Sklar's Omega: A Gaussian Copula-Based Framework for Assessing Agreement

John Hughes [*]

Abstract

The statistical measurement of agreement is important in a number of fields, e.g., content analysis, education, computational linguistics, biomedical imaging. We propose Sklar's Omega, a Gaussian copula-based framework for measuring intra-coder, inter-coder, and inter-method agreement as well as agreement relative to a gold standard. We demonstrate the efficacy and advantages of our approach by applying it to both simulated and experimentally observed datasets, including data from two medical imaging studies. Application of our proposed methodology is supported by our open-source R package, `sklarsomega`, which is available for download from the Comprehensive R Archive Network.

Keywords: Agreement coefficient; Composite likelihood; Distributional transform; Gaussian copula

1 Introduction

We develop a model-based alternative to Krippendorff's α (Hayes and Krippendorff, 2007), a well-known nonparametric measure of agreement. In keeping with the naming convention that is evident in the literature on agreement (e.g., Spearman's ρ, Cohen's κ, Scott's π), we call our approach Sklar's ω. Although Krippendorff's α is intuitive, flexible, and subsumes a number of other coefficients of agreement, we will argue that Sklar's ω improves upon α in (at least) the following ways. Sklar's ω

- permits practitioners to simultaneously assess intra-coder agreement, inter-coder agreement, agreement with a gold standard, and, in the context of multiple scoring methods, inter-method agreement;

- identifies the above mentioned types of agreement with intuitive, well-defined population parameters;

- can accommodate any number of coders, any number of methods, any number of replications (per coder and/or per method), and missing values;

---

[*] jphughesjr@gmail.com

- allows practitioners to use regression analysis to reveal important predictors of agreement (e.g., coder experience level, or time effects such as learning and fatigue);

- provides complete inference, i.e., point estimation, interval estimation, diagnostics, model selection; and

- performs more robustly in the presence of unusual coders, units, or scores.

The rest of this article is organized as follows. In Section 2 we present an overview of the agreement problem, and state our assumptions. In Section 3 we present three example applications of both Sklar's ω and Krippendorff's α. These case studies showcase various advantages of our methodology. In Section 4 we specify the flexible, fully parametric statistical model upon which Sklar's ω is based. In Section 5 we describe four approaches to frequentist inference for ω, namely, maximum likelihood, distributional transform approximation, and composite marginal likelihood. We also consider a two-stage semiparametric method that first estimates the marginal distribution nonparametrically and then estimates the copula parameter(s) by conditional maximum likelihood. In Section 6 we use an extensive simulation study to assess the performance of Sklar's ω relative to Krippendorff's α. In Section 7 we briefly describe our open-source R (Ihaka and Gentleman, 1996) package, `sklarsomega`, which is available for download from the Comprehensive R Archive Network (R Core Team, 2018). Finally, in Section 8 we point out potential limitations of our methodology, and posit directions for future research on the statistical measurement of agreement.

2 Measuring agreement

We feel it necessary to define the problem we aim to solve, for the literature on agreement contains two broad classes of methods. Methods in the first class seek to measure agreement while also explaining disagreement—by, for example, assuming differences among coders (as in Aravind and Nawarathna (2017), for example). Although our approach permits one to use regression to explain systematic variation away from a gold standard, we are not, in general, interested in explaining disagreement. Our methodology is for measuring agreement, and so we do not typically accommodate (i.e., model) disagreement. For example, we assume that coders are exchangeable (unless multiple scoring methods are being considered, in which case we assume coder exchangeability within each method). This modeling orientation allows disagreement to count fully against agreement, as desired.

Although our understanding of the agreement problem aligns with that of Krippendorff's α and other related measures, we adopt a subtler interpretation of the results. According to Krippendorff (2012), social scientists often feel justified in relying on data for which agreement is at or above 0.8, drawing tentative conclusions from data for which agreement is at or above 2/3 but less than 0.8, and discarding data for which agreement is less than 2/3. We use the

2

following interpretations instead (Table 1), and suggest—as do Krippendorff and others (Artstein and Poesio, 2008; Landis and Koch, 1977)—that an appropriate reliability threshold may be context dependent.

Range of Agreement	Interpretation
$\omega \leq 0.2$	Slight Agreement
$0.2 < \omega \leq 0.4$	Fair Agreement
$0.4 < \omega \leq 0.6$	Moderate Agreement
$0.6 < \omega \leq 0.8$	Substantial Agreement
$\omega > 0.8$	Near-Perfect Agreement

Table 1: Guidelines for interpreting values of an agreement coefficient.

3 Case studies

In this section we present three case studies that highlight some of the various ways in which Sklar's ω can improve upon Krippendorff's α. The first example involves nominal data, the second example interval data, and the third example ordinal data.

3.1 Nominal data analyzed previously by Krippendorff

Consider the following data, which appear in (Krippendorff, 2013). These are nominal values (in $\{1, \ldots, 5\}$) for twelve units and four coders. The dots represent missing values.

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	u_{11}	u_{12}
c_1	1	2	3	3	2	1	4	1	2	•	•	•
c_2	1	2	3	3	2	2	4	1	2	5	•	3
c_3	•	3	3	3	2	3	4	2	2	5	1	•
c_4	1	2	3	3	2	4	4	1	2	5	1	•

Figure 1: Some example nominal outcomes for twelve units and four coders, with a bit of missingness.

Note that all columns save the sixth are constant or nearly so. This suggests near-perfect agreement, yet a Krippendorff's α analysis of these data leads to a weaker conclusion. Specifically, using the discrete metric $d(x, y) = 1\{x \neq y\}$ yields $\hat{\alpha} = 0.74$ and bootstrap 95% confidence interval (0.39, 1.00). (We used a bootstrap sample size of $n_b = 1{,}000$, which yielded Monte Carlo standard errors (MCSE) (Flegal et al., 2008) smaller than 0.001.) This point estimate

3

indicates merely substantial agreement, and the interval implies that these data are consistent with agreement ranging from moderate to nearly perfect.

Our method produces $\hat{\omega} = 0.89$ and $\omega \in (0.70, 0.98)$ ($n_b = 1{,}000$; MCSEs < 0.004), which indicate near-perfect agreement and at least substantial agreement, respectively. And our approach, being model based, furnishes us with estimated probabilities for the marginal categorical distribution of the response:

$$\hat{\boldsymbol{p}} = (\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4, \hat{p}_5)' = (0.25, 0.24, 0.23, 0.19, 0.09)'.$$

Because we estimated ω and $\boldsymbol{p}$ simultaneously, our estimate of $\boldsymbol{p}$ differs substantially from the empirical probabilities, which are 0.22, 0.32, 0.27, 0.12, and 0.07, respectively.

The marked difference in these results can be attributed largely to the codes for the sixth unit. The relevant influence statistics are

$$\delta_\alpha(\bullet, -6) = \frac{|\hat{\alpha}_{\bullet, -6} - \hat{\alpha}|}{\hat{\alpha}} = 0.15$$

and

$$\delta_\omega(\bullet, -6) = \frac{|\hat{\omega}_{\bullet, -6} - \hat{\omega}|}{\hat{\omega}} = 0.09,$$

where the notation "$\bullet, -6$" indicates that all rows are retained and column 6 is left out. And so we see that column 6 exerts 2/3 more influence on $\hat{\alpha}$ than it does on $\hat{\omega}$. Since $\hat{\alpha}_{\bullet, -6} = 0.85$, inclusion of column 6 draws us away from what seems to be the correct conclusion for these data.

