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

John Hughes[1*]

[1*]College of Health, Lehigh University, USA.

Corresponding author(s). E-mail(s): drjphughesjr@gmail.com;

Abstract

The statistical measurement of agreement—the most commonly used form of which is inter-coder agreement (also called inter-rater reliability), i.e., consistency of scoring among two or more coders for the same units of analysis—is important in a number of fields, e.g., content analysis, education, computational linguistics, sports. We propose Sklar's Omega, a Gaussian copula-based framework for measuring not only inter-coder agreement but also intra-coder agreement, inter-method agreement, and agreement relative to a gold standard. We demonstrate the efficacy and advantages of our approach by applying both Sklar's Omega and Krippendorff's Alpha (a well-established nonparametric agreement coefficient) to simulated data, to nominal data previously analyzed by Krippendorff, and to continuous data from an imaging study of hip cartilage in femoroacetabular impingement. 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. The package permits users to apply the Omega methodology to nominal scores, ordinal scores, percentages, counts, amounts (i.e., non-negative real numbers), and balances (i.e., any real number); and can accommodate any number of units, any number of coders, and missingness. Classical inference is available for all levels of measurement while Bayesian inference is available for continuous outcomes only.

Keywords: agreement coefficient, Bayesian, biomedical imaging, composite likelihood, distributional transform, femoroacetabular impingement, Gaussian copula, Markov chain Monte Carlo

1 Introduction

By 'agreement' we mean consistency of scoring by two or more coders for the same units of analysis. Inter-coder agreement, which also goes by 'inter-rater reliability' and various other names, is surely the most commonly measured form of agreement (see Gwet (2014) for a book-length treatment of inter-rater reliability), but intra-coder agreement, inter-method agreement (i.e., consistency of scoring among two or more scoring systems), and/or agreement with a gold standard may be of interest in some contexts.

In 2007, Hayes and Krippendorff put forth Krippendorff's α as a standard measure of agreement. Although Krippendorff's α is intuitive, flexible, and subsumes a number of other coefficients of agreement, we argue that α can be improved upon in a number of ways. To that end, we develop a parametric alternative (α is nonparametric). In keeping with the naming convention that is evident in the literature on agreement (e.g., Spearman's ρ (1904), Cohen's κ (1960), Scott's π (1955)), we call our approach Sklar's ω (after Sklar's theorem (1959), which establishes

the theoretical basis for the application of copulas). Sklar's ω improves upon Krippendorff's 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;
- 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 cover essential preliminaries. In Section 3 we specify the flexible, fully parametric statistical model upon which Sklar's ω is based. In Section 4 we describe four approaches to frequentist inference for ω. In Section 5 we develop Bayesian inference for continuous scores. In Section 6 we use an extensive simulation study to assess the performance of Sklar's ω relative to Krippendorff's α. In Sections 7.1 and 7.2 we apply both Sklar's ω and Krippendorff's α to nominal data previously considered by Krippendorff and to continuous data from an imaging study of cartilage in the context of femoroacetabular impingement, a deformity of the human hip joint. Finally, in Section 8 we point out potential limitations of our methodology, and posit directions for future research. In an appendix 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, 2021).

2 Preliminaries

In this section we briefly review the literature on agreement coefficients, compare and contrast Sklar's ω and Krippendorff's α, provide a well-established scale according to which we interpret values of ω and α, and describe the level-of-measurement typology employed in this article and supported by our R package.

2.1 Measuring agreement

An inter-coder agreement coefficient—which takes a value in the unit interval, with 0 indicating no agreement and 1 indicating perfect agreement—is a statistical measure of the extent to which two or more coders agree regarding the same units of analysis. The agreement problem has a long history and is important in many fields of inquiry, and numerous agreement statistics have been proposed.

Scott (1955) proposed the π coefficient for measuring agreement between two coders. Cohen (1960) criticized π and proposed the κ coefficient, which is still widely used despite its well-known shortcomings (Feinstein and Cicchetti, 1990; Cicchetti and Feinstein, 1990). Other oft-used measures of agreement are Gwet's AC_1 (Gwet, 2008) and Krippendorff's α (Hayes and Krippendorff, 2007), the latter of which is a contrast object for the ω coefficient developed in this article. For more comprehensive reviews of the literature on agreement, we refer the interested reader to the article by Banerjee et al. (1999), the article by Artstein and Poesio (2008), and the book by Gwet (2014).

2.2 Sklar's ω and Krippendorff's α

We propose Sklar's ω as an alternative to Krippendorff's α. Although Krippendorff's α is nonparametric, the method finds its genesis in a fully parametric setting, namely, the one-way mixed-effects ANOVA model, wherein agreement is modeled as positive intraclass (Pearson) correlation. Indeed, Krippendorff's α *is* the intraclass correlation coefficient (restricted to the unit interval) for scores that conform to the one-way mixed-effects ANOVA model. And the more general, nonparametric formulation of Krippendorff's α can then be obtained by dropping the assumption of Gaussianity and replacing squared Euclidean distance with an abstract distance function. We

refer the curious reader to a recent article by Hughes (2021a), who carefully develops Krippendorff's α from first principles and situates α among statistical procedures.

Sklar's ω likewise generalizes the one-way mixed-effects ANOVA model, but ω does so in a fully parametric fashion. Specifically, Sklar's ω models agreement as positive within-unit correlation/association, pairing a Gaussian copula for the joint distribution with an appropriate marginal distribution for the scores. We describe the Sklar's ω model in detail in Section 3.

