Asymptotic theory of the Concordance Correlation Coefficient CCC

Haidara Mohamed Cheikh
LERSTAD, Université Gaston Berger (Saint-Louis, SENEGAL, chheikhh@yahoo.fr)

Gane Samb Lo
LSTA, Université Pierre et Marie Curie (Paris, France)
LERSTAD, Université Gaston Berger (UGB, Saint-Louis, SENEGAL, gane-samb.lo@ugb.edu.sn)
University of Sciences and Technology (AUST, Abuja, NIGERIA, gslo@aust.edu.ng,
ganesamblo@ganesamblo.net)

Abstract

The Concordance Correlation Coefficient (ccc or c3) is a reproducibility index which was introduced by Lin (1989) for the evaluation of reproducibility, in assessing whether a new method or instrument can reproduce the results from a traditional gold standard approach. Since this measure has been studied in relation with other indices like the Pearson and the kappa one. It has been applied in a number of disciplines (food and grud analysis, biology, etc.) and to different types of data (ordinal, functional, etc.) But, as in the original paper of Lin, most of the results use Monte Carlo experiments to find confidence intervals. In this paper we use the modern functional empirical processes theory to establish the asymptotic normality in multivariate forms and aggregate them to get the final results. A simulation studies address the applicability of the asymptotic results for small sizes.

Keywords: Concordance Correlation coefficient; agreement measures; functional empirical processes; asymptotic normality.

1. Introduction

Let us begin to cite Li and Chow(2005) as an easy introduction to our paper : Evaluation of reproducibility is needed for many scientific research problems. For example, when a new instrument is developed, it is of interest to assess whether the new instrument can reproduce the results obtained by using a traditional gold standard criterion. Indeed, the need to quantify agreement arises in many research fields when two approaches or two raters simultaneously evaluate a response. There are some traditional criteria for measuring agreement between two rating approaches, such as Pearson's correlation coefficient and paired t-test when the responses are continuous. Even though these criteria had been used in practice, they fail to detect poor agreement in some situations (see, for example, Li (1989)). Thus, the topic of assessing agreement for measurements by two approaches has become an interesting research topic. Lin et al. (2002) gives a review and comparison of various measures of recent developments in this area.

The work of Lin LI (1989) is the original point of an intense research activity of the so-called $c3$-index, the concordance Correlation coefficient, defined as follows. Let $(X,Y),(X_i,Y_i), i = 1,2,..$ be e sequence of independent and identically random couples, defined on the same probability space $(\Omega,\mathcal{A},\mathbb{P})$,with bi-variate common cumulative distribution function $G(x,y) = \mathbb{P}(X \geq x, Y \geq y)$, $(x,y) \in \mathbb{R}^2$.

The $c3$-index associated with G is defined by

$$c_3 = \frac{2\rho_{xy}\sigma_x\sigma_y}{1+\sigma_x^2+\sigma_y^2+(\mu_x-\mu_y)^2} = \frac{2\sigma_{xy}}{1+\sigma_x^2+\sigma_y^2+(\mu_x-\mu_y)^2} \tag{1}$$

where

$$\mu_x = \int x dG(x,y), \mu_y = \int y dG(x,y) \tag{2}$$

$$\sigma_x^2 = \int (x-\mu_x)^2 dG(x,y), \sigma_y^2 = \int (y-\mu_y)^2 dG(x,y) \tag{3}$$

$$\sigma_{xy} = \int (x-\mu_x)(y-\mu_y) dG(x,y), \rho_{xy} = \sigma_{xy}/(\sigma_x \sigma_y). \tag{4}$$

The $c3$-index is used now in a variety of disciplines, particularly in Food and Drug Administration (FAD) tests (see Liao(2003) and Liao and Lewis(1999)) and to different types of data (see Li and Chow (2005) for functional data). The present papers aims at providing asymptotic methodologies in using and in applying it.

Finding the exact distribution of $\widehat{c}_3$ or the exact joint distribution of $(\overline{x}-\overline{y}, s_x/s_y, \widehat{\rho}_{xy})$ is the way to deal with statistical estimation of that index. But this may be very difficult in if the data are Gaussian. As an alternative, we suggest to use of the asymptotic approach in parallel with Monte-Carlo methods that are already extensively used. As a first tentative, we may simply consider the empirical plug-in estimators, defined as follows

$$\widehat{c}_3 = \frac{2s_{xy}}{1+s_x^2+s_y^2+(\overline{x}-\overline{y})^2}, \tag{5}$$

where

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} X_i; \overline{y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{6}$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2; s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y})^2 \tag{7}$$

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y}), \widehat{\rho}_{xy} = s_{xy}/(s_x s_y). \tag{8}$$

From there, the modern theory of functional empirical process may help in concluding very quickly. The use of such a tool is explained LO (2016) and Lo et al.(2016) (See Chapter 5).