3.2   Interval data from an imaging study of hip cartilage

The data for this example, some of which appear in Figure 2, are 323 pairs of T2* relaxation times (a magnetic resonance quantity) for femoral cartilage (Nissi et al., 2015) in patients with femoroacetabular impingement (Figure 3), a hip condition that can lead to osteoarthritis. One measurement was taken when a contrast agent was present in the tissue, and the other measurement was taken in the absence of the agent. The aim of the study was to determine whether raw and contrast-enhanced T2* measurements agree closely enough to be interchangeable for the purpose of quantitatively assessing cartilage health. The Bland–Altman plot (Altman and Bland, 1983) in Figure 4 suggests good agreement: small bias, no trend, consistent variability.

	u_1	u_2	u_3	u_4	u_5	$\ldots$	u_{319}	u_{320}	u_{321}	u_{322}	u_{323}
c_1	27.3	28.5	29.1	31.2	33.0	$\ldots$	19.7	21.9	17.7	22.0	19.5
c_2	27.8	25.9	19.5	27.8	26.6	$\ldots$	18.3	23.1	18.0	25.7	21.7

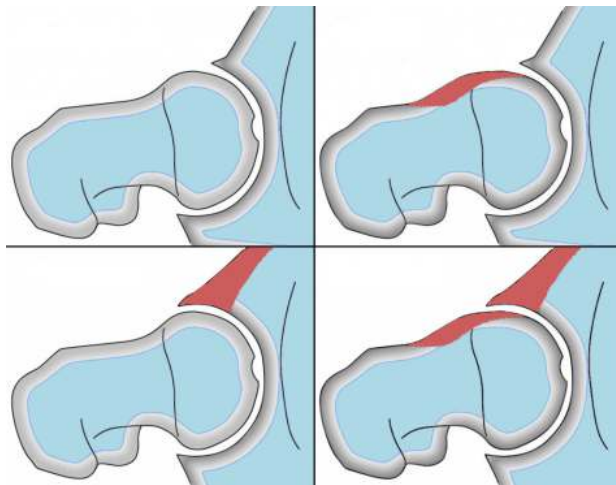Figure 2: Raw and contrast-enhanced T2* values for femoral cartilage.

Figure 3: An illustration of femoroacetabular impingement (FAI). Top left: normal hip joint. Top right: cam type FAI. Bottom left: pincer type FAI. Bottom right: mixed type.

We applied our procedure for each of three choices of parametric marginal distribution: Gaussian, Laplace, and Student's t with noncentrality. The results are shown in Table 2, where the fourth and fifth columns give the estimated location and scale parameters for the three marginal distributions, the sixth column provides values of Akaike's information criterion (AIC) (Akaike, 1974), and the final column shows model probabilities (Burnham et al., 2011) relative to the t model (since that model yielded the smallest value of AIC).

Marginal Model	$\hat{\omega}$	$\omega \in$	Location	Scale	AIC	Model Probability
Gaussian	0.837	(0.803, 0.870)	24.9	5.30	3,605	0.0002
Laplace	0.858	(0.829, 0.886)	24.0	4.34	3,643	≈ 0
t	0.862	(0.833, 0.890)	23.3	11.21	3,588	1
Empirical	0.846	(0.808, 0.869)	—	—	—	—

Table 2: Results from applying Sklar's ω to the femoral-cartilage data.

We see that the estimates are comparable for the three choices of marginal distribution, yet the t distribution is far superior to the others in terms of model probabilities. Figure 5 provides visual corroboration: it is clear that the t assumption proves more appropriate because it is able to capture the mild asymmetry of the marginal distribution. The t assumption also yielded the largest estimate of ω and the narrowest confidence interval (arrived at using the method of maximum likelihood). In any case, we must conclude that there is
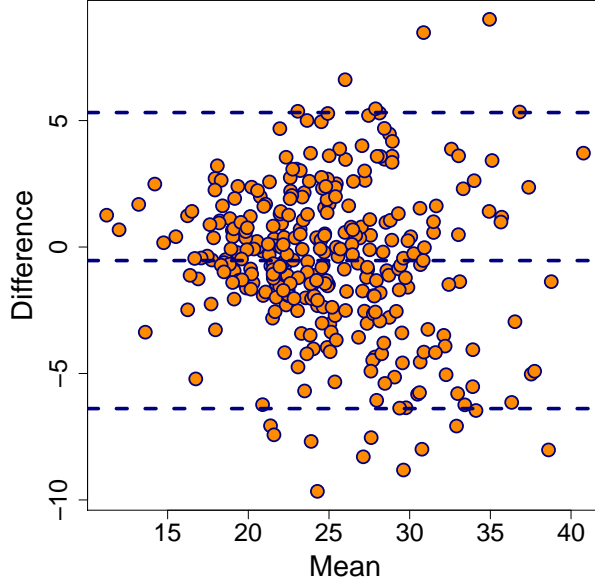
5

Figure 4: A Bland–Altman plot for the femoral cartilage data.

near-perfect agreement between raw T2* and contrast-enhanced T2*.

Note that, for a sufficiently large sample, it may be advantageous to employ a nonparametric estimate of the marginal distribution (see Section 5 for details). Results for this approach are shown in the final row of Table 2. We used a bootstrap sample size of 1,000 in computing the confidence interval. This yielded MCSEs smaller than 0.002.

A Krippendorff's α analysis gave $\hat{\alpha} = 0.837$ and $\alpha \in (0.802, 0.864)$ ($n_b = 1,000$; MSCEs ≈ 0.001). Since α implicitly assumes Gaussianity—α is the intraclass correlation for the ordinary one-way mixed-effects ANOVA model, and so $\hat{\alpha}$ is a ratio of sums of squares (Artstein and Poesio, 2008)—it is not surprising that the α results are quite similar to the results obtained by Sklar's ω with Gaussian marginals.

3.3 Ordinal data from an imaging study of liver herniation

The data for this example, some of which are shown in Figure 6, are liver-herniation scores (in $\{1, \ldots, 5\}$) assigned by two coders (radiologists) to magnetic resonance images (MRI) of the liver in a study pertaining to congenital diaphragmatic hernia (CDH) (Dannull et al., 2017). The five grades are described in Table 3.