2.3 Interpreting values of an agreement coefficient

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

Table 1 Guidelines for interpreting values of an agreement coefficient.

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

2.4 The Mosteller–Tukey typology

It is important to note that we employ the Mosteller–Tukey level-of-measurement typology (Mosteller and Tukey, 1977) in this article and in R package `sklarsomega`. Specifically, we support nominal scores, ordinal scores, percentages, counts, amounts (i.e., non-negative real numbers), and balances (i.e., any real number). These levels of measurement map quite naturally to marginal distributions for Sklar's ω. The mapping is given in Table 2. The table also defines our notation for the distributions. Additionally, we will need to refer to the chi-squared, lognormal, multinormal, and uniform distributions. We denote these distributions, respectively, as χ^2_q, $\textsc{Lognormal}(\mu, \sigma)$, $\textsc{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\textsc{Uniform}(a, b)$.

Table 2 Mapping between levels of measurement and marginal distributions for Sklar's ω. The second column also defines our notation for these distributions.

Level	Distribution
nominal, ordinal	$\textsc{Categorical}(\boldsymbol{p})$
percentage	$\textsc{Beta}(\alpha, \beta)$ $\textsc{Kumaraswamy}(a, b)$
count	$\textsc{Poisson}(\lambda)$ $\textsc{NegativeBinomial}(\mu, r)$
amount	$\textsc{Gamma}(\alpha, \beta)$
balance	$\textsc{Normal}(\mu, \sigma)$ $\textsc{Laplace}(\mu, \sigma)$ $\mathrm{T}(\nu, \mu)$ with noncentrality

Alternative typologies exist—see, e.g., Stevens (1946) and Chrisman (1998)—but are less well suited for use with Sklar's ω. In any case, the matter of typologies is still debated, and no typology appears to be entirely satisfactory.

3 Our model

In this section we first specify the direct Gaussian copula model, of which the Sklar's ω model is a special case. Then we discuss in more detail the marginal distributions for Sklar's ω. We conclude the section by describing the various task-specific forms of the copula correlation matrix that are employed in ω analyses.

3.1 The direct Gaussian copula 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 direct—as opposed to hierarchical (Musgrove et al., 2016; Han and De Oliveira, 2016; Hughes, 2021b)—Gaussian copula model is given by

$$\begin{aligned} \boldsymbol{Z} = (Z_1, \ldots, Z_n)' &\sim \text{NORMAL}(\boldsymbol{0}, \boldsymbol{\Omega}) \\ U_i = \Phi(Z_i) &\sim \text{UNIFORM}(0, 1) \\ Y_i = F_i^{-1}(U_i) &\sim F_i, \end{aligned} \tag{1}$$

where $i = 1, \ldots, n$; $\boldsymbol{\Omega}$ is a correlation matrix; Φ is the standard Gaussian cdf; F_i is the cdf for the ith outcome Y_i; and F_i^{-1} is the quantile function for F_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.

3.2 Marginal distributions for ω

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

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

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, Poisson, or negative binomial), 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 4, and evaluate through simulation in Section 6.

For amounts, i.e., non-negative real numbers, we support the gamma distribution. For balances, i.e., any real number, F can be practically any continuous distribution supported on the reals. Our R package supports the Gaussian, Laplace, and noncentral t distributions. The Laplace and t distributions are useful for accommodating heavier-than-Gaussian tails, and the t distribution can also accommodate asymmetry. Finally, two natural choices for percentages are the beta and Kumaraswamy distributions, the two-parameter versions of which are supported by our package.

Perhaps the reader can envision more "exotic" possibilities for continuous scores, e.g., mixture distributions (to handle multimodality or excess zeros, for example). Some such more complicated marginal distributions can be accommodated by first estimating F nonparametrically, and then estimating the copula parameters in a second stage. In Section 4 we will provide details regarding this semiparametric approach.

3.3 The copula correlation matrix

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,

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

where

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

for n_c coders $c_1, \ldots, c_{n_c}$.

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 $Y_{ij'}$ $(j \neq j')$, 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, Φ^{-1} denotes the quantile function for the standard normal distribution, 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

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

In this scheme ω_g captures agreement with the gold standard, and ω_c captures inter-coder agreement. We note that large ω_g implies large ω_c, but the converse is not necessarily true, i.e., the coders may agree with one another yet perform poorly relative to the gold standard.

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 $\mathbf{\Omega}$ might look like

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

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 $\mathbf{\Omega}_i$ is given by

$$\begin{array}{c} \\ g_1 \\ c_{111} \\ c_{112} \\ c_{121} \\ c_{122} \\ c_{211} \\ c_{212} \\ c_{221} \\ c_{222} \end{array} \begin{array}{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{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{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 $\omega_{\bullet\bullet\bullet}$ 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 may be infeasible (see the next section for details).

4 Approaches to classical inference for $\boldsymbol{\omega}$

When the response is continuous it is straightforward to do maximum likelihood (ML) inference for Sklar's ω since the likelihood (see (3) below) is meta-Gaussian. When the marginal distribution is discrete, 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 chosen marginal distribution.

