An Alternative Representation of Broad Sense Agreement for Complete Data

Bo Wei, Tian Dai, Limin Peng, PhD, Ying Guo, PhD, and Amita Manatunga, PhD

Technical Report 18-01
September, 2018

Department of Biostatistics and Bioinformatics
Rollins School of Public Health
Emory University
Atlanta, Georgia 30322

Department of Biostatistics, Rollins School of Public Health,
Emory University, 1518 Clifton Road, NE, Atlanta, Georgia 30322 USA

Telephone: (404) 712-9643  FAX: (404) 727-1370
e-mail: bo.wei@emory.edu
Technical report: An Alternative Representation of Broad Sense Agreement for complete data

Bo Wei, Tian Dai, Limin Peng, Ying Guo, Amita Manatunga

Abstract

The concept of broad sense agreement (BSA) has been developed for evaluating the agreement between a continuous measurement and an ordinary measurement (Peng et al., JASA, 2011). They provided a non-parametric estimator of BSA index based on the stratified sampling without replacement. In this work, we propose a new functional representation of BSA index based on the conditional survival function. This new representation form would give a non-parametric estimator via plugging in the empirical conditional cumulative distribution function. We prove the new proposed estimation method with empirical conditional c.d.f. is equivalent to the estimation method in Peng et al. (2011) almost surely. Besides, the new proposed estimation method would gain significant computation efficiency when the sample size or levels of the ordinary measurement increases. Extensive simulation studies are conducted to evaluate the computation efficiency of two methods.

1 Introduction

In Biomedical research, the agreement between a new instrument and a 'gold standard' instrument, or the measurements from different raters is often evaluated by researchers. When two measurements are categorical, the kappa coefficient (Cohen, 1960; Fleiss, 1971; Williamson et al., 2000; Guo and Manatunga, 2009) has been well developed and extended to dependent samples and censored data. For comparing two continuous measurements, the concordance correlation coefficients (CCC) is widely used, and has been developed to accommodate for different data structures, such as replicated samples, repeated measures, survival outcomes and multivariate observations. (Lawrence and Lin, 1989; Lin et al., 2002, 2007; Quiroz, 2005; Guo and Manatunga, 2007; Janson and Olsson, 2001). These well-developed methods could not be used to compare one continuous measurement and one categorical measurement.

Comparing measurements with different scale is common in some studies. For example, the clinician-administered Hamilton Depression Scale (HAM-D) and self-report dimensional scale (Carroll-D) are both used to evaluate the depression in Melanoma and Depression Study (Musselman et al., 2001). In this
study, the Carroll-D is a continuous measurement while the HAM-D is a categorical measurement. To assess the correspondence between HAM-D and Carroll-D, Peng et al. (2011) developed the the a sensible broad sense agreement (BSA) index to evaluate the agreement between a continuous measurement and an ordinal measurement. A non-parametric estimator of BSA index is proposed. The non-parametric estimation procedure has been demonstrated with simulation studies for its stability and utility, and one crucial step in this procedure is based on stratified sampling without replacement. The computation burden of this step increases dramatically when the sample size or the levels of ordinal scale increase.

In this technical report, we propose an alternative representation of BSA index based on conditional cumulative distribution function (or survival function equivalently) for complete data. A non-parametric estimator based on this BSA functional representation is developed via plugging in non-parametric cumulative distribution function estimator. We show that the proposed estimator is coincide with the non-parametric estimator in Peng et al. (2011) when we plug in the empirical c.d.f as the estimator of c.d.f. The simulation studies shows the estimator based on empirical c.d.f gains computation benefits compared to the estimators based on stratified sampling without replacement in Peng et al. (2011).

This technical report is organized as follows. In section 2, a brief review of BSA and the estimation procedure of the BSA index in Peng et al. (2011) are provided, and then a new functional representation of BSA index is proposed. The equivalence between the estimator from functional representation via plugging in empirical cumulative distribution function (c.d.f.) and the non-parametric estimator in Peng et al. (2011) is also shown. In section 3, the results of simulation studies are shown to compare the performance of computation time between the proposed estimator and the estimator from Peng et al. (2011) in different sample sizes and levels.

