A NOVEL ALTERNATIVE ESTIMATOR OF pAUC TURNS OUT TO BE THE SAME GOOD OLD ONE

Mai Zhou

University of Kentucky, USA

Abstract: We show that a new alternative estimator of pAUC, described by Yang, Lu and Zhao (2017), is in fact the same good old Mann-Whitney estimator, with linear interpolation. Several existing `R` software for computing the estimator are compared.

Key words and phrases: Mann-Whitney estimator, Partial area under AUC curve, Smoothing.

1. Introduction and Conclusion

In the paper of Yang, Lu and Zhao (2017) they studied two different estimators of Partial Area Under ROC Curve (pAUC) given a sample of m X_i's and another sample of n Y_j's : namely the classic Mann-Whitney estimator and a new alternative estimator which they attributed to Wang and Chang (2011). They subsequently used this alternative estimator to obtain the jackknife pseudo-values, which in turn leading to the estimation of the variance, and construction of an empirical likelihood.

On the surface the two estimators look quite different. However, we found that these two estimators of pAUC are actually identical in the sense that the alternative estimator is a linearly interpolated Mann-Whitney estimator:

Main Result *(1) If the partial value P equals to one of those values in the set $\mathcal{A}$ below then the two estimators, (defined in (2.2) and (2.3)), are exactly equal. (2) Suppose P takes a value in between two consecutive values of this set $\mathcal{A}$, i.e. $a_{(j)} < P < a_{(j+1)}$ where $a_{(k)}$ being the ordered values of the set $\mathcal{A}$. Then the Mann-Whitney estimator (as defined in (2.2)) equals to its value at $a_{(j)}$; and the alternative estimator (defined in (2.3)) equals to a linear interpolation between the Mann-Whitney values at $a_{(j)}$ and $a_{(j+1)}$.*

The set $\mathcal{A}$ is defined by

$$\mathcal{A} = \{P;\ for\ \ some\ x \in \mathbb{R}^1,\ P = \frac{1}{n}\sum_{j=1}^{n} I[Y_j > x]\ \}, \tag{1.1}$$

which is just $\{0, 1/n, 2/n, \cdots, 1\}$ if there is no tie in the Y_j sample of n observations. If there are ties in the Y_j observations, then some of the values in the list will be missing.

Smoothing, when sample quantile functions are involved, is a common practice. For example, the `R` function `quantile()` has 9 different options for different smoothing choices. We point out that the classic Mann-Whitney estimator of pAUC, see (2.2), involves sample quantile and is a step function in P_0. Therefore smoothing is often used. Aside from the linear interpolation smoothing, Zhao, Ding and Zhou (2022) used higher order smoothing for the Mann-Whitney estimator of pAUC.

The `R` software package `pROC` Robin et al. (2023) actually also used a linear smoothing for the Mann-Whitney estimator of pAUC, therefore the Mann-Whitney estimator obtained by using function provided by package `pROC` will get you exactly the same result as the so called alternative estimator for *all* P_0. See numeric examples later.

2. Definition, a Lemma and Proof of the Main Result

We shall keep using all the notation of Yang, Lu and Zhao (2017). We recall the definition

$$S_{G,n}(t) = \frac{1}{n}\sum_{j=1}^{n} I[Y_j > t] \ ,$$

and

$$S_{G,n}^{-1}(P_0) = \inf\{x \in \mathbb{R}^1;\ P_0 \geq S_{G,n}(x)\} \ .$$

Using their notation, the definition of the two estimators are

$$\widehat{pAUC}(0,\, P_0) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} I[X_i \geq Y_j] I[Y_j > S_{G,n}^{-1}(P_0)] \quad \text{Mann-Whitney} \tag{2.2}$$

and

$$\widetilde{pAUC}(0,\, P_0) = P_0 - \frac{1}{m}\sum_{i=1}^{m} \min\{S_{G,n}(X_i),\, P_0\} \qquad \text{Alternative.} \tag{2.3}$$