The DT-based approximate likelihood performs very well for Poisson or negative binomial outcomes, and is even practically exact when the marginal variance is sufficiently large and agreement is not too strong (Hughes, 2021b). When the marginal distribution is categorical, composite marginal likelihood is indicated since the DT approach tends to perform poorly for such data. Specifically, the DT estimator tends to exhibit substantial bias, even for larger samples, which leads to unacceptably low coverage rates for confidence intervals. We demonstrate this by simulation in Section 6.

4.1 The method of maximum likelihood for Sklar's $\boldsymbol{\omega}$

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

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

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. Should the initial attempt at optimization fail, we revert to the more stable bounded Hooke–Jeeves algorithm (Varadhan et al., 2020; Hooke and Jeeves, 1961).

To estimate an asymptotic confidence ellipsoid for the ML approach we of course use the observed Fisher information matrix:

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

where $\hat{\boldsymbol{\mathcal{I}}}_{\text{ML}}$ denotes the observed information, $q = \dim(\boldsymbol{\theta})$, and $\chi^2_{1-\alpha,q}$ denotes the $1-\alpha$ quantile of the χ^2 distribution with q degrees of freedom.

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 $\boldsymbol{\psi}_0 = (\mu_0, \nu_0)' = (\text{med}_n, \text{mad}_n)'$ (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

$$\begin{aligned}\alpha_0 &= \bar{Y}^2/S^2 \\ \beta_0 &= \bar{Y}/S^2,\end{aligned}$$

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

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

when the marginal distribution is beta.

4.2 The distributional transform method

When the marginal distribution is discrete, 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 has on the order of 2^n terms and is thus intractable unless the sample is rather small. For some choices of marginal distribution 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 $W \sim \text{UNIFORM}(0, 1)$, 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:

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

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 seems crude, but it performs well in a wide variety of circumstances (Kazianka, 2013) and is even practically exact for some variants of the direct Gaussian copula model (Hughes, 2021b). Moreover, the DT approach is, for most marginal distributions, much more efficient computationally than the composite likelihood method described in the next section. For Sklar's ω we recommend using the DT approach for counts.

Since the DT-based objective function is, in general, misspecified, using $\hat{\boldsymbol{\mathcal{I}}}_{\text{DT}}$ alone usually leads to optimistic inference. This can be overcome by using a sandwich estimator (Godambe, 1960) or by doing a bootstrap (Davison and Hinkley, 1997). Sandwich estimation is described below in Section 4.4.

4.3 Composite marginal likelihood

For nominal or ordinal outcomes 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.

Note that we provide sensible starting values for optimization of ℓ_{DT} and ℓ_{CML}. Specifically, we supply the sample mean for Poisson scores, the sample mean and 1 for negative binomial scores, and empirical probabilities for categorical scores.

4.4 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

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{CML}} - \boldsymbol{\theta}) &\Rightarrow_n \text{NORMAL}\{\mathbf{0}, \boldsymbol{\mathcal{V}}_{\text{CML}}(\boldsymbol{\theta})\} \\ \sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{DT}} - \boldsymbol{\theta}) &\Rightarrow_n \text{NORMAL}\{\mathbf{0}, \boldsymbol{\mathcal{V}}_{\text{DT}}(\boldsymbol{\theta})\},\end{aligned}$$

where $\Rightarrow_n$ denotes convergence in distribution and $\boldsymbol{\mathcal{V}}_{\bullet} = \boldsymbol{\mathcal{I}}_{\bullet}^{-1}(\boldsymbol{\theta})\boldsymbol{\mathcal{J}}_{\bullet}(\boldsymbol{\theta})\boldsymbol{\mathcal{I}}_{\bullet}^{-1}(\boldsymbol{\theta})$. Here $\boldsymbol{\mathcal{I}}_{\bullet}$ is the appropriate sensitivity matrix:

$$\boldsymbol{\mathcal{I}}_{\bullet}(\boldsymbol{\theta}) = -\mathbb{E}\nabla^2\ell_{\bullet}(\boldsymbol{\theta} \mid \boldsymbol{Y}).$$

An estimate of this matrix can be produced as a side effect of optimization. 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 for DT-based inference but tends to lead to inflated standard errors for the CML method. Consequently, our R package supports sandwich estimation, as well as "full" parametric bootstrap inference, for the DT approach but not for the CML approach. For the CML approach package `sklarsomega` supports bootstrap inference only.

We note that sandwich estimation is considerably more efficient computationally than bootstrapping because it is much faster to approximate the score using numerical differentiation (Gilbert and Varadhan, 2019) than to optimize the objective function. Both bootstrapping and sandwich estimation can be made even more efficient through embarrassing parallelization, which is supported by our R package. Running-time comparisons are provided below in Section 6.

4.5 A two-stage semiparametric approach for amounts and balances

For the amount and balance levels of measurement our R package supports a two-stage semiparametric method (SMP). In the first stage one estimates F nonparametrically. The empirical distribution function $\hat{F}_n(y) = n^{-1}\sum_i 1\{Y_i \le y\}$ is a natural choice for our estimator of F, but other sensible choices exist. For example, one might employ the Winsorized estimator

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

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 $\boldsymbol{Y}_j^*$.

Although this approach may be necessary when the marginal distribution does not appear to take a familiar form, two-stage estimation can have a significant drawback, even for larger samples. If agreement is high, 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.