Each coder scored each of the 47 images twice, and so we are interested in assessing both intra-coder and inter-coder agreement. We can accomplish both goals with a single analysis by choosing an appropriate form for our copula
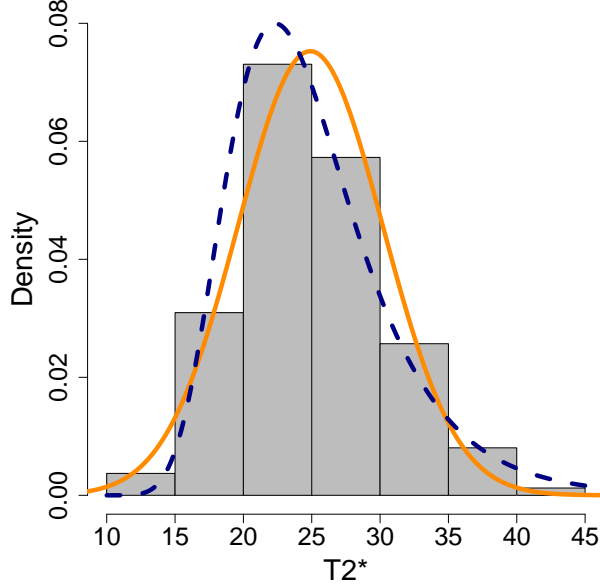
Figure 5: For the T2* data: histogram and fitted Gaussian and t densities. The solid, orange curve is the fitted Gaussian density, and the dashed, blue curve is the fitted t density.

correlation matrix $\boldsymbol{\Omega}$. Specifically, we let $\boldsymbol{\Omega}$ be block diagonal: $\boldsymbol{\Omega} = \text{diag}(\boldsymbol{\Omega}_i)$, where the block for unit i is given by

$$\boldsymbol{\Omega}_i = \begin{array}{c} \\ c_{11} \\ c_{12} \\ c_{21} \\ c_{22} \end{array} \begin{array}{c} c_{11} \quad c_{12} \quad c_{21} \quad c_{22} \\ \begin{pmatrix} 1 & \omega_1 & \omega_{12} & \omega_{12} \\ \omega_1 & 1 & \omega_{12} & \omega_{12} \\ \omega_{12} & \omega_{12} & 1 & \omega_2 \\ \omega_{12} & \omega_{12} & \omega_2 & 1 \end{pmatrix} \end{array},$$

ω_1 being the intra-coder agreement for the first coder, ω_2 being the intra-coder agreement for the second coder, and ω_{12} being the inter-coder agreement. See Section 4 for more information regarding useful correlation structures for agreement.

Our results (from a single analysis), and Krippendorff's α results (three separate analyses), are shown in Table 4. We see that the point estimates for the two approaches are comparable (since the outcomes are approximately Gaussian). Our method produced considerably wider intervals, however. This is not surprising given that Sklar's ω estimates all three correlation parameters simultaneously and accounts for our uncertainty regarding the marginal probabilities. By contrast, Krippendorff's α can yield optimistic inference, as we show by simulation in Section 6.

7

	u_1	u_2	u_3	u_4	u_5	$\ldots$	u_{43}	u_{44}	u_{45}	u_{46}	u_{47}
c_{11}	2	4	4	4	4	$\ldots$	2	1	2	1	1
c_{12}	2	4	5	4	4	$\ldots$	2	1	2	1	1
c_{21}	3	5	5	5	4	$\ldots$	2	2	2	1	1
c_{22}	3	5	5	4	4	$\ldots$	2	2	2	1	1

Figure 6: Ordinal scores for MR images of the liver. Each coder scored each unit twice.

Grade	Description
1	No herniation of liver into the fetal chest
2	Less than half of the ipsilateral thorax is occupied by the fetal liver
3	Greater than half of the thorax is occupied by the fetal liver
4	The liver dome reaches the thoracic apex
5	The liver dome not only reaches the thoracic apex but also extends across the thoracic midline

Table 3: Liver-herniation grades for the CDH study.

4 Our model

The statistical model underpinning Sklar's ω is a Gaussian copula model (Xue-Kun Song, 2000). We begin by specifying the model in full generality. Then we consider special cases of the model that speak to the tasks listed in Section 1 and the assumptions and levels of measurement presented in Section 2.

The stochastic form of the Gaussian copula model is given by

$$
\begin{aligned}
\boldsymbol{Z} = (Z_1, \ldots, Z_n)' &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega}) \\
U_i = \Phi(Z_i) &\sim \mathcal{U}(0,1) \qquad (i = 1, \ldots, n) \\
Y_i = F_i^{-1}(U_i) &\sim F_i,
\end{aligned}
\tag{1}
$$

where $\boldsymbol{\Omega}$ is a correlation matrix, Φ is the standard Gaussian cdf, and F_i is the cdf for the ith outcome Y_i. Note that $\boldsymbol{U} = (U_1, \ldots, U_n)'$ is a realization of the Gaussian copula, which is to say that the U_i are marginally standard uniform and exhibit the Gaussian correlation structure defined by $\boldsymbol{\Omega}$. Since U_i is standard uniform, applying the inverse probability integral transform to U_i produces outcome Y_i having the desired marginal distribution F_i.

In the form of Sklar's ω that most closely resembles Krippendorff's α, we assume that all of the outcomes share the same marginal distribution F. The choice of F is then determined by the level of measurement. While Krippendorff's α typically employs two different metrics for nominal and ordinal out-

Method	Agreement	Interval (MCSEs)
α (Intra-Agreement for Coder 1)	$\hat{\alpha}_1 = 0.987$	$(0.957, 1.000)\ (\approx 0.002)$
α (Intra-Agreement for Coder 2)	$\hat{\alpha}_2 = 0.988$	$(0.958, 1.000)\ (\approx 0.001)$
α (Inter-Agreement)	$\hat{\alpha}_{12} = 0.956$	$(0.917, 0.980)\ (\leq 0.002)$
ω (Intra-Agreement for Coder 1)	$\hat{\omega}_1 = 0.987$	$(0.897, 0.995)\ (\leq 0.002)$
ω (Intra-Agreement for Coder 2)	$\hat{\omega}_2 = 0.991$	$(0.835, 0.994)\ (\leq 0.005)$
ω (Inter-Agreement)	$\hat{\omega}_{12} = 0.965$	$(0.838, 0.996)\ (\leq 0.006)$

Table 4: Results from applying Krippendorff's α and Sklar's ω to the liver scores. The intervals are 95% bootstrap intervals, where the bootstrap sample size was $n_b = 1,000$.

comes, we assume the categorical distribution

$$p_k = \mathbb{P}(Y = k) \qquad (k = 1, \ldots, K)$$
$$\sum_k p_k = 1 \qquad\qquad (2)$$

for both levels of measurement, where K is the number of categories. For $K = 2$, (2) is of course the Bernoulli distribution.

Note that when the marginal distributions are discrete (in our case, categorical), the joint distribution corresponding to (1) is uniquely defined only on the support of the marginals, and the dependence between a pair of random variables depends on the marginal distributions as well as on the copula. Genest and Neslehova (2007) described the implications of this and warned that, for discrete data, "modeling and interpreting dependence through copulas is subject to caution." But Genest and Neslehova go on to say that copula parameters may still be interpreted as dependence parameters, and estimation of copula parameters is often possible using fully parametric methods. It is precisely such methods that we recommend in Section 5, and evaluate through simulation in Section 6.

For interval outcomes F can be practically any continuous distribution. Our R package supports the Gaussian, Laplace, Student's t, and gamma distributions. The Laplace and t distributions are useful for accommodating heavier-than-Gaussian tails, and the t and gamma distributions can accommodate asymmetry. Perhaps the reader can envision more "exotic" possibilities such as mixture distributions (to handle multimodality or excess zeros, for example).

