniformity of Cell Shape`+`Marginal Adhesion` +
# + `Single Epithelial Cell Size` + `Bare Nuclei` + `Bland Chromatin`+
#  + `Normal Nucleoli` + `Mitoses`"
# > labels <- my_data$Class
# > CI_AUC(model_formula, mydata, labels, 0.95, B = 1000)
         2.5 %    97.5 %
# [1,] 0.9932559 0.9993937
```

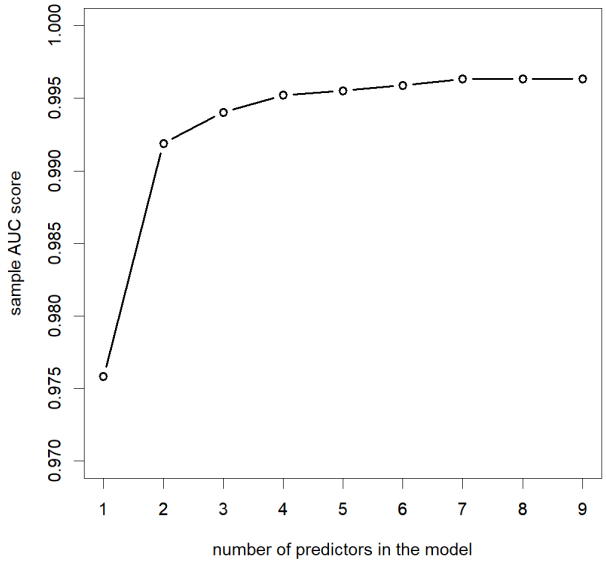

Figure 1: AUC scores of different candidate models.

In practice, one can easily compute the AUC scores of different models and

estimate the variance of each sample AUC score using the functions in "aucvar". Figure 1 displays the sample AUC scores of 9 different models, where each is the optimal model of a given size in the BIC sense (Schwarz 1978) (here the "size" of a model represents its number of predictors).

It is apparent that Model 7 (with 7 predictor variables) has the highest AUC estimate. However, its AUC score is not significantly larger than some of the more parsimonious models. To help identify the model that is truly optimal, one can consider applying the one-standard-error (1-SE) rule (James et al. 2013), selecting the most parsimonious model whose AUC score is within 1 standard error of the largest AUC estimate. Figure 2 presents the implementation of the 1-SE rule, where the standard error was computed based on the unbiased variance estimator using the "varAUC()" function. According to the 1-SE rule, one would consider Model 4, instead of Model 7, as the best model.

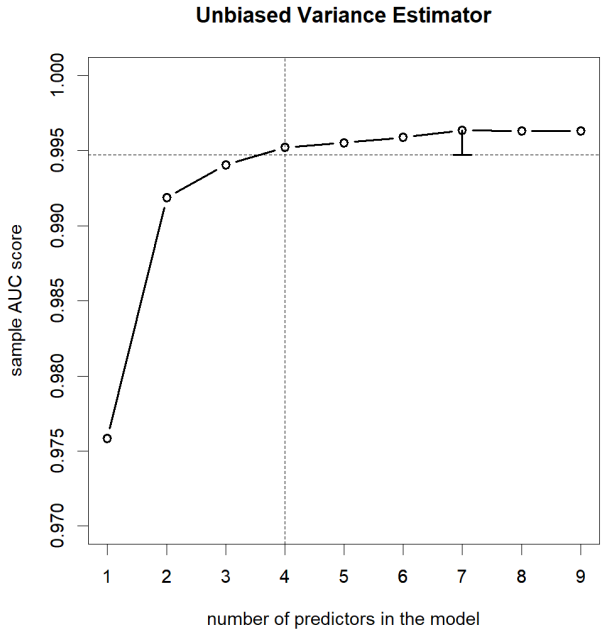

Figure 2: 1-SE model selection rule realized by the unbiased variance estimator of AUC.

Acknowledgements

We want to thank Alexandria Guo for her help during the development of this *R* package. We would also like to acknowledge the Wellesley College research support for this project.

References

Bradley, AP. 1997. "The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms." *Pattern Recognition* 30: 1145–59.

Efron, B. 1979. "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics* 7: 1–26.

Efron, B, and C Stein. 1981. "The Jackknife Estimate of Variance." *The Annals of Statistics* 9: 586–96.

James, G, D Witten, T Hastie, and R Tibshirani. 2013. *An Introduction to Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer.

Lee, AJ. 1990. *U-Statistics: Theory and Practice*. Marcel Dekker.

Mann, HB, and DR Whitney. 1947. "On a Test of Whether One of Two Random Variables Is Stochastically Larger Than the Other." *Annals of Mathematical Statistics* 18: 50–60.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Repository, UCI Machine Learning. 1995. "Breast Cancer Wisconsin (Diagnostic) Data Set." https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic).

Schwarz, GE. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6: 461–64.

Wang, Q, and A Guo. 2020. "An Efficient Variance Estimator of AUC and Its Applications to Binary Classification." *Statistics in Medicine* 39: 4281–4300.