2 Methods

2.1 A review of BSA framework

We firstly give a review of the broad sense agreement (BSA) concept, the BSA index and its estimation procedure in Peng et al. (2011). Let $X$ and $Y$ denote a continuous scale and an ordinal scale of an outcome from the same object respectively. $D_X$ and $D_Y$ are the corresponding domain of $X$ and $Y$. From the definition of Peng et al. (2011), $X$ and $Y$ are in perfect broad sense agreement (or disagreement) if and only if there exists an increasing (or decreasing) step function $\Psi$ from $D_X$ to $D_Y$ such that $Y = \Psi(X)$ with probability 1. This definition could be interpreted as there exists a set of cut-off points in the continuous $X$ that can produce the discretized $X$, and an exact one-to-one concordance (or discordance) exists between the discretized $X$ and $Y$. From
the concept of BSA, the BSA index, $\rho_{BSA}$ in Peng et al. (2011) is proposed as

$$
\rho_{BSA} = 1 - \frac{E\{\sum_{t=1}^{L}(l - R_t)^2\}}{E\{\sum_{t=1}^{L}(l - R_t)^2 | X \perp\!\!\!\perp Y\}},
$$

where $(R_1, \cdots, R_L)$ is the observed ranks of $\{X_{(x_1)}, \cdots, X_{(x_L)}\}$. The BSA index assesses the extent of departure from the relationship between $X$ and $Y$ to perfect broad sense agreement. The value of $\rho_{BSA}$ is between $-1$ and 1, and a larger value of $\rho_{BSA}$ indicates a better BSA.

By adopting the notation in Peng et al. (2011), suppose the observed data consist of $n$ complete random samples of $(X, Y)$, denoted by $\{X_i, Y_i\}_{i=1}^n$, and they are arranged based on the ordering of $Y$ as follows,

$$(X_1, Y_1 = 1), \cdots, (X_{n_1}, Y_{n_1} = 1), (X_{n_{1}+1}, Y_{n_{1}+1} = 2), \cdots, (X_{n_{1}+n_2}, Y_{n_{1}+n_2} = 2), \cdots,$n_{s} + 1, Y_{n_{s} + 1} = L), \cdots, (X_{n_{1}+\cdots+n_{L}}, Y_{n_{1}+\cdots+n_{L}} = L),$$

where $n_i = \sum_{t=1}^{n} I(Y_t = i)$ and $\sum_{i=1}^{L} n_i = n$. For $1 \leq s \leq L$, further denote $X_{i,s,t} = X_{\sum_{i=1}^{j} n_i + t}$, where $1 \leq t_i \leq n_i$. Peng et al. (2011) showed that

$$
E\{\sum_{i=1}^{L}(l - R_t)^2 | X \perp\!\!\!\perp Y\} = (L^3 - L)/6.
$$

The expected distance $E\{\sum_{i=1}^{L}(l - R_t)^2\}$ is estimated by the sample mean of the mean square distance based on stratified samples without replacement. The non-parametric estimator in Peng et al. (2011) is

$$
\hat{\rho}_{BSA} = 1 - \frac{\left(\prod_{i=1}^{L}\right)^{-1}\sum_{j_1=1}^{n_1} \cdots \sum_{j_L=1}^{n_L} \sum_{t=1}^{l} \left(l - \sum_{r=1}^{l} I(X_{i,j_t} \geq X_{r,j_r})\right)^2}{(L^3 - L)/6}.
$$

(1)

### 2.2 A functional representation of BSA index and the proposed estimator

Suppose $S_x(z) = \Pr(X > z | Y = l)$ denote the conditional survival function of $X$ given $Y = l$. Note that $E\left[\sum_{i=1}^{L}(l - R_t)^2\right] = 2 \sum_{i=1}^{L} l^2 - 2E\left[\sum_{i=1}^{L} \sum_{r=1}^{L} I(\hat{X}_r \leq X_i)\right]$.

By law of total expectation,

$$
E\left[\sum_{i=1}^{L} \sum_{r=1}^{L} I(\hat{X}_r \leq X_i)\right] = \mathbb{E}_{X_1} \left\{ E_{X_2, \cdots, X_L|X_1} \left[ \sum_{i=1}^{L} I(X_1 \leq X_i) + \sum_{r=2}^{L} I(X_r \leq X_1) + \sum_{r=2}^{L} I(X_r \leq X_1) \right] \right\}
$$

$$
= L - \sum_{i=2}^{L} (i - 1) \int_{0}^{\infty} S_l(x_1) dS_l(x_1) + E \left[ \sum_{i=2}^{L} \sum_{r=2}^{L} I(X_r \leq X_i) \right].
$$

(3)
After using the law of total expectation inductively, we have

\[
E \left[ \sum_{l=1}^{L} \sum_{r=1}^{L} I(X_r \leq X_l) \right] = \sum_{l=1}^{L} l(L - l + 1) - \sum_{r=1}^{L-1} \sum_{l=r+1}^{L} (l - r) \int_{0}^{\infty} S_t(x_r) dS_r(x_r)
\]

Based on the above equations, the BSA index could be represented based on the conditional survival function \(\{S_1(\cdot), \cdots, S_L(\cdot)\}\). The representation is

\[
\rho_{bas} = \frac{2 \sum_{r=1}^{L-1} \sum_{l=r+1}^{L} -(l - r) \int_{-\infty}^{\infty} S_t(x) dS_r(x)}{(L^3 - L)/6} - 1. \tag{2}
\]