Another possibility for continuous outcomes is to first estimate F nonparametrically, and then estimate the copula parameters in a second stage. In Section 5 we will provide details regarding this approach.

Finally, the natural choice for ratio outcomes is the beta distribution, the two-parameter version of which is supported by package `sklarsomega`.

Now we turn to the copula correlation matrix $\boldsymbol{\Omega}$, the form of which is de-

termined by the question(s) we seek to answer. If we wish to measure only inter-coder agreement, as is the case for Krippendorff's α, our copula correlation matrix has a very simple structure: block diagonal, where the ith block corresponds to the ith unit $(i = 1, \ldots, n_u)$ and has a compound symmetry structure. That is,

$$\boldsymbol{\Omega} = \mathrm{diag}(\boldsymbol{\Omega}_i),$$

where

$$\boldsymbol{\Omega}_i = \begin{array}{c} \\ c_1 \\ c_2 \\ \vdots \\ c_{n_c} \end{array} \begin{array}{cccc} c_1 & c_2 & \ldots & c_{n_c} \\ \begin{pmatrix} 1 & \omega & \ldots & \omega \\ \omega & 1 & \ldots & \omega \\ \vdots & \vdots & \ddots & \vdots \\ \omega & \omega & \ldots & 1 \end{pmatrix} \end{array}$$

On the scale of the outcomes, ω's interpretation depends on the marginal distribution. If the outcomes are Gaussian, ω is the Pearson correlation between Y_{ij} and $Y_{ij'}$, and so the outcomes carry exactly the correlation structure codified in $\boldsymbol{\Omega}$. If the outcomes are non-Gaussian, the interpretation of ω (still on the scale of the outcomes) is more complicated. For example, if the outcomes are Bernoulli, ω is often called the tetrachoric correlation between those outcomes. Tetrachoric correlation is constrained by the marginal distributions. Specifically, the maximum correlation for two binary random variables is

$$\min\left\{ \sqrt{\frac{p_1(1 - p_2)}{p_2(1 - p_1)}}, \sqrt{\frac{p_2(1 - p_1)}{p_1(1 - p_2)}} \right\},$$

where p_1 and p_2 are the expectations (Prentice, 1988). More generally, the marginal distributions impose bounds, called the Fréchet–Hoeffding bounds, on the achievable correlation (Nelsen, 2006). For most scenarios, the Fréchet–Hoeffding bounds do not pose a problem for Sklar's ω because we typically assume that our outcomes are identically distributed, in which case the bounds are -1 and 1. (We do, however, impose our own lower bound of 0 on ω since we aim to measure agreement.)

In any case, ω has a uniform and intuitive interpretation for suitably transformed outcomes, irrespective of the marginal distribution. Specifically,

$$\omega = \rho\left[\Phi^{-1}\{F(Y_{ij})\}, \ \Phi^{-1}\{F(Y_{ij'})\} \right],$$

where ρ denotes Pearson's correlation and the second subscripts index the scores within the ith unit $(j, j' \in \{1, \ldots, n_c\} : j \neq j')$.

By changing the structure of the blocks $\boldsymbol{\Omega}_i$ we can use Sklar's ω to measure not only inter-coder agreement but also a number of other types of agreement. For example, should we wish to measure agreement with a gold standard, we

might employ

$$
\boldsymbol{\Omega}_i =
\begin{array}{c@{\,}c}
& \begin{array}{ccccc} g & c_1 & c_2 & \ldots & c_{n_c} \end{array} \\
\begin{array}{c} g \\ c_1 \\ c_2 \\ \vdots \\ c_{n_c} \end{array} &
\left(\begin{array}{ccccc}
1 & \omega_g & \omega_g & \ldots & \omega_g \\
\omega_g & 1 & \omega_c & \ldots & \omega_c \\
\omega_g & \omega_c & 1 & \ldots & \omega_c \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\omega_g & \omega_c & \omega_c & \ldots & 1
\end{array}\right)
\end{array}.
$$

In this scheme ω_g captures agreement with the gold standard, and ω_c captures inter-coder agreement.

In a more elaborate form of this scenario, we could include a regression component in an attempt to identify important predictors of agreement with the gold standard. This could be accomplished by using a cdf to link coder-specific covariates with ω_g. Then the blocks in $\boldsymbol{\Omega}$ might look like

$$
\boldsymbol{\Omega}_i =
\begin{array}{c@{\,}c}
& \begin{array}{ccccc} g & c_1 & c_2 & \ldots & c_{n_c} \end{array} \\
\begin{array}{c} g \\ c_1 \\ c_2 \\ \vdots \\ c_{n_c} \end{array} &
\left(\begin{array}{ccccc}
1 & \omega_{g1} & \omega_{g2} & \ldots & \omega_{gn_c} \\
\omega_{g1} & 1 & \omega_c & \ldots & \omega_c \\
\omega_{g2} & \omega_c & 1 & \ldots & \omega_c \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\omega_{gn_c} & \omega_c & \omega_c & \ldots & 1
\end{array}\right)
\end{array},
$$

where $\omega_{gj} = H(\boldsymbol{x}'_j\boldsymbol{\beta})$, H being a cdf, $\boldsymbol{x}_j$ being a vector of covariates for coder j, and $\boldsymbol{\beta}$ being regression coefficients.

For our final example we consider a complex study involving a gold standard, multiple scoring methods, multiple coders, and multiple scores per coder. In the interest of concision, suppose we have two methods, two coders per method, two scores per coder for each method, and gold standard measurements for the first method. Then $\boldsymbol{\Omega}_i$ is given by

$$
\boldsymbol{\Omega}_i =
\begin{array}{c@{\,}c}
& \begin{array}{ccccccccc} g_1 & c_{111} & c_{112} & c_{121} & c_{122} & c_{211} & c_{212} & c_{221} & c_{222} \end{array} \\
\begin{array}{c} g_1 \\ c_{111} \\ c_{112} \\ c_{121} \\ c_{122} \\ c_{211} \\ c_{212} \\ c_{221} \\ c_{222} \end{array} &
\left(\begin{array}{ccccccccc}
1 & \omega_{g1} & \omega_{g1} & \omega_{g1} & \omega_{g1} & 0 & 0 & 0 & 0 \\
\omega_{g1} & 1 & \omega_{11\bullet} & \omega_{1\bullet\bullet} & \omega_{1\bullet\bullet} & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} \\
\omega_{g1} & \omega_{11\bullet} & 1 & \omega_{1\bullet\bullet} & \omega_{1\bullet\bullet} & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} \\
\omega_{g1} & \omega_{1\bullet\bullet} & \omega_{1\bullet\bullet} & 1 & \omega_{12\bullet} & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} \\
\omega_{g1} & \omega_{1\bullet\bullet} & \omega_{1\bullet\bullet} & \omega_{12\bullet} & 1 & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} \\
0 & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} & 1 & \omega_{21\bullet} & \omega_{2\bullet\bullet} & \omega_{2\bullet\bullet} \\
0 & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} & \omega_{21\bullet} & 1 & \omega_{2\bullet\bullet} & \omega_{2\bullet\bullet} \\
0 & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} & \omega_{2\bullet\bullet} & \omega_{2\bullet\bullet} & 1 & \omega_{22\bullet} \\
0 & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} & \omega_{\bullet\bullet\bullet} & \omega_{2\bullet\bullet} & \omega_{2\bullet\bullet} & \omega_{22\bullet} & 1
\end{array}\right)
\end{array},
$$

where the subscript mcs denotes score s for coder c of method m. Thus ω_{g1} represents agreement with the gold standard for the first method, $\omega_{11\bullet}$ represents intra-coder agreement for the first coder of the first method, $\omega_{12\bullet}$ represents intra-coder agreement for the second coder of the first method, $\omega_{1\bullet\bullet}$ represents

inter-coder agreement for the first method, and so on, with $\omega_{\bullet\bullet}$ representing inter-method agreement.