Below we show the two estimators are identical for $P_0 =$ any values in the set $\mathcal{A}$ of (1.1). Let us begin the proof by re-write the alternative estimator:

$$\widetilde{pAUC}(0,\, P_0) = \frac{1}{m}\sum_{i=1}^{m} P_0 - \frac{1}{m}\sum_{i=1}^{m} \min\{S_{G,n}(X_i),\, P_0\}$$

$$= \frac{1}{m}\sum_{i=1}^{m} [P_0 - \min\{S_{G,n}(X_i),\, P_0\}] = \frac{1}{m}\sum_{i:\, S_{G,n}(X_i) \leq P_0} P_0 - S_{G,n}(X_i)$$

$$= \frac{1}{m}\sum_{i:\, S_{G,n}(X_i) \leq P_0} \left\{ P_0 - 1/n \sum_{j=1}^{n} I[X_i < Y_j] \right\} .$$

Next, we want to substitute in the above

$$P_0 = \frac{1}{n}\sum_{j=1}^{n} I[Y_j > S_{G,n}^{-1}(P_0)] = S_{G,n}(S_{G,n}^{-1}(P_0))\,, \tag{2.4}$$

for those P_0 values specified in the set $\mathcal{A}$ of (1.1) in view of the following Lemma:

Lemma This identity (2.4) is not valid for all P_0 values (left hand side is continuous in P_0, right hand side is a step function of P_0) but it is valid when P_0 takes any value in the set $\mathcal{A}$ specified in (1.1).

We shall give a formal proof at the end, but include here a plot that illustrate the validity of our Lemma.

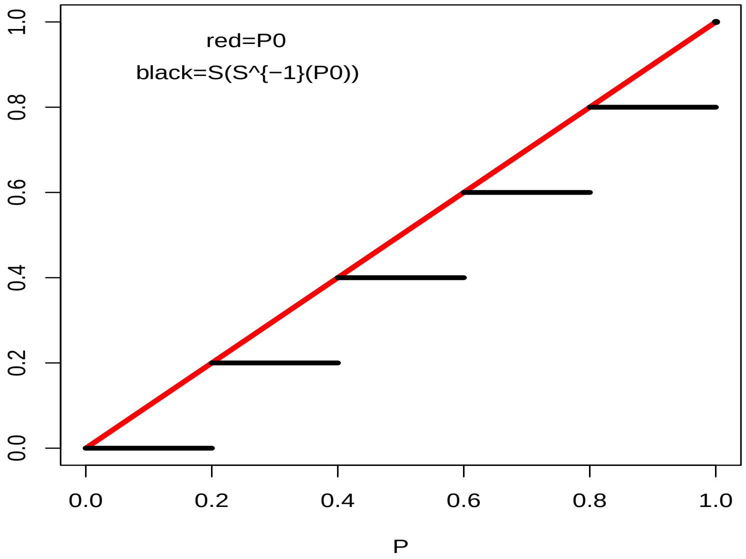

Figure 1: Red equals black at $P = 0, 0.2, 0.4, 0.6, 0.8, 1$.

Therefore, for those P_0 values in the set $\mathcal{A}$, we finally have

$$\widetilde{pAUC}(0,\, P_0) = \frac{1}{m} \sum_{i:\, S_{G,n}(X_i) \le P_0} \left\{ \frac{1}{n} \sum_{j=1}^{n} I[Y_j > S_{G,n}^{-1}(P_0)] - \frac{1}{n} \sum_{j=1}^{n} I[X_i < Y_j] \right\}$$

$$= \frac{1}{mn} \sum_{i:\, S_{G,n}(X_i) \le P_0} \sum_{j=1}^{n} I[Y_j > S_{G,n}^{-1}(P_0)] - I[X_i < Y_j] \ .$$

Next we point out $I[S_{G,n}(X_i) \leq P_0] = I[X_i \geq S_{G,n}^{-1}(P_0)]$ for $i = 1, \cdots, m$ by the definition of the inverse. Thus we can continue

$$\widetilde{pAUC}(0,\, P_0) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} I[X_i \geq S_{G,n}^{-1}(P_0)]\{I[Y_j > S_{G,n}^{-1}(P_0)] - I[X_i < Y_j]\}\,.$$

To finish the proof, we use Vann diagrams to confirm the set relation

$$[X_i \geq S_{G,n}^{-1}(P_0)][Y_j > S_{G,n}^{-1}(P_0)] - [X_i < Y_j][X_i \geq S_{G,n}^{-1}(P_0)] = [X_i \geq Y_j][Y_j > S_{G,n}^{-1}(P_0)]\,.$$

What about the other values of P_0 not specified in the set $\mathcal{A}$? Well, apparently the so called alternate estimator is continuous in P_0 and linear for $a_{(j)} < P_0 < a_{(j+1)}$; while the Mann-Whitney is a step function with jumps at $a_{(j)}$. □

Yang, Lu and Zhao (2017) also studied two similar estimators for pODC. Parallel results also hold there, in that the two estimators are actually the same.