5 Bayesian inference for continuous scores

Since the Sklar's ω likelihood is not available in the case of discrete (i.e., nominal, ordinal, or count) 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), the performance of which was recently investigated for direct Gaussian copula models by Henn (2021), 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.

Package `sklarsomega` does support Bayesian inference for continuous scores, however. As we mentioned above, the package currently supports beta, gamma, Gaussian, Kumaraswamy, Laplace, and noncentral t marginal distributions for continuous 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

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

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 $\textsc{Gamma}(0.01, 0.01)$ 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. Consider parameter α, for example. To generate a proposal for α, we begin by drawing $\eta^* = \eta + \textsc{Normal}(0, \sigma_1)$, 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

$$\frac{\textsc{Lognormal}\{\alpha;\ \log(\alpha^*), \sigma_1\}}{\textsc{Lognormal}\{\alpha^*;\ \log(\alpha), \sigma_1\}}$$

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, 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: $\mu^* = \mu + \textsc{Normal}(0, \sigma_j)$, where $j = 1$ if the marginal distribution is Gaussian or Laplace, or $j = 2$ if the marginal distribution is $\mathrm{T}(\nu, \mu)$. The acceptance rate can be modulated via control parameter `sigma.1` (for a Normal or Laplace marginal) or `sigma.2` (for a 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, $\eta_k^* = \eta_k + \textsc{Normal}(0, \sigma_{\omega_k})$. 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, e.g., `control = list(sigma.omega = c(0.1, 0.1, 0.3))` (in the case of $\dim(\boldsymbol{\omega}) = 3$).

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. In the fixed-width approach, one chooses a small positive threshold ϵ and terminates sampling when all estimated coefficients of variation are smaller than said threshold, where the estimated coefficient of variation for parameter θ_j is $\widehat{\text{cv}}_j = \text{mcse}(\hat{\theta}_j)/|\hat{\theta}_j|$, with 'mcse' denoting Monte Carlo standard error. That is, sampling terminates when $\widehat{\text{cv}}_j < \epsilon$ 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, the sampler computes Monte Carlo standard errors (using package `mcmcse` (Flegal et al., 2021)) infrequently.

6 Application to simulated data

To investigate the performance of Sklar's ω, and ω's performance relative to Krippendorff's α, we applied both methods to simulated outcomes. The study plan is shown in Table 3. We carried out a study for each level of measurement, for various realistic sample geometries ($n_u = 15, n_c = 3$; $n_u = 30, n_c = 3$; and $n_u = 15, n_c = 6$), and for a few consequential values of ω (0.6, 0.7, 0.8, 0.85, 0.9, and 0.95). The chosen sample geometries allow us to reveal changes in performance as both the number of units n_u and the number of coders n_c vary. The values for ω were chosen at random, but the dependence strength ranges from substantial to strong, which makes inference challenging for all of the procedures. The values of the marginal parameters also pose challenges. Specifically, the marginal distributions for the count, amount, and percentage levels of measurement are skewed; the marginal distribution for the balance level of measurement has heavier-than-Gaussian tails; and the marginal distribution for the first scenario has a small number of categories.

As we mentioned earlier, the first scenario necessitates CML estimation owing to the small number of categories, and our R package supports only bootstrap inference for the CML method. The purpose of the second scenario is to show that DT-based inference tends to be poor for categorical data even when the number of categories is larger. The DT approach performs very well for counts, however, which are represented in the third scenario. For the DT approach we computed both bootstrap intervals and sandwich intervals. For the fourth, fifth, and sixth scenarios we applied the method of maximum likelihood, and computed both bootstrap intervals and asymptotic intervals. For the amount and balance levels of measurement we also investigated the performance of the semiparametric method described in Section 4.5. For the SMP method only bootstrap interval estimation is supported. We used a bootstrap sample size of 1,000 throughout.

We applied Krippendorff's α—using R package `krippendorffsalpha` (Hughes, 2021a)—and our procedure to each of 500 simulated datasets for each of the three sample geometries within each scenario, and so we analyzed $500 \cdot 3 \cdot 6 = 9{,}000$ datasets in total.

It is important to note which distance functions we chose for Krippendorff's α. For the nominal/ordinal levels of measurement we used the discrete metric $d(x, y) = 1\{x \neq y\}$. For counts and amounts we used the ratio metric

$$d^2(x, y) = \left(\frac{x - y}{x + y}\right)^2.$$

For balances we used squared Euclidean distance: $d^2(x, y) = (x - y)^2$. And for percentages we employed the bipolar distance function

$$d^2(x, y) = \frac{(x - y)^2}{(x + y + 2a)(2b - x - y)}$$

with $a = 0$ and $b = 1$. Krippendorff (2013) informed these choices.

The results are shown in Table 4. We report the median estimate, the percent bias, the variance, the mean squared error (MSE), coverage rates for 95% intervals, and average running times. When two coverage rates or average running times appear in a single cell of the table, the first value is for bootstrapping, the second for asymptotic inference. All of the intervals for α are bootstrap intervals. The running times are for a 3.6 GHz 10-core Intel Core i9 CPU. We parallelized the code (Tierney et al., 2018) for bootstrapping and sandwich estimation.

Table 3 Our simulation scenarios.