Note that, for a study involving multiple methods, it may be reasonable to assume a different marginal distribution for each method. In this case, the Fréchet–Hoeffding bounds may be relevant, and, if some marginal distributions are continuous and some are discrete, maximum likelihood inference is infeasible (see the next section for details).

5 Approaches to inference for ω

When the response is continuous, i.e., when the level of measurement is interval or ratio, we recommend maximum likelihood (ML) inference for Sklar's ω. When the marginal distribution is a categorical distribution (for nominal or ordinal level of measurement), maximum likelihood inference is infeasible because the log-likelihood, having $\Theta(2^n)$ terms, is intractable for most datasets. In this case we recommend the distributional transform (DT) approximation or composite marginal likelihood (CML), depending on the number of categories and perhaps the sample size. If the response is binary, composite marginal likelihood is indicated even for large samples since the DT approach tends to perform poorly for binary data. If there are three or four categories, the DT approach may perform at least adequately for larger samples, but we still favor CML for such data. For five or more categories, the DT approach performs well and yields a more accurate estimator than does the CML approach. The DT approach is also more computationally efficient than the CML approach.

5.1 The method of maximum likelihood for Sklar's ω

For correlation matrix $\boldsymbol{\Omega}(\boldsymbol{\omega})$, marginal cdf $F(y \mid \boldsymbol{\psi})$, and marginal pdf $f(y \mid \boldsymbol{\psi})$, the log-likelihood of the parameters $\boldsymbol{\theta} = (\boldsymbol{\omega}', \boldsymbol{\psi}')'$ given observations $\boldsymbol{y}$ is

$$\ell_{\text{ML}}(\boldsymbol{\theta} \mid \boldsymbol{y}) = -\frac{1}{2}\log|\boldsymbol{\Omega}| - \frac{1}{2}\boldsymbol{z}'(\boldsymbol{\Omega}^{-1} - \mathbf{I})\boldsymbol{z} + \sum_i \log f(y_i), \qquad (3)$$

where $z_i = \Phi^{-1}\{F(y_i)\}$ and $\mathbf{I}$ denotes the $n \times n$ identity matrix. We obtain $\hat{\boldsymbol{\theta}}_{\text{ML}}$ by minimizing $-\ell_{\text{ML}}$. For all three approaches to inference—ML, DT, CML—we use the optimization algorithm proposed by Byrd et al. (1995) so that $\boldsymbol{\omega}$, and perhaps some elements of $\boldsymbol{\psi}$, can be appropriately constrained. To estimate an asymptotic confidence ellipsoid we of course use the observed Fisher information matrix, i.e., the estimate of the Hessian matrix at $\hat{\boldsymbol{\theta}}_{\text{ML}}$:

$$\{\boldsymbol{\theta} : (\hat{\boldsymbol{\theta}}_{\text{ML}} - \boldsymbol{\theta})' \hat{\boldsymbol{\mathcal{I}}}_{\text{ML}} (\hat{\boldsymbol{\theta}}_{\text{ML}} - \boldsymbol{\theta}) \leq \chi^2_{1-\alpha,q}\},$$

where $\hat{\boldsymbol{\mathcal{I}}}_{\text{ML}}$ denotes the observed information and $q = \dim(\boldsymbol{\theta})$.

Optimization of ℓ_{ML} is insensitive to the starting value for $\boldsymbol{\omega}$, but it can be important to choose an initial value $\boldsymbol{\psi}_0$ for $\boldsymbol{\psi}$ carefully. For example, if the

assumed marginal family is t, we recommend $\boldsymbol{\psi}_0 = (\mu_0, \nu_0)' = (\mathrm{med}_n, \mathrm{mad}_n)'$ (Serfling and Mazumder, 2009), where μ is the noncentrality parameter, ν is the degrees of freedom, med_n is the sample median, and mad_n is the sample median absolute deviation from the median. For the Gaussian and Laplace distributions we use the sample mean and standard deviation. For the gamma distribution we recommend $\boldsymbol{\psi}_0 = (\alpha_0, \beta_0)'$, where

$$\alpha_0 = \bar{Y}^2/S^2$$
$$\beta_0 = \bar{Y}/S^2,$$

for sample mean $\bar{Y}$ and sample variance S^2. Similarly, we provide initial values

$$\alpha_0 = \bar{Y}\left\{\frac{\bar{Y}(1-\bar{Y})}{S^2} - 1\right\}$$
$$\beta_0 = (1 - \bar{Y})\left\{\frac{\bar{Y}(1-\bar{Y})}{S^2} - 1\right\}$$

when the marginal distribution is beta. Finally, for a categorical distribution we use the empirical probabilities.

5.2   The distributional transform method

When the marginal distribution is discrete (in our case, categorical), the log-likelihood does not have the simple form given above because $z_i = \Phi^{-1}\{F(y_i)\}$ is not standard Gaussian (since $F(y_i)$ is not standard uniform if F has jumps). In this case the true log-likelihood is intractable unless the sample is rather small. An appealing alternative to the true log-likelihood is an approximation based on the distributional transform.

It is well known that if $Y \sim F$ is continuous, $F(Y)$ has a standard uniform distribution. But if Y is discrete, $F(Y)$ tends to be stochastically larger, and $F(Y^-) = \lim_{x \nearrow Y} F(x)$ tends to be stochastically smaller, than a standard uniform random variable. This can be remedied by stochastically "smoothing" F's discontinuities. This technique goes at least as far back as Ferguson (1967), who used it in connection with hypothesis tests. More recently, the distributional transform has been applied in a number of other settings—see, e.g., Rüschendorf (1981), Burgert and Rüschendorf (2006), and Rüschendorf (2009).

Let $W \sim \mathcal{U}(0, 1)$, and suppose that $Y \sim F$ and is independent of W. Then the distributional transform

$$G(W, Y) = WF(Y^-) + (1 - W)F(Y)$$

follows a standard uniform distribution, and $F^{-1}\{G(W, Y)\}$ follows the same distribution as Y.

