sklarsomega: An R Package for Measuring Agreement Using Sklar's Omega Coefficient

John Hughes *

September 10, 2018

Abstract

R package **sklarsomega** provides tools for measuring agreement using Sklar's ω coefficient, which subsumes Krippendorff's α coefficient, which in turn subsumes a number of other well-known agreement coefficients. The package permits users to apply the ω methodology to nominal, ordinal, interval, or ratio scores; can accommodate any number of units, any number of coders, and missingness; and can measure intra-coder agreement, inter-coder agreement, and agreement relative to a gold standard. Classical inference is available for all levels of measurement while Bayesian inference is available for interval data and ratio data only.

Keywords: Agreement coefficient; Bayesian; Composite likelihood; Distributional transform; Gaussian copula; Markov chain Monte Carlo; R

1   Introduction: Measuring agreement in R

Sklar's ω (Hughes, 2018) is a model-based alternative to Krippendorff's α (Hayes and Krippendorff, 2007), a well-known nonparametric measure of agreement. Although Krippendorff's α is intuitive, flexible, and subsumes a number of other coefficients of agreement, 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; www.johnhughes.org

- 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;

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

- permits Bayesian inference for interval or ratio scores.

In R (Ihaka and Gentleman, 1996), Krippendorff's α can be applied using function `kripp.alpha()` of package **irr** (Gamer et al., 2012) and function `kripp.boot()` of package **kripp.boot** (Proutskova and Gruszczynski, 2017).

2 The agreement problem

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)). 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 (up to randomness) against agreement, as desired.

3 Models and software

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

The stochastic form of the Gaussian copula model is given by

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

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 $\boldsymbol{U} = (U_1, \ldots, U_n)^\top$ 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 outcomes, we assume the categorical distribution

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

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, and support in package **sklarsomega**.

For interval outcomes F can be practically any continuous distribution. Our package supports the Gaussian, Laplace, Student's t, and gamma distributions, which we denote as $\mathcal{N}(\mu, \sigma)$, $\mathcal{L}(\mu, \sigma)$, $\mathcal{T}(\nu, \mu)$, and $\mathcal{G}(\alpha, \beta)$, respectively. The Laplace and t distributions are useful for accommodating heavier-than-Gaussian tails, and the t and gamma distributions can accommodate asymmetry.

Another possibility for continuous outcomes is to first estimate F nonparametrically, and then estimate the copula parameters in a second stage. In Section 3.2 we will provide details regarding this approach.

Finally, two natural choices for ratio outcomes are the beta and Kumaraswamy distributions, the two-parameter versions of which are supported by our package. We denote these distributions as $\mathcal{B}(\alpha, \beta)$ and $\mathcal{K}(a, b)$.

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

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

3

where

$$
\mathbf{\Omega}_i = \begin{array}{c} \\ c_1 \\ c_2 \\ \vdots \\ c_{n_c} \end{array} \begin{array}{cccc} c_1 & c_2 & \ldots & c_{n_c} \\ \left(\begin{array}{cccc} 1 & \omega & \ldots & \omega \\ \omega & 1 & \ldots & \omega \\ \vdots & \vdots & \ddots & \vdots \\ \omega & \omega & \ldots & 1 \end{array}\right) \end{array}
$$

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 $\mathbf{\Omega}$. 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 $\mathbf{\Omega}_i$ 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

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

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

$$
\boldsymbol{\Omega}_i = \begin{matrix}
 & \begin{matrix} g & c_1 & c_2 & \ldots & c_{n_c} \end{matrix} \\
\begin{matrix} g \\ c_1 \\ c_2 \\ \vdots \\ c_{n_c} \end{matrix} &
\begin{pmatrix}
1 & \omega_{g1} & \omega_{g2} & \ldots & \omega_{gn_c} \\
\omega_{g1} & 1 & \omega_c & \ldots & \omega_c \\
\omega_{g2} & \omega_c & 1 & \ldots & \omega_c \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\omega_{gn_c} & \omega_c & \omega_c & \ldots & 1
\end{pmatrix}
\end{matrix},
$$

where $\omega_{gj} = H(\boldsymbol{x}_j^\top \boldsymbol{\beta})$, 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{matrix}
 & \begin{matrix} g_1 & c_{111} & c_{112} & c_{121} & c_{122} & c_{211} & c_{212} & c_{221} & c_{222} \end{matrix} \\
\begin{matrix} g_1 \\ c_{111} \\ c_{112} \\ c_{121} \\ c_{122} \\ c_{211} \\ c_{212} \\ c_{221} \\ c_{222} \end{matrix} &
\begin{pmatrix}
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{pmatrix}
\end{matrix},
$$

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 $\omega_{\bullet\bullet\bullet}$ 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).