Suppose \((X, Y)\) are both observed, we could propose an non-parametric estimator via plugging-in the empirical c.d.f

\[
\hat{\rho}_{bas} = \frac{2 \sum_{r=1}^{L-1} \sum_{l=r+1}^{L} -(l - r) \int_{-\infty}^{\infty} \hat{S}_t(x) d\hat{S}_r(x)}{(L^3 - L)/6} - 1 \tag{3}
\]

where \(\hat{S}_t(x) = 1 - \tilde{F}_t(x)\) and \(\hat{F}_t(x) = \frac{1}{n_t} I(X_i \leq x \& Y_i = l)\) is the empirical conditional c.d.f of \(X\) given \(Y = t, t = 1, \cdots, L\).

The proposed non-parametric estimator based on functional representation is equivalent to non-parametric estimator in Peng et al. (2011). Since \(X\) is continuous given \(Y\), no ties exist in \((X, Y)\) almost surely. Without loss of generality, we assume there are no ties among \(\{(X_i, Y_i), i = 1, \cdots, n\}\). Under this assumption, then \(\hat{S}_t(X_{i,t-1}) - \hat{S}_t(X_{i,t}) = 1/n_t\).

To prove the equivalence between these two estimators, it is sufficient to prove the following equation.

\[
(\prod_{l=1}^{L} n_l)^{-1} \sum_{j_1=1}^{n_1} \cdots \sum_{j_L=1}^{n_L} \sum_{l=1}^{L} (l - R_{t,(j_1,\cdots,j_L)})^2 = (L^3 - L)/3 + 2 \sum_{r=1}^{L-1} \sum_{l=r+1}^{L} (l - r) \sum_{j_r=1}^{n_r} \hat{S}_t(X_{r,j_r}) (\hat{S}_r(X_{r,j_r}) - \hat{S}_r(X_{r,j_r-1})) \tag{4}
\]
The left-hand side of Equation (4) can be rewritten as

\[
\left( \prod_{l=1}^{L} n_l \right)^{-1} \sum_{j_1=1}^{n_1} \cdots \sum_{j_L=1}^{n_L} \sum_{l=1}^{L} (l - R_{l, (j_1, \ldots, j_L)})^2
\]

\[
= 2 \left( \prod_{l=1}^{L} n_l \right)^{-1} \sum_{j_1=1}^{n_1} \cdots \sum_{j_L=1}^{n_L} \sum_{l=1}^{L} (l^2 - l) \sum_{r=1}^{L} I(X_{r, j_r} \leq X_{l, j_l})
\]

\[
= 2 \left( \prod_{l=1}^{L} n_l \right)^{-1} \sum_{j_1=1}^{n_1} \cdots \sum_{j_L=1}^{n_L} \sum_{l=1}^{L} (l^2 - \sum_{r=1}^{L} I(X_{r, j_r} \leq X_{l, j_l}))
\]

\[
= 2 \left( \prod_{l=1}^{L} n_l \right)^{-1} \sum_{j_1=1}^{n_1} \cdots \sum_{j_L=1}^{n_L} \left( \sum_{l=1}^{L} (l^2 - \sum_{r=1}^{L} I(X_{r, j_r} \leq X_{l, j_l})) \right)
\]

Equation (5) follows from

\[
\sum_{l=1}^{L} \sum_{r=1}^{L} U(l(X_{r, j_r} \leq X_{l, j_l}))
\]

\[
= \sum_{l=1}^{L} U(X_{l, j_l} \leq X_{l, j_l}) + \sum_{r=2}^{L} I(X_{r, j_r} \leq X_{1, j_1}) + \sum_{l=2}^{L} \sum_{r=2}^{L} U(X_{r, j_r} \leq X_{l, j_l})
\]

\[
= L + (l - 1) \sum_{l=2}^{L} I(X_{l, j_l} \leq X_{l, j_l}) + \sum_{l=2}^{L} \sum_{r=2}^{L} U(X_{r, j_r} \leq X_{l, j_l})
\]

\[
= \sum_{l=1}^{L} l(L - l + 1) + \sum_{r=1}^{L-1} \sum_{l=r+1}^{L} (l - r) I(X_{r, j_r} \leq X_{l, j_l})
\]
Then,

\[
(L^3 - L) / 3 - 2 \left( \prod_{l=1}^{L} n_l \right)^{-1} \sum_{j_1=1}^{n_1} \cdots \sum_{j_L=1}^{n_L} \sum_{l=1}^{L} (l - R_i(j_1, \ldots, j_L))^2
\]