Scenario	Level of Measurement	Marginal Distribution	ω	Inference Method	Interval Type
1	nominal/ordinal	Categorical$(0.2, 0.5, 0.3)$	0.8	CML	bootstrap
2	nominal/ordinal	Categorical$(0.1, 0.3, 0.2, 0.05, 0.35)$	0.7	DT	bootstrap, sandwich
3	count	Poisson(3)	0.9	DT	bootstrap, sandwich
4	amount	Gamma$(2, 0.5)$	0.85	ML SMP	bootstrap, asymptotic bootstrap
5	balance	Laplace$(12, 10)$	0.95	ML SMP	bootstrap, asymptotic bootstrap
6	percentage	Beta$(4, 1)$	0.6	ML	bootstrap, asymptotic

For Scenario 1 we see that the CML approach for Sklar's ω performed very well and bested Krippendorff's α in every respect except running time. It appears that Krippendorff's α is too stringent for categorical scores (when Sklar's ω is the data-generating mechanism). For $\omega = 0.8$ we obtained $\hat{\alpha} \approx 0.5$ for all three sample geometries. Thus α analyses would lead us to conclude that agreement was merely moderate when, in fact, agreement was substantial or nearly perfect. This is not surprising given that α employs the discrete metric for nominal outcomes and, being nonparametric, must take the scores at face value, so to speak. Sklar's ω, by contrast, models agreement as a latent construct and so is not unduly influenced by marginal variation.

Neither procedure performed well in Scenario 2. Krippendorff's α performed poorly in this scenario for the same reasons as in Scenario 1. And the DT approach for Sklar's ω struggled to perform well because the marginal distribution was categorical. Indeed, Scenario 2 was included in the study precisely to show that DT-based inference, although considerably better than α-based inference, is somewhat poor for categorical outcomes, even for a larger number of categories. This is why package `sklarsomega` supports only the CML approach for categorical scores.

The DT approach shined in Scenario 3, though, as expected. The DT-based approximation is known to perform well for counts, and is even practically exact for some variants of the direct Gaussian copula model with count marginals (see Hughes (2021b) for details). The DT approach did underperform in one respect for this scenario, however: bootstrap confidence intervals did not have the desired 95% coverage rate. This is because the bootstrap distribution is slightly biased downward. This deficiency can be remedied by iterating the bootstrap or using the Gaussian method to compute bootstrap intervals. Alternatively, one can use sandwich intervals, which have better than 95% coverage and permit more efficient computation.

Krippendorff's α performed much better for counts than for nominal scores, yet α's performance still fell far short of ω's for Scenario 3. In this setting α tends to mistake near-perfect agreement for merely substantial agreement.

The results for simulation Scenarios 4 and 5, along with the analysis of the cartilage data presented in Section 7.2, allow us to draw an easy conclusion about the two-stage semiparametric procedure for Sklar's ω: the SMP approach performs well only for large samples, and is somewhat inferior to Krippendorff's α for small to medium sized samples. Since the cartilage dataset considered later in Section 7.2 is large, we will see that the SMP method produces plausible results for those data. For the data sizes employed in our simulation study, the SMP method clearly underperformed.

Finally, the results for Scenarios 4, 5, and 6 show that the method of maximum likelihood performs very well for amounts, balances, and percentages. For those levels of measurement the ML method for Sklar's ω clearly bested α in every respect except running time. Nearly all of the coverage rates for the ML approach were at or near the desired 95%, with the asymptotic intervals performing a bit better than the bootstrap intervals overall.

The running time for Krippendorff's α is considerably shorter than the running time for ω,

unless the ML method with asymptotic confidence interval is used for ω. This is not surprising since computation of $\hat{\alpha}$ does not require function optimization.

The running time for the Bayesian procedure described in Section 5 prohibited the procedure's inclusion in the simulation studies. But small auxiliary simulation studies revealed that the Bayesian method tends to yield inference that is quite similar to ML inference owing to the use of non-informative prior distributions.

7 Application to experimentally observed data

In this section we present two case studies. In the first study we apply both Sklar's ω and Krippendorff's α to nominal data previously analyzed by Krippendorff. In the second study we apply both methods to magnetic resonance imaging data of human hip cartilage. These studies show that Sklar's ω permits more nuanced analyses than does Krippendorff's α, and may lead to more plausible conclusions.

7.1 Application to nominal data analyzed previously by Krippendorff

Consider the data shown in Figure 1, which appear in (Krippendorff, 2013). (We display the data because visual inspection is necessary for understanding the analysis that follows.) These are nominal values (in $\{1, \ldots, 5\}$) for twelve units and four coders. The dots represent missing values.

Note that the scores for all units 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 $\alpha \in (0.39, 1.00)$ for a bootstrap sample size of 1,000. This point estimate 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 $\omega \in (0.70, 0.98)$, 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:

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

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 is not surprising in light of our simulation studies and 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.

7.2 Application to continuous data from an imaging study of hip cartilage

In this section we analyze experimentally observed bioimaging data to demonstrate how Sklar's ω can permit richer and more nuanced analyses of continuous data than Krippendorff's α can. The data for this application, which are included in R package `sklarsomega`, are 323 pairs of T2* relaxation times (a magnetic resonance quantity) for femoral cartilage (Nissi et al., 2015) in patients with femoroacetabular impingement (Figure 2), 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. We note that two subsequent studies, both of

	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	•

Fig. 1 Nominal scores previously analyzed by Krippendorff, for twelve units and four coders. The dots represent missing values.

which compared T2* measurements to gold standard arthroscopic evaluations, established the usefulness of T2* for assessing cartilage health (Henn et al., 2017; Morgan et al., 2018).