3.1 Classical inference for all levels of measurement

When the response is continuous, i.e., when the level of measurement is interval or ratio, we recommend maximum likelihood (ML) or Bayesian inference for

Sklar's ω. When the marginal distribution is a categorical distribution (for nominal or ordinal level of measurement), likelihood-based 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. 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.

3.1.1   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 $\boldsymbol{\theta} = (\boldsymbol{\omega}^\top, \boldsymbol{\psi}^\top)^\top$ given observations $\boldsymbol{y}$ is

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

where $z_i = \Phi^{-1}\{F(y_i)\}$ and $\mathbf{I}$ denotes the $n \times n$ identity matrix. We obtain $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ by minimizing $-\ell_{\mathrm{ML}}$. 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 $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$:

$$\{\boldsymbol{\theta} : (\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta})^\top \hat{\boldsymbol{\mathcal{I}}}_{\mathrm{ML}} (\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}) \le \chi^2_{1-\alpha,q}\},$$

where $\hat{\boldsymbol{\mathcal{I}}}_{\mathrm{ML}}$ 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 $\boldsymbol{\psi}_0 = (\mu_0, \nu_0)^\top = (\mathrm{med}_n, \mathrm{mad}_n)^\top$ (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 $\boldsymbol{\psi}_0 = (\alpha_0, \beta_0)^\top$, 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. For the Kumaraswamy distribution we use $a_0 = 1$ and $b_0 = 1$. Finally, for a categorical distribution we use the empirical probabilities.

3.1.2 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:

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

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 $\hat{\mathcal{I}}_{\text{DT}}$ 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).

3.1.3 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

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

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.

3.1.4 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 $\hat{\boldsymbol{\theta}}_{\text{DT}}$ and $\hat{\boldsymbol{\theta}}_{\text{CML}}$ have sandwich forms (Godambe, 1960; Geyer, 2013). Specifically, we have

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{CML}} - \boldsymbol{\theta}) &\Rightarrow_n \mathcal{N}\{\mathbf{0}, \, \mathcal{I}_{\text{CML}}^{-1}(\boldsymbol{\theta}) \mathcal{J}_{\text{CML}}(\boldsymbol{\theta}) \mathcal{I}_{\text{CML}}^{-1}(\boldsymbol{\theta})\} \\
\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{DT}} - \boldsymbol{\theta}) &\Rightarrow_n \mathcal{N}\{\mathbf{0}, \, \mathcal{I}_{\text{DT}}^{-1}(\boldsymbol{\theta}) \mathcal{J}_{\text{DT}}(\boldsymbol{\theta}) \mathcal{I}_{\text{DT}}^{-1}(\boldsymbol{\theta})\},
\end{aligned}$$

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

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

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 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). Package **sklarsomega** makes the procedure even more efficient through parallelization.

3.2 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

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

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 (see the introduction to Section 4 for information regarding interpretation of agreement coefficients), 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 will be the aim of a future project.

3.3 Bayesian inference for interval or ratio scores

Since the Sklar's ω likelihood is not available in the case of nominal or ordinal scores, *true* Bayesian inference is infeasible for those levels of measurement. It is possible, however, to do *pseudo*-Bayesian inference for discrete scores. This entails using the appropriate CML or DT-based objective function in place of the likelihood. Although sound theory supports this approach (Ribatet et al., 2012), package **sklarsomega** does not support pseudo-Bayesian inference, for two reasons. First, pseudo-Bayesian inference requires a curvature correction because both the CML and the DT-based objective functions have too large a curvature relative to the true likelihood; unfortunately, the curvature adjustment is based on a time-consuming frequentist procedure. Second, we have no reason to suspect that the (curvature-adjusted) pseudo-posterior will have (at least approximately) the same shape as the true posterior.

Via function `sklars.omega.bayes()`, package **sklarsomega** does support Bayesian inference for interval or ratio scores. The function's signature appears below.

```
sklars.omega.bayes(data, level = c("interval", "ratio"), verbose = FALSE,
  control = list())
```

As we mentioned above, package **sklarsomega** currently supports gamma, Gaussian, Laplace, and t marginal distributions for interval outcomes; and beta and Kumaraswamy marginals for ratio outcomes.