\[
= (L^3 - L) / 3 - 2 \left( \prod_{l=1}^{L} n_l \right)^{-1} \sum_{j_1=1}^{n_1} \cdots \sum_{j_L=1}^{n_L} \sum_{r=1}^{L-1} (l - r) I(X_{r,j_r} \leq X_{l,j_l})
\]

\[
= (L^3 - L) / 3 - 2 \sum_{r=1}^{L-1} \sum_{l=r+1}^{L} (l - r) \frac{1}{n_r} I(X_{r,j_r} \leq X_{l,j_l})
\]

\[
= (L^3 - L) / 3 - 2 \sum_{r=1}^{L-1} \sum_{l=r+1}^{L} (l - r) \frac{1}{n_r} \sum_{j_r=1}^{n_r} \sum_{j_l=1}^{n_l} I(X_{r,j_r} \leq X_{l,j_l})
\]

\[
= (L^3 - L) / 3 - 2 \sum_{r=1}^{L-1} \sum_{l=r+1}^{L} (l - r) \frac{1}{n_r} \sum_{j_r=1}^{n_r} I(X_{r,j_r} \leq X_{l,j_l})
\]

\[
= (L^3 - L) / 3 - 2 \sum_{r=1}^{L-1} \sum_{l=r+1}^{L} (l - r) \frac{1}{n_r} \sum_{j_r=1}^{n_r} \hat{S}_l(X_{r,j_r})
\]

\[
= (L^3 - L) / 3 + 2 \sum_{r=1}^{L-1} \sum_{l=r+1}^{L} (l - r) \sum_{j_r=1}^{n_r} \hat{S}_l(X_{r,j_r})(\hat{S}_l(X_{r,j_r}) - \hat{S}_l(X_{r,j_r-1})�).
\]

Thus, the equivalence of two non-parametric estimators is proved. Besides, since the proposed estimator from functional representation only based on the empirical conditional c.d.f of $F_l(x), l = 1, \ldots, L$ and stratified sampling without replacement is not required, it would be more efficient in computation than the non-parametric estimators calculated from the stratified sampling in Peng et al. (2011).

3 Simulation studies

We conducted Monte Carlo simulations to evaluate the computation efficiency of the proposed estimator calculated from empirical c.d.f and stratified sampling. We compared the computation times between two methods for sample sizes $n = 80, 120$ or 240 and $L = 3$ or 4. The simulated data was generated as follows. For $L$ levels, $Y$ was generated from $\{1, \ldots, L\}$ with equal probability, and the continuous variable $X$ was generated from $Y + \text{Normal}(0, 0.4^2)$. For each simulation, we simulated 400 datasets. The average computation time for each simulated dataset is shown in Table 1. From Table 1, the estimation method based empirical c.d.f. has huge advantages than the estimation method in Peng et al. (2011) in all settings, especially for $L = 4$ and $n = 240$. When
$L = 4$ and $n = 240$, the computation time of estimation method in Peng et al. (2011) is about 9000 times than the proposed estimation method based empirical c.d.f.. The computation time of estimation method in Peng et al. (2011) would increase dramatically when sample size $n$ or $L$ becomes larger, while the computation time of estimation method based on the empirical c.d.f. would have lower increasing rate. In generally, the computation time of the stratified sampling methods in Peng et al. (2011) and the new proposed method are about $O((n/L)^L)$, and $O(nL(L-1))$ respectively.

Table 1: Comparing average computation times (in seconds) for the estimation of DSA index and jackknife standard error in one dataset.

<table>
<thead>
<tr>
<th>$L$</th>
<th>$N$</th>
<th>Peng et al. (2011)'s method</th>
<th>Method based empirical c.d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>80</td>
<td>23.1</td>
<td>1.1</td>
</tr>
<tr>
<td>3</td>
<td>120</td>
<td>113.3</td>
<td>1.8</td>
</tr>
<tr>
<td>3</td>
<td>240</td>
<td>2643.0</td>
<td>4.7</td>
</tr>
<tr>
<td>4</td>
<td>80</td>
<td>168.6</td>
<td>2.0</td>
</tr>
<tr>
<td>4</td>
<td>120</td>
<td>2557.1</td>
<td>2.8</td>
</tr>
<tr>
<td>4</td>
<td>240</td>
<td>63072.0</td>
<td>7.2</td>
</tr>
</tbody>
</table>

4 Conclusion

In this report, we propose a functional representation of broad sense agreement index. Specially, we give a non-parametric estimator based on conditional empirical cumulative distribution function, and we show that the new proposed estimator is equivalent to the non-parametric estimator in Peng et al. (2011) almost surely. Besides, the new proposed method based on the empirical c.d.f. would gain significant efficiency than the method based on the stratified sampling proposed in Peng et al. (2011).

References