Prior to analyzing the T2* data we produced the Bland–Altman plot (Altman and Bland, 1983) shown in Figure 3. The plot suggests good agreement: small bias, no trend, consistent variability.

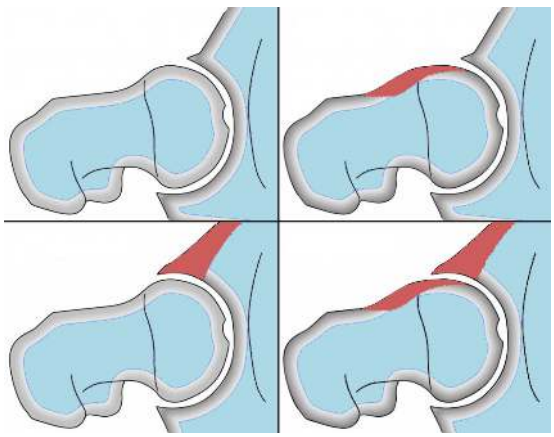

Fig. 2 An illustration of femoroacetabular impingement (FAI). Top left: normal hip joint. Top right: cam type FAI (deformed femoral head). Bottom left: pincer type FAI (deformed acetabular rim). Bottom right: mixed type (both deformities).

Because T2* is a relaxation time, the amount level of measurement is most appropriate for the cartilage data. This implies a gamma marginal distribution for Sklar's ω and the ratio distance function for Krippendorff's α. Since the data are continuous and the sample is large, we also applied the semiparametric version of Sklar's ω. For the gamma model we computed both asymptotic and bootstrap intervals. For the semiparametric model and Krippendorff's α we computed bootstrap intervals. The bootstrap sample size was 1,000 for all procedures.

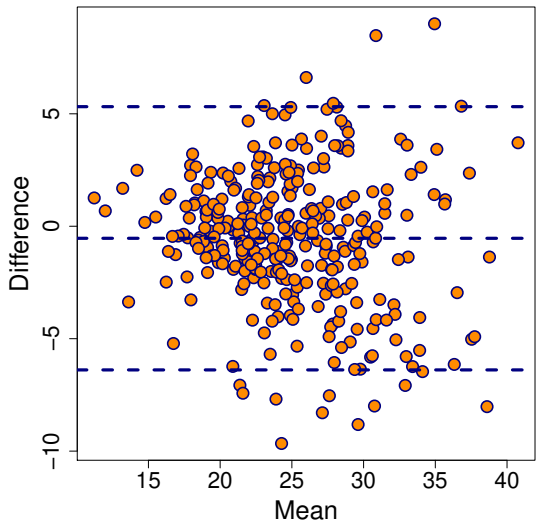

Fig. 3 A Bland–Altman plot for the femoral cartilage data.

Although T2* values are amounts rather than balances, we also applied Sklar's ω with a noncentral t marginal distribution, for the sake of comparison with the gamma fit. Since the T2* outcomes are noticeably right skewed (see Figure 4), we did not apply either the Gaussian model or the Laplace model.

The results are shown in Table 5, where the fourth column provides values of Akaike's information criterion (AIC) (Akaike, 1974) for the gamma and t fits, and the final column shows model probabilities (Burnham et al., 2011).

We see that the estimates and intervals are roughly comparable for all of these methods because the sample size is large and the marginal distribution is not too far from Gaussian. Yet the gamma distribution is far superior to the t distribution in terms of model probabilities. Figure 4 provides visual corroboration: it is clear that the gamma fit (with estimated shape parameters of 22.1 and 0.885) proves more compelling than the t fit (with estimated degrees of freedom of 11.2 and

estimated noncentrality parameter 23.3), as the estimated gamma pdf is quite close to the kernel density estimate while the fitted t density imposes too much asymmetry.

In any case, we must conclude that there is near-perfect agreement between raw T2* and contrast-enhanced T2*. This finding has clinical significance since the use of gadolinium-based contrast agents (GBCAs) is not free of risk to patients, particularly pregnant women and patients having impaired kidney function. For additional information regarding the potential risks associated with the use of GBCAs, we refer the interested reader to the University of California, San Francisco's policy on contrast-enhanced magnetic resonance imaging: https://tinyurl.com/rwnes6ku.

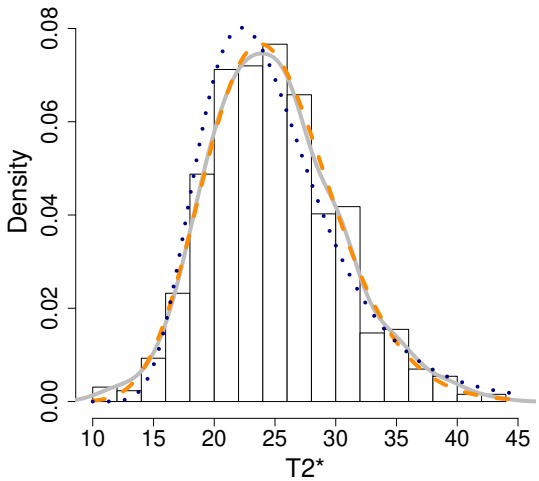

Fig. 4 For the T2* data: histogram, kernel density estimate, and fitted gamma and t densities. The solid (gray) curve is the kernel density estimate, the dashed (orange) curve is the fitted gamma density, and the dotted (blue) curve is the fitted t density. The fitted gamma density and the kernel density estimate are very close.