The Sklar's ω posterior is given by

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto L(\boldsymbol{\theta} \mid \boldsymbol{y})p(\boldsymbol{\theta}),$$

where $p(\cdot)$ denotes a prior distribution and

$$L(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{\frac{1}{|\boldsymbol{\Omega}|^{1/2}} \exp[-\frac{1}{2}\boldsymbol{z}^{\top}\{\boldsymbol{\Omega}(\boldsymbol{\omega})^{-1} - \mathbf{I}\}\boldsymbol{z}]}{\prod_i f(y_i \mid \boldsymbol{\psi})}.$$

In the interest of striking a sensible balance between flexibility and usability, we do not permit the user to specify $p(\boldsymbol{\theta})$. Instead, we assign an independent, noninformative prior to each element of $\boldsymbol{\theta}$—i.e., $p(\boldsymbol{\theta}) = p(\boldsymbol{\omega})p(\boldsymbol{\psi}) = \{\prod_{k=1}^{m} p(\omega_k)\}p(\psi_1)p(\psi_2)$. The prior for ω_k $(k = 1, \ldots, m)$ is standard uniform. Each of α, β, σ, ν, a, and b is given a $\mathcal{G}(0.01, 0.01)$ prior distribution. And the prior for μ is Gaussian with mean zero and standard deviation 1,000.

As for sampling, we use a Gaussian random walk for each parameter, and transform when necessary. The user can control the acceptance rates by adjusting the standard deviations of the perturbations, which can be supplied to function `sklars.omega.bayes()` via argument `control`. Consider parameter α, for example. To generate a proposal for α, we begin by drawing $\eta^* = \eta + \mathcal{N}(0, \sigma_1)$, where η was obtained during the previous iteration. Then we take $\alpha^* = \exp(\eta^*)$, which of course yields a log-normal proposal (necessitating the inclusion of the ratio $\mathcal{LN}\{\alpha; \log(\alpha^*), \sigma_1\}/\mathcal{LN}\{\alpha^*; \log(\alpha), \sigma_1\}$ in the Metropolis–Hastings acceptance probability). The proposal standard deviation σ_1 can be set using the syntax `control = list(sigma.1 = 0.2)`, for example. This proposal scheme is employed for all of the non-negative parameters, with σ_1 the tuning parameter for α, ν, and a; and σ_2 the tuning parameter for β, σ, and b. Again, these standard deviations can be set straightforwardly in the function call: `control = list(sigma.1 = 0.2, sigma.2 = 0.3)`; or they can be omitted, in which case they default to the value 0.1.

Updates for the μ chain take the form of a Gaussian random walk: $\mu^* = \mu + \mathcal{N}(0, \sigma_j)$, where $j = 1$ if the marginal distribution is Gaussian or Laplace, or $j = 2$ if the marginal distribution is $\mathcal{T}(\nu, \mu)$. The acceptance rate can be modulated via control parameter `sigma.1` (for a $\mathcal{N}$ or $\mathcal{L}$ marginal) or `sigma.2` (for a $\mathcal{T}$ marginal).

Although we propose values for the ω_k independently, we accept or reject those proposals jointly so that $|\boldsymbol{\Omega}|$ and $\boldsymbol{\Omega}^{-1}$ need not be computed too frequently. Each proposal begins with a Gaussian random step, $\eta_k^* = \eta_k + \mathcal{N}(0, \sigma_{\omega_k})$. Then we apply the logistic function to map into the unit interval: $\omega_k^* = \exp(\eta_k^*)/\{1 + \exp(\eta_k^*)\}$. This of course requires us to include the Jacobian $\exp(\eta_k^*)/\{1 + \exp(\eta_k^*)\}^2$ in the Metropolis–Hastings acceptance probability. The acceptance rates can be adjusted by passing a vector of proposal standard deviations to `sklars.omega.bayes()`, e.g., `control = list(sigma.omega = c(0.1, 0.1, 0.3))` (in the case of $\dim(\boldsymbol{\omega}) = 3$).

The remaining control parameters are `dist` (for selecting the marginal distribution), `minit`, `maxit`, and `tol`. Function `sklars.omega.bayes()` draws at least `minit` posterior samples. We use 1,000 as the default, and minimum, value for `minit`. Similarly, `sklars.omega.bayes()` draws at most `maxit` samples, with 10,000 as the default value.