Kazianka and Pilz (2010) suggested approximating $G(W, Y)$ by replacing it

with its expectation with respect to W:

$$
\begin{aligned}
G(W, Y) &\approx \mathbb{E}_W G(W, Y) \\
&= \mathbb{E}_W \{ W F(Y^-) + (1 - W) F(Y) \} \\
&= \mathbb{E}_W W F(Y^-) + \mathbb{E}_W (1 - W) F(Y) \\
&= F(Y^-) \mathbb{E}_W W + F(Y) \mathbb{E}_W (1 - W) \\
&= \frac{F(Y^-) + F(Y)}{2}.
\end{aligned}
$$

To construct the approximate log-likelihood for Sklar's ω, we replace $F(y_i)$ in (3) with

$$
\frac{F(y_i^-) + F(y_i)}{2}.
$$

If the distribution has integer support, this becomes

$$
\frac{F(y_i - 1) + F(y_i)}{2}.
$$

This approximation is crude, but it performs well as long as the discrete distribution in question has a sufficiently large variance (Kazianka, 2013). For Sklar's ω, we recommend using the DT approach when the scores fall into five or more categories.

Since the DT-based objective function is misspecified, using $\hat{\mathcal{I}}_{\mathrm{DT}}$ alone leads to optimistic inference unless the number of categories is large. This can be overcome by using a sandwich estimator (Godambe, 1960) or by doing a bootstrap (Davison and Hinkley, 1997).

5.3 Composite marginal likelihood

For nominal or ordinal outcomes falling into a small number of categories, we recommend a composite marginal likelihood (Lindsay, 1988; Varin, 2008) approach to inference. Our objective function comprises pairwise likelihoods (which implies the assumption that any two pairs of outcomes are independent). Specifically, we work with log composite likelihood

$$
\ell_{\mathrm{CML}}(\boldsymbol{\theta} \mid \boldsymbol{y}) = \sum_{\substack{i \in \{1, \ldots, n-1\} \\ j \in \{i+1, \ldots, n\} \\ \boldsymbol{\Omega}_{ij} \neq 0}} \log \left\{ \sum_{j_1=0}^{1} \sum_{j_2=0}^{1} (-1)^k \Phi_{\boldsymbol{\Omega}^{ij}}(z_{ij_1}, z_{jj_2}) \right\},
$$

where $k = j_1 + j_2$, $\Phi_{\boldsymbol{\Omega}^{ij}}$ denotes the cdf for the bivariate Gaussian distribution with mean zero and correlation matrix

$$
\boldsymbol{\Omega}^{ij} = \begin{pmatrix} 1 & \boldsymbol{\Omega}_{ij} \\ \boldsymbol{\Omega}_{ij} & 1 \end{pmatrix},
$$

$z_{\bullet 0} = \Phi^{-1}\{F(y_\bullet)\}$, and $z_{\bullet 1} = \Phi^{-1}\{F(y_\bullet - 1)\}$. Since this objective function, too, is misspecified, bootstrapping or sandwich estimation is necessary.

5.4 Sandwich estimation for the DT and CML procedures

As we mentioned above, the DT and CML objective functions are misspecified, and so the asymptotic covariance matrices of $\hat{\boldsymbol{\theta}}_{\mathrm{DT}}$ and $\hat{\boldsymbol{\theta}}_{\mathrm{CML}}$ have sandwich forms (Godambe, 1960; Geyer, 2013). Specifically, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathrm{CML}} - \boldsymbol{\theta}) \;\Rightarrow_n\; \mathcal{N}\{\mathbf{0},\, \boldsymbol{\mathcal{I}}^{-1}_{\mathrm{CML}}(\boldsymbol{\theta})\boldsymbol{\mathcal{J}}_{\mathrm{CML}}(\boldsymbol{\theta})\boldsymbol{\mathcal{I}}^{-1}_{\mathrm{CML}}(\boldsymbol{\theta})\}$$
$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathrm{DT}} - \boldsymbol{\theta}) \;\Rightarrow_n\; \mathcal{N}\{\mathbf{0},\, \boldsymbol{\mathcal{I}}^{-1}_{\mathrm{DT}}(\boldsymbol{\theta})\boldsymbol{\mathcal{J}}_{\mathrm{DT}}(\boldsymbol{\theta})\boldsymbol{\mathcal{I}}^{-1}_{\mathrm{DT}}(\boldsymbol{\theta})\},$$

where $\boldsymbol{\mathcal{I}}_\bullet$ is the appropriate Fisher information matrix and $\boldsymbol{\mathcal{J}}_\bullet$ is the variance of the score:

$$\boldsymbol{\mathcal{J}}_\bullet(\boldsymbol{\theta}) = \mathbb{V}\nabla\ell_\bullet(\boldsymbol{\theta} \mid \boldsymbol{Y}).$$

We recommend that $\boldsymbol{\mathcal{J}}_\bullet$ be estimated using a parametric bootstrap, i.e, our estimator of $\boldsymbol{\mathcal{J}}_\bullet$ is

$$\hat{\boldsymbol{\mathcal{J}}}_\bullet(\boldsymbol{\theta}) = \frac{1}{n_b}\sum_{j=1}^{n_b} \nabla\nabla'\ell_\bullet(\hat{\boldsymbol{\theta}}_\bullet \mid \boldsymbol{Y}^{(j)}),$$

where n_b is the bootstrap sample size and the $\boldsymbol{Y}^{(j)}$ are datasets simulated from our model at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_\bullet$. This approach performs well, as our simulation results show, and is considerably more efficient computationally than a "full" bootstrap (it is much faster to approximate the score than to optimize the objective function). What is more, $\hat{\boldsymbol{\mathcal{J}}}_\bullet(\boldsymbol{\theta})$ is accurate for small bootstrap sample sizes (100 in our simulations). The procedure can be made even more efficient through parallelization.

5.5 A two-stage semiparametric approach for continuous measurements

If the sample size is large enough, a two-stage semiparametric method (SMP) may be used. In the first stage one estimates F nonparametrically. The empirical distribution function $\hat{F}_n(y) = n^{-1}\sum_i 1\{Y_i \leq y\}$ is a natural choice for our estimator of F, but other sensible choices exist. For example, one might employ the Winsorized estimator

$$\tilde{F}_n(y) = \begin{cases} \epsilon_n & \text{if } \hat{F}_n(y) < \epsilon_n \\ \hat{F}_n(y) & \text{if } \epsilon_n \leq \hat{F}_n(y) \leq 1 - \epsilon_n \\ 1 - \epsilon_n & \text{if } \hat{F}_n(y) > 1 - \epsilon_n, \end{cases}$$

where ϵ_n is a truncation parameter (Klaassen et al., 1997; Liu et al., 2009). A third possibility is a smoothed empirical distribution function

$$\breve{F}_n(y) = \frac{1}{n} \sum_i K_n(y - Y_i),$$

where K_n is a kernel (Fernholz, 1991).