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 frequentist approach does not exist for the form of Sklar's ω presented in Section 3, wherein the copula is applied directly to the outcomes. We also developed Bayesian inference for continuous outcomes.

We demonstrated the advantages of Sklar's ω by pitting it against Krippendorff's α in an extensive simulation study. We also applied both approaches to previously analyzed nominal data and to magnetic resonance imaging data of hip cartilage. We envision ω's use in many other fields, e.g., sports, medical triage, social media.

As we mentioned in Section 4.5, when the outcomes are continuous, the marginal distribution is complicated, and dependence is strong, it may be desirable to estimate the marginal distribution and the copula parameter(s) simultaneously instead of applying a two-stage procedure. For example, Chen et al. (2004) developed a sieve maximum likelihood method for semiparametric copula models. Chen et al. (2004) studied two cases, one of which applies to Sklar's ω, namely, models for which the marginal distributions are equal but otherwise unspecified.

In addition to a sieve approach, other promising possibilities exist for continuous outcomes. For example, Szabó et al. (2007) described an approach they called "Gaussianization." In the first stage of this two-stage approach the outcomes for a given coder are ranked and then transformed as

$$\tilde{Z}_{ij} = \Phi^{-1}\left(\frac{R_{ij}}{n_u + 1}\right) \quad (i = 1, \ldots, n_u),$$

where R_{ij} is the rank of the ith unit for coder j. In the second stage one optimizes

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

to obtain $\hat{\boldsymbol{\omega}}$. Clearly, this approach is similar to the SMP method described in Section 4.5. And additional, similar approaches, based on Spearman's ρ (1904) and Kendall's τ (1938), were explored by Singh and Póczos (2017). These approaches

may be supported by future versions of package `sklarsomega`.

Another potential addition to package `sklarsomega` is support for Conway–Maxwell–Poisson (Huang, 2017; Sellers et al., 2012; Shmueli et al., 2005; Conway and Maxwell, 1962) marginal distributions for the count level of measurement. This would allow Sklar's ω to accommodate underdispersed as well as equidispersed and overdispersed counts.

Here we briefly introduce our R package, `sklarsomega`, version 3.0 of which is available for download from the Comprehensive R Archive Network.

Appendix A R package `sklarsomega`

We introduce our R package by way of a brief usage example. Additional examples are provided in the package documentation.

We apply our Bayesian methodology to a subset of the cartilage data, assuming first a $\textsc{Laplace}(\mu, \sigma)$ and then a $\mathrm{T}(\nu, \mu)$ marginal distribution. First we load the cartilage data, which are included in the package.

```
R> data(cartilage)
R> data = as.matrix(cartilage)[1:100, ]
R> colnames(data) = c("c.1.1", "c.2.1")
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 $\widehat{\mathrm{cv}}_j < 0.01$ for $j \in \{1, 2, 3\}$. As a second check we examine the plot given in Figure A1, 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.

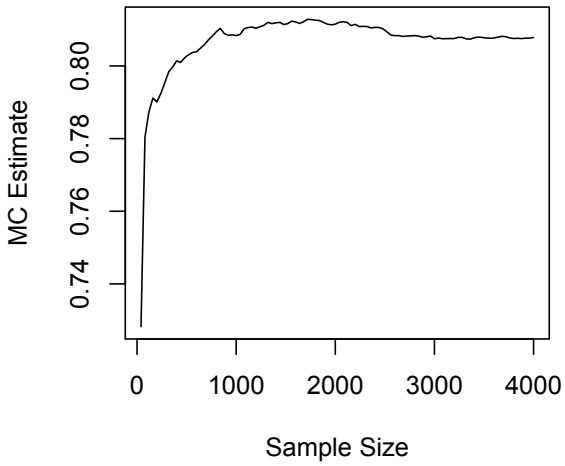

Fig. A1 A plot of estimated posterior mean versus sample size for ω, having assumed a Laplace marginal distribution.

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> 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
```

Much additional functionality is supported by package `sklarsomega`, e.g., plotting, simulation, influence statistics. 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) for the CML method. Future versions of the package will employ C++ (Eddelbuettel and Francois, 2011).

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.

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

Banerjee, M., Capozzoli, M., McSweeney, L., and Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, 27(1):3–23.

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.

Chen, X., Fan, Y., and Tsyrennikov, V. (2004). Efficient estimation of semiparametric multivariate copula models. Technical Report 04-W20, Vanderbilt University, Nashville, TN.

Chrisman, N. R. (1998). Rethinking levels of measurement for cartography. *Cartography and Geographic Information Systems*, 25(4):231–242.

Cicchetti, D. V. and Feinstein, A. R. (1990). High agreement but low kappa: II. resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6):551–558.

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

Conway, R. W. and Maxwell, W. L. (1962). Network dispatching by the shortest-operation discipline. *Operations Research*, 10(1):51–73.

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.

Feinstein, A. R. and Cicchetti, D. V. (1990). High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

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

Flegal, J. M., Hughes, J., Vats, D., Dai, N., Gupta, K., and Maji, U. (2021). *mcmcse: Monte Carlo Standard Errors for MCMC*. Riverside, CA, and Kanpur, India. R package version 1.5-0.