Since the Markov chain tends to mix well, between 1,000 and 10,000 samples are usually sufficient for obtaining stable estimates (of posterior means and of DIC (Spiegelhalter et al., 2002)). We recommend using the fixed-width method (Flegal et al., 2008) for determining when to stop sampling, and we provide control parameter `tol` for this purpose. In the fixed-width approach, one chooses a small positive threshold ϵ and terminates sampling when all estimated coeffi-

cients of variation are smaller than said threshold, where the estimated coefficient of variation for parameter θ_j is $\widehat{cv}_j = \text{mcse}(\hat{\theta}_j)/|\hat{\theta}_j|$, with 'mcse' denoting Monte Carlo standard error. That is, sampling terminates when $\widehat{cv}_j < \epsilon$ for all $j \in \{1, \ldots, \dim(\boldsymbol{\theta})\}$. The user can set the threshold value via control parameter `tol`, which defaults to 0.1. In the interest of computational efficiency, function `sklars.omega.bayes()` computes Monte Carlo standard errors (using package **mcmcse**) only every `minit` samples.

4   Illustrations

Here we illustrate the use of **sklarsomega** by applying Sklar's ω to a couple of datasets. Although our understanding of the agreement problem aligns with that of Krippendorff's α and other related measures (e.g., Spearman's ρ, Cohen's κ, Scott's π (Spearman, 1904; Cohen, 1960; Scott, 1955)), we shall 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 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.

4.1   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.

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

Now we apply our methodology, first loading package **sklarsomega**.

	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 seven missing values.

```
R> library(sklarsomega)
```

```
sklarsomega: Measuring Agreement Using Sklar's Omega Coefficient
Version 2.0 created on 2018-06-18.
Copyright (c) 2018 John Hughes
For citation information, type citation("sklarsomega").
Type help(package = sklarsomega) to get started.
```

Now we create the dataset as a matrix. Then we supply appropriate column names. This is a necessary step since package function build.R() uses the column names to create the copula correlation matrix (which we denoted as $\mathbf{\Omega}$ above but is denoted as $\mathbf{R}$ in the package documentation).

```
R> 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)
R> colnames(data) = c("c.1.1", "c.2.1", "c.3.1", "c.4.1")
R> data

      c.1.1 c.2.1 c.3.1 c.4.1
 [1,]     1     1    NA     1
 [2,]     2     2     3     2
 [3,]     3     3     3     3
 [4,]     3     3     3     3
 [5,]     2     2     2     2
 [6,]     1     2     3     4
 [7,]     4     4     4     4
 [8,]     1     1     2     1
 [9,]     2     2     2     2
[10,]    NA     5     5     5
[11,]    NA    NA     1     1
[12,]    NA     3    NA    NA
```

Function check.colnames() can be used to perform a (rudimentary) check of the column names. This function returns a list comprising exactly two el-

ements, `success` and `cols`. The former is a boolean that indicates appropriateness or inappropriateness of the column names. If the column names are appropriate, element `cols` is equal to `NULL`. Otherwise, `col` contains the numbers of the columns that failed the check.

```
R> (check.colnames(data))

$success
[1] TRUE

$cols
NULL
```

In the next example we introduce errors for columns 1 and 4. Then we provide column names that pass the check but are illogical. Finally, we revert to the correct names.

```
R> colnames(data) = c("c.a.1", "c.2.1", "c.3.1", "C.4.1")
R> (check.colnames(data))

$success
[1] FALSE

$cols
[1] 1 4

R> colnames(data) = c("g", "c.2.1", "c.1.47", "c.2.1")
R> (check.colnames(data))

$success
[1] TRUE

$cols
NULL

R> colnames(data) = c("c.1.1", "c.2.1", "c.3.1", "c.4.1")
```

Now we create the copula correlation matrix, display the portion of the matrix that corresponds to the first two units, and verify that the matrix contains exactly one parameter, namely, the inter-coder agreement coefficient (which has the dummy value 0.1).

```
R> temp = build.R(data)
R> names(temp)

[1] "R"      "onames"
```