In the rest of the paper, we make a quick reminder of the Functional Empirical Process (*fep*) tool. In Section 2, we expose our asymptotic results for the $c3$-index and its components. In Section 3, we deal with simulation studies. We finish the paper with concluding remarks.

The *fep* is defined for $f : \mathbb{R} \to \mathbb{R}$ measurable as follows

$$\mathbb{G}_n(f) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} f(X,Y) - \mathbb{E}f(X,Y) \tag{9}$$

Although the *fep* $\{\mathbb{G}_n(f), f \in \mathcal{F}\}$ extensively studied for Donsker's classes $\mathcal{F}$, we do not need here all the armada. Indeed, we only need these two points.

(a) Finite-distribution laws : For any functions f_i, $1 \le i \le k$, satisfying

$$\mathbb{E}\,|f_i(X,Y)| = \int |f_i|\, dG < \infty \tag{10}$$

and

$$\mathbb{E}\left(f_i(X,Y) - Ef_i(X,Y)\right)^2 = \int \left(f_i - \mathbb{E}f_i(X,Y)\right)^2 dG < \infty, \tag{11}$$

$(\mathbb{G}_n(f_1), ..., \mathbb{G}_n(f_k))$ converges to the finite distributions $(\mathbb{G}(f_1), ..., \mathbb{G}(f_k))$ of a functional Gaussian distribution whose variance-covariance function is

$$\sigma^2(f_i, f_j) = \int (f_i - \mathbb{E}f_i(X,Y))(f_j - \mathbb{E}f_j(X,Y))\, dG \tag{12}$$

(b) Linearity : for any functions f_i, $1 \leq i \leq k$, satisfying (10) and for any $a = (a_1, ..., a_k) \in R^k$

$$\mathbb{G}_n(\sum_{1 \leq i \leq k} a_i f_i) = \sum_{1 \leq i \leq k} a_i \mathbb{G}_n(f_i). \tag{13}$$

2. Results

We state the asymptotic law of $\widehat{c}_3$ using the functional Gaussian process whose variance-covariance function is given by 12. We begin by some notation that may seem heavy. But this does not matter so much since the computations are handled by machines.

Notations

Define

$$f_1(x,y) = x; f_2(x,y) = y \tag{14}$$

$$g_1(x,y) = x^2; g_2(x,y) = y^2 \tag{15}$$

$$G_i = g_i - \mathbb{P}_{(X,Y)}(f_i) f_i, i = 1, 2 \tag{16}$$

$$h(x,y) = xy; K = h - \mathbb{P}_{(X,Y)}(f_1) f_2 - \mathbb{P}_{(X,Y)}(f_2) f_1 \tag{17}$$

$$L = 1 + G_1 + G_2 - 2 - (\mathbb{P}_{(X,Y)}(f_1) - \mathbb{P}_{(X,Y)}(f_2))(f_1 - f_2) \tag{18}$$

$$H = \sigma_y^{-4}(\sigma_y^2(g_1 - 2\mu_x f_1) - \sigma_x^2(g_2 - 2\mu_y f_2) \tag{19}$$

$$b = 1 + \sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2; M = b^{-2}(bK - \sigma_{xy} L) \tag{20}$$

Throughout the paper, we must always remind the following notations :

$$\mu_x = \mathbb{P}_{(X,Y)}(f_1), \mu_y = \mathbb{P}_{(X,Y)}(f_2), \tag{21}$$

$$\sigma_x^2 = \mathbb{P}_{(X,Y)}(g_1) - \mathbb{P}^2_{(X,Y)}(f_1), \sigma_y^2 = \mathbb{P}_{(X,Y)}(g_2) - \mathbb{P}^2_{(X,Y)}(f_2) \tag{22}$$

$$\sigma_{xy} = \mathbb{P}_{(X,Y)}(h) - \mathbb{P}_{(X,Y)}(f_1)\mathbb{P}_{(X,Y)}(f_2) \tag{23}$$

Our results are stated in the following lines.