Armed with an estimate of F—$\hat{F}_n$, say—we compute $\hat{\boldsymbol{z}}$, where $\hat{z}_i = \Phi^{-1}\{\hat{F}_n(y_i)\}$, and optimize

$$\ell_{\text{ML}}(\boldsymbol{\omega} \mid \hat{\boldsymbol{z}}) = -\frac{1}{2}\log|\boldsymbol{\Omega}| - \frac{1}{2}\hat{\boldsymbol{z}}'\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{z}}$$

to obtain $\hat{\boldsymbol{\omega}}$. This approach is advantageous when the marginal distribution is complicated, but has the drawback that uncertainty regarding the marginal distribution is not reflected in the (ML) estimate of $\hat{\boldsymbol{\omega}}$'s variance. This deficiency can be avoided by using a bootstrap sample $\{\hat{\boldsymbol{\omega}}_1^*, \ldots, \hat{\boldsymbol{\omega}}_{n_b}^*\}$, the jth element of which can be generated by (1) simulating $\boldsymbol{U}_j^*$ from the copula at $\boldsymbol{\omega} = \hat{\boldsymbol{\omega}}$; (2) computing a new response $\boldsymbol{Y}_j^*$ as $Y_{ji}^* = \hat{F}_n^{-1}(U_{ji}^*)$ $(i = 1, \ldots, n)$, where $\hat{F}_n^{-1}(p)$ is the empirical quantile function; and (3) applying the estimation procedure to $\boldsymbol{Y}_j^*$. We compute sample quantiles using the median-unbiased approach recommended by Hyndman and Fan (1996). It is best to compute the bootstrap interval using the Gaussian method since that interval tends to have the desired coverage rate while the quantile method tends to produce an interval that is too narrow. This is because the upper-quantile estimator is inaccurate while the bootstrap estimator of $\mathbb{V}\hat{\boldsymbol{\omega}}$ is rather more accurate. To get adequate performance using the quantile method, a much larger sample size is required.

Although this approach may be necessary when the marginal distribution does not appear to take a familiar form, two-stage estimation does have a significant drawback, even for larger samples. If agreement is at least fair, dependence may be sufficient to pull the empirical marginal distribution away from the true marginal distribution. In such cases, simultaneous estimation of the marginal distribution and the copula should perform better. Development of such a method is beyond the scope of this article.

6 Application to simulated data

To investigate the performance of Sklar's ω relative to Krippendorff's α, we applied both methods to simulated outcomes. We carried out a study for each level of measurement, for various sample sizes, and for a few values of ω. The study plan is shown in Table 5, where $\mathcal{B}eta(\alpha, \beta)$ denotes a beta distribution, $\mathcal{L}(\mu, \sigma)$ denotes a Laplace distribution, $\phi(\mu, \sigma)$ denotes a Gaussian density function, $cat(\boldsymbol{p})$ denotes a categorical distribution, and $\mathcal{B}er(p)$ denotes a Bernoulli distribution.

We applied Krippendorff's α and our procedure to each of 500–1,000 simulated datasets for each scenario. The results are shown in Table 6. The coverage

	Margins		ω	Sample Geometry n_u, n_c	Method	Interval
$\mathcal{Beta}(1.5, 2)$			0.70	30, 3	ML	ML
$\mathcal{Beta}(13, 2)$			0.95	10, 5	ML	ML
$\mathcal{L}(12, 4)$			0.65	40, 2	ML	ML
$0.3\,\phi(0, 1) + 0.7\,\phi(3, 0.5)$			0.80	100, 4	SMP	Bootstrap
$cat(0.1, 0.3, 0.2, 0.05, 0.35)$			0.90	20, 10	DT	Sandwich
$\mathcal{Ber}(0.7)$			0.40	300, 6	CML	Sandwich

Table 5: Our simulation scenarios. The images show the shapes of the beta and Gaussian-mixture densities.

rates are for 95% intervals. All of the intervals for α are bootstrap intervals.

For every simulation scenario, our estimator exhibited smaller bias, variance (excepting the final scenario), and mean squared error, and a significantly higher coverage rate. Our method proved especially advantageous for the first, fifth, and sixth scenarios. This is not surprising in light of the fact that Krippendorff's α implicitly assumes Gaussianity; the marginal distributions for the first, fifth, and sixth scenarios are far from Gaussian, and so Krippendorff's α performs relatively poorly for those scenarios. Krippendorff's α performed best for the second and third scenarios because the $\mathcal{Beta}(13, 2)$ and Laplace distributions do not depart too markedly from Gaussianity. Note that our estimator had a much larger variance than did $\hat{\alpha}$ for the final scenario. This is because we employed composite likelihood, which is a must for binary data since the DT estimator is badly biased for binary outcomes.

7   R package `sklarsomega`

Here we briefly introduce our R package, `sklarsomega`, by way of a usage example. The package is available for download from the Comprehensive R Archive Network.

The following example applies Sklar's ω to the nominal data from the first case study of Section 3. We provide the value `"asymptotic"` for argument `confint`. This results in sandwich estimation for the DT approach (see Section 5.4). We estimate $\mathcal{J}$ using a parallel bootstrap, where $n_b = 1,000$ and six CPU cores are employed (control parameter `type` takes the default value of `"SOCK"`). Since argument `verbose` was set to `TRUE`, a progress bar appears (Solymos and Zawadzki, 2017). We see that $\hat{\mathcal{J}}$ was computed in one minute and 49 seconds.

Margins	ω	Median Est.	Bias	Variance	MSE	Coverage
$\mathcal{B}eta(1.5, 2)$	0.70	$\hat{\omega} = 0.695$	2%	0.0065	0.0067	94%
		$\hat{\alpha} = 0.575$	19%	0.0073	0.0242	51%
$\mathcal{B}eta(13, 2)$	0.95	$\hat{\omega} = 0.942$	2%	0.0017	0.0021	95%
		$\hat{\alpha} = 0.930$	3%	0.0021	0.0031	66%
$\mathcal{L}(12, 4)$	0.65	$\hat{\omega} = 0.651$	2%	0.0098	0.0099	93%
		$\hat{\alpha} = 0.634$	4%	0.0111	0.0120	89%
$0.3\,\phi(0, 1) + 0.7\,\phi(3, 0.5)$	0.80	$\hat{\omega} = 0.788$	2%	0.0008	0.0010	95%
		$\hat{\alpha} = 0.756$	6%	0.0018	0.0040	73%
$cat(0.1, 0.3, 0.2, 0.05, 0.35)$	0.90	$\hat{\omega} = 0.900$	$< 1\%$	0.0010	0.0010	98%
		$\hat{\alpha} = 0.504$	44%	0.0052	0.1626	0%
$\mathcal{B}er(0.7)$	0.40	$\hat{\omega} = 0.397$	6%	0.0173	0.0180	93%
		$\hat{\alpha} = 0.244$	38%	0.0007	0.0240	0%

Table 6: Results from our simulation study.

```
> data = matrix(c(1,2,3,3,2,1,4,1,2,NA,NA,NA,
+                 1,2,3,3,2,2,4,1,2,5,NA,3,
+                 NA,3,3,3,2,3,4,2,2,5,1,NA,
+                 1,2,3,3,2,4,4,1,2,5,1,NA), 12, 4)
> colnames(data) = c("c.1.1", "c.2.1", "c.3.1", "c.4.1")
> fit = sklars.omega(data, level = "nominal", confint = "asymptotic", verbose = TRUE,
+                    control = list(bootit = 1000, parallel = TRUE, nodes = 6))

Control parameter 'type' must be "SOCK", "PVM", "MPI", or "NWS". Setting it to "SOCK".

   |+++++++++++++++++++++++++++++++++++++++++++++++++++| 100% elapsed = 01m 49s
> summary(fit)

Call:

sklars.omega(data = data, level = "nominal", confint = "asymptotic",
    verbose = TRUE, control = list(bootit = 1000, parallel = TRUE,
        nodes = 6))

Convergence:

Optimization converged at -40.42 after 31 iterations.

Control parameters:

bootit   1000
parallel TRUE
nodes    6
dist     categorical
type     SOCK

Coefficients:

      Estimate    Lower  Upper
```

```
inter  0.89420  0.76570 1.0230
p1     0.25170  0.01407 0.4893
p2     0.24070  0.01842 0.4631
p3     0.22740  0.04639 0.4084
p4     0.18880 -0.06007 0.4377
p5     0.09136 -0.15580 0.3385
```