14

```
R> R = temp$R
R> R[1:7, 1:7]

     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]  1.0  0.1  0.1  0.0  0.0  0.0  0.0
[2,]  0.1  1.0  0.1  0.0  0.0  0.0  0.0
[3,]  0.1  0.1  1.0  0.0  0.0  0.0  0.0
[4,]  0.0  0.0  0.0  1.0  0.1  0.1  0.1
[5,]  0.0  0.0  0.0  0.1  1.0  0.1  0.1
[6,]  0.0  0.0  0.0  0.1  0.1  1.0  0.1
[7,]  0.0  0.0  0.0  0.1  0.1  0.1  1.0

R> temp$onames

[1] "inter"
```

Next we apply Sklar's ω for nominal data. Since there are five categories, function `sklars.omega()` automatically selects the DT approach. We do a full bootstrap with sample size $n_b = 1,000$. We do the bootstrapping in embarrassingly parallel fashion, using six CPU cores on the local machine (hence `type` is equal to `"SOCK"`). Since we set `verbose` equal to `TRUE`, a progress bar is shown. The computation took 19 minutes and 21 seconds.

```
R> set.seed(99)
R> fit = sklars.omega(data, level = "nominal", confint = "bootstrap",
+                     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 = 19m 21s
```

One can view a summary by passing the fit object to function `summary.sklarsomega()`.

```
R> summary(fit)

Call:

sklars.omega(data = data, level = "nominal", confint = "bootstrap",
    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  0.77530 1.0130
p1     0.25170  0.02420 0.4792
p2     0.24070  0.09078 0.3907
p3     0.22740  0.07583 0.3790
p4     0.18880  0.03522 0.3424
p5     0.09136 -0.06572 0.2484
```

Our method produced $\hat{\omega} = 0.89$ and $\omega \in (0.78, 1.00)$, 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:

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

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.

Now we compute asymptotic (sandwich) intervals. This requires a much shorter running time.

```
R> set.seed(12)
R> 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 = 02m 03s

R> 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  0.76270 1.0260
p1      0.25170  0.01703 0.4864
p2      0.24070  0.01796 0.4635
p3      0.22740  0.04792 0.4069
p4      0.18880 -0.06715 0.4447
p5      0.09136 -0.16860 0.3513
```

The marked difference between our results and those obtained for Krippendorff's α 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.

Influence (DFBETA (Young, 2017)) statistics can be obtained using the package's `influence()` function, as illustrated below. Here we investigate the influence of units 6 and 11, and of coders 2 and 3.

```
R> (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
```

We conclude this illustration by simulating a dataset from the fitted model and then converting the resulting vector to matrix form for display. Note that row 12 was removed since, having only one score, it carries no information about ω.

```
R> sim = simulate(fit, seed = 42)
R> data.sim = t(fit$data)
R> data.sim[! is.na(data.sim)] = sim[, 1]
R> data.sim = t(data.sim)
R> data.sim

      c.1.1 c.2.1 c.3.1 c.4.1
 [1,]     3     4    NA     3
 [2,]     3     2     2     2
 [3,]     3     3     3     3
 [4,]     1     1     1     2
 [5,]     3     4     3     4
 [6,]     5     4     4     4
 [7,]     2     2     1     2
 [8,]     4     4     4     4
 [9,]     1     2     1     1
[10,]    NA     1     2     2
[11,]    NA    NA     2     2
```

4.2 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.

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

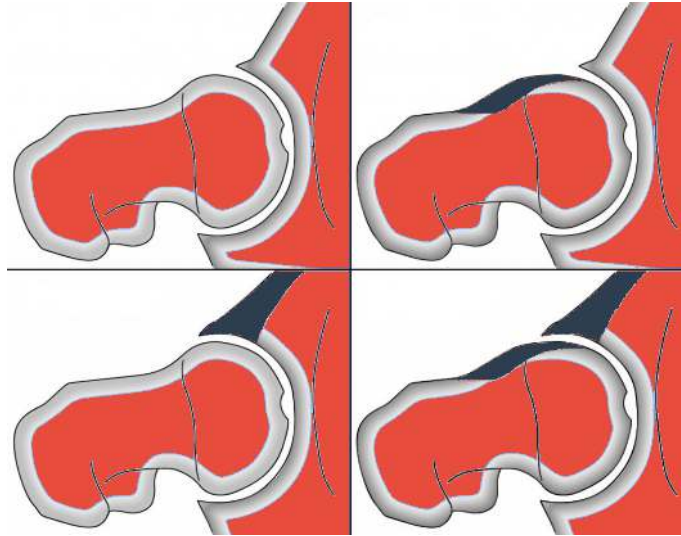Figure 2: Raw and contrast-enhanced T2* values for femoral cartilage.