Asymptotic Laws

We have the main following theorem

Theorem 1 *Let $P_{X,Y}(f_1)$, $P_{X,Y}(g_1)$,$P_{X,Y}(f_2)$ and $P_{X,Y}(g_2)$ are finite then we have as $n \to \infty$,*

$$\sqrt{n}(\widehat{c}_3 - c_3) \rightsquigarrow \mathbb{G}(M) \tag{24}$$

and

$$\sqrt{n} \begin{pmatrix} (\overline{x} - \overline{y}) - (\mu_x - \mu_y) \\ s_x/s_y - \sigma_x/\sigma_y \\ \widehat{\rho}_{xy} - \rho_{xy} \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{G}(f_1 - f_2) \\ \mathbb{G}(K) \\ \mathbb{G}(H) \end{pmatrix} \tag{25}$$

Applications. *We are able to use the main theorem to get confidence intervals of c_3 at levels alpha $\in]0,1[$ as follows :*

$$c_3 \in \widehat{c}_3 \pm \frac{z_{1-\alpha/2}}{\sqrt{n}}\sigma_M \tag{26}$$

where

$$\sigma_M^2 = \int (M(x,y) - \mathbb{E}M(x,y))^2 dG(x,y) \tag{27}$$

We are able to use $\widehat{\sigma}_M$ of σ_M as an estimator of of $(\mu_x - \mu_y, \sigma_x/\sigma_y, \rho_{xy})$ at level $\alpha \in]0,1[$. From there, we have two possibilities :

(A) : Using the S-method :

By using the same methods as in (26), we may find for each coordinate an confidence interval of level $\alpha/3$:

$$(\mu_x - \mu_y) \in \overline{x} - \overline{y} \pm \frac{z_{1-\alpha/6}}{\sqrt{n}}\sigma_{f_1-f_2} = I_1 \tag{28}$$

$$\sigma_x/\sigma_y \in s_x/s_y \pm \frac{z_{1-\alpha/6}}{\sqrt{n}}\sigma_K = I_3 \tag{29}$$

$$\rho_{xy} \in \widehat{\rho}_{xy} \pm \frac{z_{1-\alpha/6}}{\sqrt{n}}\sigma_H = I_2 \tag{30}$$

And then, with cover probability of $1-\alpha$, we get

$$(\mu_x - \mu_y, \sigma_x/\sigma_y, \rho_{xy}) \in I_1 \times I_2 \times I_3 \tag{31}$$

(B) Direct method. Put for a, b and c positive real numbers,

$$I_1(a) = \overline{x} - \overline{y} \pm \frac{a}{\sqrt{n}}\sigma_{f_1-f_2} = I_1 \tag{32}$$

$$I_2(b) = s_x/s_y \pm \frac{b}{\sqrt{n}}\sigma_K = I_3 \tag{33}$$

$$I_3(b) = \widehat{\rho}_{xy} \pm \frac{c}{\sqrt{n}}\sigma_H = I_2. \tag{34}$$

We have

$$\mathbb{P}((\mu_x - \mu_y, \sigma_x/\sigma_y, \rho_{xy}) \in I_1(a) \times I_2(b) \times I_3(c)) \tag{35}$$

$$\to \mathbb{P}(|\mathbb{G}(f_1 - f_2)| \leq a, |\mathbb{G}(fK)| \leq b, |\mathbb{G}(H)| \leq c) = p(a,b,c) \tag{36}$$

From there, we have to find a, b and c so that

$$p(a,b,c) = 1 - \alpha. \tag{37}$$

We will consider

$$I_1(a) \times I_2(b) \times I_3(c) \tag{38}$$

as a confidence set of $(\mu_x - \mu_y, \sigma_x/\sigma_y, \rho_{xy})$ at level α. Of course the solution (37) may not be unique.

5. Conclusions

We have been able to find the asymptotic law of the c3-index. Simulations studies are underway to study the applicability of the method for small sizes. Indeed, in some areas, researchers do not and can not have a considerable number of data. Our method will be tested in that context and associated logarithms will be provided..

References

Li R. and Chow. M.(2005). Evaluation of reproducibility for paired functional data. *Journal of Multivariate Analysis*, 93 81101

Lin L.I.(1989). A concordance correlation coefficients to evaluate reproducibility, *Biometrics* 45 (1989) 255268.

Lin L., Hedayat A.S, Sinha B. and Yang M.(2002). Statistical methods in assessing agreement: models, issues and tools, *J. Amer. Statist. Assoc.* 97, 257270.

Lo, G.S., Tchilabalo A. and Ngom, M.(2016). Weak Convergence (IA). Sequences of random vectors. SPAS Books Series.(2016). Doi : 10.16929/sbs/2016.0001. Arxiv : 1610.05415. ISBN : 978-2-9559183-1-9

Lo, G. S.(2016). How to use the functional empirical process for deriving asymptotic laws for functions of the sample. *Arxiv* 1607.02745.

Liao, J.Z. (2003). An improved concordance correlation coefficient. *Pharmaceutical Statistics.* V (2), Issue 4.

Liao J.J.Z. and Lewis J.W.(1999). A Note on Concordance Correlation Coefficient. *PDA Journal of Pharmaceutical Science and Technology.* vol. 54 (1) 23-26.