3. Numeric Examples

We confirm the equality of the two estimators, by using numerical examples. Since the existing Mann-Whitney estimators in `R` packages often apply some sort of smoothing, we shall write our own code to strictly follow the definition in section 2 (2.2) of this paper without smoothing. We also write our own code for the “alternative” estimator.

R code for the Mann-Whitney pAUC estimator, no smoothing.

```
 MannW <- function(xvec, yvec, partial) {
    if(partial > 1) stop("partial can not > 1")
    if(partial < 0) stop("partial can not < 0")

    m <- length(xvec)
    n <- length(yvec)

    if(partial == 0 ) Qpartial <- max(yvec)
    if(partial == 1 ) Qpartial <- -Inf
    if((partial <1)&(partial >0)) Qpartial <- QS(yvec, partial)

    iSUMYj <- rep(0, m)
    for (i in 1:m)
    iSUMYj[i] <- sum(as.numeric((xvec[i]>=yvec)&(yvec>Qpartial)))
    FV <- sum( iSUMYj/n )/m
    return(FV)
}

QS <- function(yvec, partial) {
#### Assume 1> partial >0. Omit checking.
n <- length(yvec)
prob <- (0:n)/n
sortedY <- sort(yvec)

Qpartial <- -Inf
posi <- sum(as.numeric(partial > prob))
if( partial == prob[posi+1] ) Qpartial <- sortedY[n-posi]
if((prob[posi] <= partial)&(partial < prob[posi+1]))
                   Qpartial <- sortedY[n+1-posi]
return(Qpartial)
}
```

R code for the alternative pAUC estimator.

```
  ALTernative <- function(xvec, yvec, partial) {
           if(partial > 1) stop("partial can not > 1")
           if(partial < 0) stop("partial can not < 0")

           m <- length(xvec)
```

```
          n <- length(yvec)

          Yecdf <- ecdf(yvec)
          SYecdf <- function(t){1 - Yecdf(t)}

          temp <- pmin( SYecdf(xvec), partial )
          FV <- partial - sum(temp)/m
          return(FV)
}
```

Now we are ready to compute some examples and compare:

```
   set.seed(123)
   Yvec <- rnorm(50); Xvec <- rnorm(50, mean=1)

   ALTernative(xvec=Xvec, yvec=Yvec, partial=0.3)
   ## You get 0.1492
   MannW(xvec=Xvec, yvec=Yvec, partial=0.3)
   ## You also get 0.1492,
   ## since 0.3=15/50, it is a value in the set A of (1.1)
   MannW(xvec=Xvec, yvec=Yvec, partial=0.32)
   ## You get 0.1648.
   ALTernative(xvec=Xvec, yvec=Yvec, partial=0.32)
   ## You also get 0.1648
   #### Since partial=0.32=16/50, both estimators return 0.1648.
```

If P_0 is in between the node points, the two estimators are different.

```
   ALTernative(xvec=Xvec, yvec=Yvec, partial=0.31)
   ## You get  0.157
   MannW(xvec=Xvec, yvec=Yvec, partial=0.31)
   ## You get 0.1492; different from above.
   ## Since 15/50 < 0.31 < 16/50.
```

Verify that the difference in two estimators is due to a linear interpolation:

```
    (0.1492+0.1648)/2
    ##  You get 0.157, same as ALTernative ( ..., partial=0.31)
```

Remark: If we use the R package pROC, then the Mann-Whitney estimator there are apparently already linearly smoothed, which lead to the identical estimator as the 'alternative' estimator for **all** values of P_0.