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 will carry out both maximum likelihood and Bayesian analyses for (a subset of) these data, assuming $\mathcal{L}(\mu, \sigma)$ and $\mathcal{T}(\nu, \mu)$ marginal distributions.

First we load the cartilage data, which are included in the package, and apply the method of maximum likelihood for a Laplace marginal distribution. The running time is just over one second.

```
R> data(cartilage)
R> data = as.matrix(cartilage)[1:100, ]
R> colnames(data) = c("c.1.1", "c.2.1")
R> fit1 = sklars.omega(data, level = "interval", confint = "asymptotic",
+                       control = list(dist = "laplace"))
R> summary(fit1)

Call:

sklars.omega(data = data, level = "interval", confint = "asymptotic",
    control = list(dist = "laplace"))

Convergence:

Optimization converged at -593.5 after 18 iterations.

Control parameters:
```

```
dist laplace

Coefficients:

      Estimate   Lower    Upper
inter   0.8077  0.7379   0.8776
mu     26.5100 26.3500  26.6700
sigma   4.7090  3.8630   5.5550

AIC: 1193
BIC: 1203
```

We see that $\hat{\boldsymbol{\theta}} = (\hat{\omega}, \hat{\mu}, \hat{\sigma})^\top = (0.81, 26.51, 4.71)^\top$. This suggests that the contrast-enhanced T2* values agree nearly perfectly with their raw counterparts.

Now we repeat the analysis for a t marginal distribution.

```
R> fit2 = sklars.omega(data, level = "interval", confint = "asymptotic",
+                      control = list(dist = "t"))
R> summary(fit2)

Call:

sklars.omega(data = data, level = "interval", confint = "asymptotic",
    control = list(dist = "t"))

Convergence:

Optimization converged at -608.7 after 18 iterations.

Control parameters:

dist t

Coefficients:

      Estimate   Lower    Upper
inter   0.8701  0.8224   0.9179
nu      7.0280  5.2130   8.8430
mu     23.4400 22.2400  24.6400

AIC: 1223
BIC: 1233
```

This led to a considerably larger value for $\hat{\omega}$ and, given the two confidence intervals, a stronger conclusion for these data. But we must select the Laplace model since that model yielded much smaller values of AIC and BIC. In fact, the model probability (Burnham et al., 2011) is near zero.

```
R> aic = c(AIC(fit1), AIC(fit2))
R> (pr = exp((min(aic) - max(aic)) / 2))
```

```
[1] 2.516706e-07
```

Finally, we apply the Bayesian methodology described in Section 3.3. We begin by assuming a Laplace marginal model once again.

```
R> set.seed(111111)
R> fit1 = sklars.omega.bayes(data, verbose = FALSE,
+                            control = list(dist = "laplace", minit = 1000,
+                            maxit = 5000, tol = 0.01, sigma.1 = 1, sigma.2 = 0.1,
+                            sigma.omega = 0.2))
R> summary(fit1)

Call:

sklars.omega.bayes(data = data, verbose = FALSE, control = list(dist = "laplace",
    minit = 1000, maxit = 5000, tol = 0.01, sigma.1 = 1, sigma.2 = 0.1,
    sigma.omega = 0.2))


Number of posterior samples: 4000


Control parameters:

dist        laplace
minit       1000
maxit       5000
tol         0.01
sigma.1     1
sigma.2     0.1
sigma.omega 0.2


Coefficients:

      Estimate   Lower    Upper      MCSE
inter   0.8079  0.7366   0.8695  0.002111
mu     26.4600 25.7100  27.1400  0.011310
sigma   4.7990  3.9730   5.6970  0.025410


DIC: 1193
```