Next we compute DFBETAs (Young, 2017) for units 6 and 11, and for coders 2 and 3. We see that omitting unit 6 results in a much larger value for $\hat{\omega}$, whereas unit 11 is not influential. Likewise, coder 2 is influential while coder 3 is not.

```
> (inf = influence(fit, units = c(6, 11), coders = c(2, 3)))
$dfbeta.units
         inter          p1           p2          p3          p4           p5
6  -0.07914843 0.03438538  0.052599491 -0.05540904 -0.05820757  0.026631732
11  0.01096758 0.04546670 -0.007630807 -0.01626192 -0.01514173 -0.006432246

$dfbeta.coders
          inter           p1           p2          p3         p4          p5
2  0.0579843781 -0.002743713  0.002974195 -0.02730064 0.01105672  0.01601343
3 -0.0008664934 -0.006572821 -0.048168128  0.05659853 0.02149364 -0.02335122

> fit$coef - t(inf$dfbeta.units)
               6          11
inter 0.97335265 0.88323664
p1    0.21731494 0.20623362
p2    0.18814896 0.24837926
p3    0.28280614 0.24365903
p4    0.24700331 0.20393747
p5    0.06472664 0.09779062
```

Much additional functionality is supported by `sklarsomega`, e.g., plotting, simulation. And we note that computational efficiency is supported by our use of sparse-matrix routines (Furrer and Sain, 2010) and a clever bit of Fortran code (Genz, 1992). Future versions of the package will employ C++ (Eddelbuettel and Francois, 2011).

8   Conclusion

Sklar's ω offers a flexible, principled, complete framework for doing statistical inference regarding agreement. In this article we developed various frequentist approaches for Sklar's ω, namely, maximum likelihood, distributional-transform approximation, composite marginal likelihood, and a two-stage semiparametric method. This was necessary because a single, unified approach does not exist for the form of Sklar's ω presented in Section 4, wherein the copula is applied directly to the outcomes. An appealing alternative would be a hierarchical formulation of Sklar's ω such that the copula is applied through the responses' mean structure. This would permit, for example, a well-motivated Bayesian scheme and/or expectation-maximization algorithm to be devised.

Another potentially appealing extension/refinement would focus on composite likelihood inference for the current version of the model. Perhaps one should use a different composite likelihood, for example, and/or employ well-chosen weights (Xu et al., 2016).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Altman, D. G. and Bland, J. M. (1983). Measurement in medicine: The analysis of method comparison studies. *The Statistician*, pages 307–317.

Aravind, K. R. and Nawarathna, L. S. (2017). A statistical method for assessing agreement between two methods of heteroscedastic clinical measurements using the copula method. *Journal of Medical Statistics and Informatics*, 5(1):3.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Burgert, C. and Rüschendorf, L. (2006). On the optimal risk allocation problem. *Statistics & Decisions*, 24(1/2006):153–171.

Burnham, K. P., Anderson, D. R., and Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, 65(1):23–35.

Byrd, R., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

Dannull, K., Stein, J., Hughes, J., and Kline-Fath, B. (2017). Grading of liver herniation in cases of congenital diaphragmatic hernia: Further refining neonatal mortality (without the use of liver volume). *under review*.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*, volume 1. Cambridge University Press.

Eddelbuettel, D. and Francois, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York.

Fernholz, L. T. (1991). Almost sure convergence of smoothed empirical distribution functions. *Scandinavian Journal of Statistics*, 18(3):255–262.

Flegal, J., Haran, M., and Jones, G. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science*, 23(2):250–260.

Furrer, R. and Sain, S. R. (2010). spam: A sparse matrix R package with emphasis on MCMC methods for Gaussian Markov random fields. *Journal of Statistical Software*, 36(10):1–25.

Genest, C. and Neslehova, J. (2007). A primer on copulas for count data. *Astin Bulletin*, 37(2):475.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, pages 141–149.

Geyer, C. J. (2013). Le Cam made simple: Asymptotics of maximum likelihood without the LLN or CLT or sample size going to infinity. In Jones, G. L. and Shen, X., editors, *Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton*. Institute of Mathematical Statistics, Beachwood, Ohio, USA.

Godambe, V. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, pages 1208–1211.

Hayes, A. F. and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89.

Hyndman, R. J. and Fan, Y. (1996). Sample quantiles in statistical packages. *American Statistician*, 50:361–365.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314.

Kazianka, H. (2013). Approximate copula-based estimation and prediction of discrete spatial data. *Stochastic Environmental Research and Risk Assessment*, 27(8):2015–2026.

Kazianka, H. and Pilz, J. (2010). Copula-based geostatistical modeling of continuous and discrete data including covariates. *Stochastic Environmental Research and Risk Assessment*, 24(5):661–673.

Klaassen, C. A., Wellner, J. A., et al. (1997). Efficient estimation in the bivariate normal copula model: Normal margins are least favourable. *Bernoulli*, 3(1):55–77.

Krippendorff, K. (2012). *Content Analysis: An Introduction to Its Methodology*. Sage.

Krippendorff, K. (2013). Computing Krippendorff's alpha-reliability. Technical report, University of Pennsylvania.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80(1):221–239.

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(Oct):2295–2328.

Nelsen, R. B. (2006). *An Introduction to Copulas.* Springer, New York.

Nissi, M. J., Mortazavi, S., Hughes, J., Morgan, P., and Ellermann, J. (2015). T2* relaxation time of acetabular and femoral cartilage with and without intra-articular Gd-DTPA2 in patients with femoroacetabular impingement. *American Journal of Roentgenology*, 204(6):W695.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, pages 1033–1048.

R Core Team (2018). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rüschendorf, L. (1981). Stochastically ordered distributions and monotonicity of the OC-function of sequential probability ratio tests. *Statistics*, 12(3):327–338.

Rüschendorf, L. (2009). On the distributional transform, Sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference*, 139(11):3921–3927.

Serfling, R. and Mazumder, S. (2009). Exponential probability inequality and convergence results for the median absolute deviation and its modifications. *Statistics & Probability Letters*, 79(16):1767–1773.

Solymos, P. and Zawadzki, Z. (2017). *pbapply: Adding Progress Bar to '*apply' Functions.* R package version 1.3-2.

Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis*, 92(1):1–28.

Xu, X., Reid, N., and Xu, L. (2016). Note on information bias and efficiency of composite likelihood. *arXiv preprint arXiv:1612.06967.*

Xue-Kun Song, P. (2000). Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics*, 27(2):305–320.

Young, D. S. (2017). *Handbook of Regression Methods.* CRC Press.