Remark: As a side note, the option of method="MW" for Mann-Whitney estimator inside the function proc in the R package tpAUC Version 2.1.1 gives a different result as above. There must be a bug. I was unable (email bounce) to contact the maintainer of the package to alert him.

```
library(pROC)
roc(c(rep("Good",50),rep("Poor",50)), c(Yvec,Xvec),
    partial.auc=c(1-0.3, 1), partial.auc.focus="sp",
    progress="none", ci=FALSE)
 ## You get the Mann-Whitney est., exactly same as above: 0.1492
roc(c(rep("Good",50),rep("Poor",50)), c(Yvec,Xvec),
    partial.auc=c(1-0.31, 1), partial.auc.focus="sp",
    progress="none", ci=FALSE)
 ## You get est.=0.157, same as ALTernative(..., partial=0.31).

library(tpAUC)
proc(c(rep("Good",50),rep("Poor",50)), c(Yvec,Xvec),
       threshold=0.3, method="expect")
##You get an alternative est. pAUC(0, 0.3)=0.1492, same as others.
proc(c(rep("Good",50),rep("Poor",50)), c(Yvec,Xvec),
       threshold=0.3, method="MW")
##You get Mann-Whitney est. 0.1648 (something wrong, must be a bug)
```

As a final note, our proof do not invalidate the work of Yang, Lu and Zhao (2017), rather we show that their jackknife method and empirical likelihood approach are really for the (smoothed) classic Mann-Whitney estimator.

References

Yang, H., Lu, K. and Zhao, Y. (2017). A nonparametric approach for partial areas under roc curves and ordinal dominance curves. *Statistica Sinica* **27**, 357–371.

Wang, Z. and Chang, Y. (2011). Marker selection via maximizing the partial area under the ROC curve of linear risk scores. *Biostatistics* **12**, 369–385.

Zhao, Y., Ding, X. and Zhou, M. (2022). Statistical Inference of AUC and pAUC by Empirical Likelihood. Preprint. `https://www.ms.uky.edu/∼mai/research/eAUC1.pdf`

Yang, H., Lu, K., Lyu, X., Hu, F. and Zhao, Y. (2017). tpAUC: Estimation and Inference of Two-Way pAUC, pAUC and pODC. `R` package `tpAUC` version 2.1.1 `https://CRAN.R-project.org/package=tpAUC`

Robin, X. et. al. (2023). Display and Analyze ROC Curves. `R` package `pROC` version 1.18.5 `https://xrobin.github.io/pROC/`

4. Proof of Lemma

It is easy to verify the Lemma when $P_0 = 0$ (in this case $S^{-1}_{G,n}(0) = \max Y_j$, and $S_{G,n}(\max Y_j) = 0$). Also for $P_0 = 1$ (in this case $S^{-1}_{G,n}(1) = -\infty$ and $S_{G,n}(-\infty) = 1$).

Now suppose p' is one of those values as specified in $\mathcal{A}$ as in (1.1) and $0 < p' < 1$. This means

$$p' = \frac{1}{n}\sum_{j=1}^{n} I[Y_j > x'] \quad \text{for some } x'.$$

In fact there are many x that also satisfy the above equality. We write out all those x: for any $x \in [A, I)$, it will satisfy the above equality as well as the x'. (this is just the flat interval of $S_{G,n}(\cdot)$ contain x')

$$A = \max(Y_j;\ s.t.\ Y_j \le x') \qquad \text{and} \qquad I = \min(Y_j;\ s.t.\ Y_j > x')\ .$$

By definition of $S^{-1}_{G,n}(P_0)$, when $P_0 = p'$, is

$$\inf\{x \in \mathbb{R}^1;\ P_0 = p' \ge \frac{1}{n}\sum I[Y_j > x]\}\ .$$

This x set, as we argued, obviously include the interval $[A, I)$. Due to monotonicity of $S_{G,n}(\cdot)$, this x set also includes anything above it but nothing below it. Thus the infimum is seen to equal to A which means $S^{-1}_{G,n}(p') = A$. Finally $S_{G,n}(A) = p'$ since $\#\{Y_j > A\} = \#\{Y_j > x'\}$.

So we have verified $P_0 = S_{G,n}(S^{-1}_{G,n}(P_0))$ for all P_0 values in the set $\mathcal{A}$. □

University of Kentucky, USA

E-mail: (maizhou@gmail.com)