We see that sampling terminated when 4,000 samples had been drawn, since that sample size yielded $\widehat{cv}_j < 0.01$ for $j \in \{1, 2, 3\}$. As a second check we examine the plot given in Figure 4, which shows the estimated posterior mean for ω as a function of sample size. The estimate evidently stabilized after approximately 2,500 samples had been drawn.
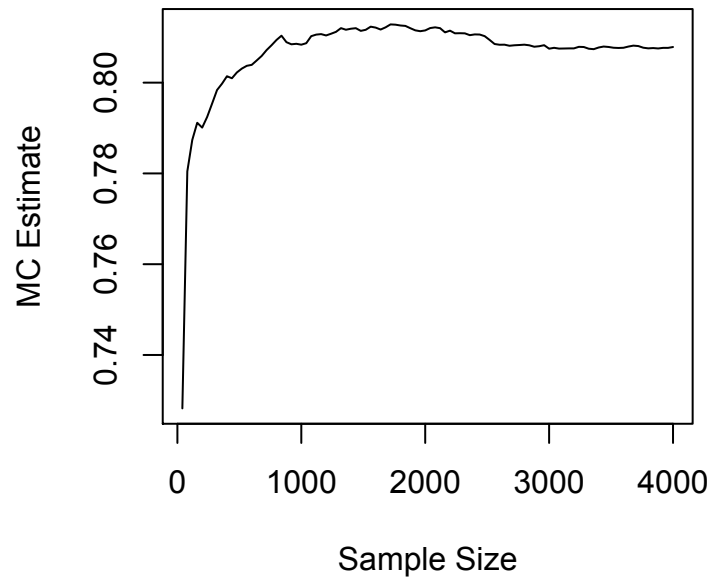
Figure 4: A plot of estimated posterior mean versus sample size for ω, having assumed a Laplace marginal distribution.

The proposal standard deviations (1 for μ, 0.1 for σ, and 0.2 for ω) led to sensible acceptance rates of 40%, 60%, and 67%.

```
R> fit1$accept

    inter        mu     sigma
0.6694174 0.4013503 0.5951488
```

For a t marginal distribution only 3,000 samples were required.

```
R> set.seed(4565)
R> fit2 = sklars.omega.bayes(data, verbose = FALSE,
+                            control = list(dist = "t", minit = 1000,
+                            maxit = 5000, tol = 0.01, sigma.1 = 0.2,
+                            sigma.2 = 2, sigma.omega = 0.2))
R> summary(fit2)

Call:

sklars.omega.bayes(data = data, verbose = FALSE, control = list(dist = "t",
    minit = 1000, maxit = 5000, tol = 0.01, sigma.1 = 0.2, sigma.2 = 2,
    sigma.omega = 0.2))
```

```
Number of posterior samples: 3000

Control parameters:

dist        t
minit       1000
maxit       5000
tol         0.01
sigma.1     0.2
sigma.2     2
sigma.omega 0.2

Coefficients:

        Estimate   Lower   Upper      MCSE
inter      0.874  0.8283   0.919  0.002054
nu         6.720  5.0210   8.424  0.053200
mu        23.450 22.2600  24.690  0.028070

DIC: 1224
```

Note that the Laplace model yielded a much smaller value of DIC, and hence a very small relative likelihood for the t model.

```
R> dic = c(fit1$DIC, fit2$DIC)
R> (pr = exp((min(dic) - max(dic)) / 2))

[1] 1.852924e-07
```

5   Summary and discussion

Sklar's ω offers a flexible, principled, complete framework for doing statistical inference regarding agreement. In this article we described three frequentist approaches to inference for Sklar's ω as well as a Bayesian methodology for interval or ratio outcomes. All of these approaches are supported by R package **sklarsomega**, version 2.0 of which is available from the Comprehensive R Archive Network. As illustrated in the preceding section, package **sklarsomega** also offers much useful related functionality.

Computational details

The results in this paper were obtained using R 3.3.3 with the **extraDistr** 1.8.9 package, the **hash** 2.2.6 package, the **LaplacesDemon** 16.0.1 package, the **Matrix** 1.2.8 package, the **mcmcse** 1.3.2 package, the **numDeriv** 2014.2.1 package, the **spam** 1.3.0 package, and the **pbapply** 1.3.2 package. R itself

and all packages used are available from the Comprehensive R Archive Network (CRAN) at `http://CRAN.R-project.org`.

References

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.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

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

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.

Gamer, M., Lemon, J., and Singh, I. F. P. (2012). *irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.84.

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

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.

Hughes, J. (2018). Sklar's Omega: A Gaussian copula-based framework for assessing agreement. *ArXiv e-prints*.

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.

Proutskova, P. and Gruszczynski, M. (2017). *kripp.boot: Bootstrap Krippendorff's Alpha Intercoder Reliability Statistic*. R package version 1.0.0.

Ribatet, M., Cooley, D., and Davison, A. C. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, pages 813–845.

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.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, pages 321–325.

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.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

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

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.