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.

Gilbert, P. and Varadhan, R. (2019). *numDeriv: Accurate Numerical Derivatives*. R package version 2016.8-1.1.

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

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Gwet, K. L. (2014). *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Advanced Analytics, LLC, Gaithersburg, MD, 4th edition.

Han, Z. and De Oliveira, V. (2016). On the correlation structure of Gaussian copula models for geostatistical count data. *Australian & New Zealand Journal of Statistics*.

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.

Henn, L. L. (2021). Limitations and performance of three approaches to Bayesian inference for Gaussian copula regression models of discrete data. *Computational Statistics*, pages 1–38.

Henn, L. L., Hughes, J., Iisakka, E., Ellermann, J., Mortazavi, S., Ziegler, C., Nissi, M. J., and Morgan, P. (2017). Disease severity classification using quantitative magnetic resonance imaging data of cartilage in femoroacetabular impingement. *Statistics in Medicine*, 36(9):1491–1505.

Hooke, R. and Jeeves, T. A. (1961). "Direct search" solution of numerical and statistical problems. *Journal of the ACM*, 8(2):212–229.

Huang, A. (2017). Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts. *Statistical Modelling*, 17(6):359–380.

Hughes, J. (2021a). krippendorffsalpha: An R package for measuring agreement using Krippendorff's Alpha coefficient. *The R Journal*,

13(1):413–425.

Hughes, J. (2021b). On the occasional exactness of the distributional transform approximation for direct Gaussian copula models with discrete margins. *Statistics & Probability Letters*, 177:109159.

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.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

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.

Morgan, P., Nissi, M. J., Hughes, J., Mortazavi, S., and Ellermann, J. (2018). T2* mapping provides information that is statistically comparable to an arthroscopic evaluation of acetabular cartilage. *Cartilage*, 9(3):237–240.

Mosteller, F. and Tukey, J. (1977). *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley series in behavioral science. Addison-Wesley Publishing Company.

Musgrove, D., Hughes, J., and Eberly, L. (2016). Hierarchical copula regression models for areal data. *Spatial Statistics*, 17:38–49.

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 (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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*, 19:321–325.

Sellers, K. F., Borle, S., and Shmueli, G. (2012). The COM–Poisson model for count data: a survey of methods and applications. *Applied Stochastic Models in Business and Industry*, 28(2):104–116.

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.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., and Boatwright, P. (2005). A useful distribution for fitting discrete data: Revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):127–142.

Singh, S. and Póczos, B. (2017). Nonparanormal information estimation. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th*

International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, pages 3210–3219. PMLR.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231.

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

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.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684):677–680.

Szabó, Z., Póczos, B., Szirtes, G., and Lőrincz, A. (2007). Post nonlinear independent subspace analysis. In *International Conference on Artificial Neural Networks*, pages 677–686. Springer.

Tierney, L., Rossini, A. J., Li, N., and Sevcikova, H. (2018). *snow: Simple Network of Workstations*. R package version 0.4-3.

Varadhan, R., University, J. H., Borchers, H. W., Research, A. C., Bechard, V., and Montreal, H. (2020). *dfoptim: Derivative-Free Optimization*. R package version 2020.10-1.

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.

Table 4 Results from our simulation study.

Scenario	n_u	n_c	Inference Method	Median Estimate	Bias	Variance	MSE	Coverage Rate	Average Time
1	15	3	CML	$\hat{\omega} = 0.792$	-3.8%	0.0160	0.0169	95%	132.6 s
			α	$\hat{\alpha} = 0.476$	-40.2%	0.0218	0.1254	34%	1.6 s
	30	3	CML	$\hat{\omega} = 0.795$	-1.9%	0.0073	0.0075	95%	253.0 s
			α	$\hat{\alpha} = 0.482$	-39.7%	0.0099	0.1107	7%	1.6 s
	15	6	CML	$\hat{\omega} = 0.763$	-5.5%	0.0113	0.0133	94%	328.5 s
			α	$\hat{\alpha} = 0.454$	-42.9%	0.0130	0.1307	9%	1.6 s
2	15	3	DT	$\hat{\omega} = 0.740$	2.9%	0.0133	0.0138	93%, 90%	138.7 s, 25.3 s
			α	$\hat{\alpha} = 0.269$	-59.9%	0.0134	0.1894	11%	1.6 s
	30	3	DT	$\hat{\omega} = 0.742$	4.9%	0.0054	0.0065	88%, 90%	261.3 s, 43.7 s
			α	$\hat{\alpha} = 0.291$	-58.3%	0.0066	0.1732	1%	1.7 s
	15	6	DT	$\hat{\omega} = 0.768$	8.3%	0.0060	0.0094	77%, 83%	296.2 s, 43.3 s
			α	$\hat{\alpha} = 0.287$	-59.1%	0.0073	0.1782	1%	1.6 s
3	15	3	DT	$\hat{\omega} = 0.885$	-2.2%	0.0011	0.0014	91%, 99%	12.6 s, 2.9 s
			α	$\hat{\alpha} = 0.706$	-22.9%	0.0225	0.0648	65%	1.5 s
	30	3	DT	$\hat{\omega} = 0.881$	-2.3%	0.0006	0.0010	70%, 98%	19.1 s, 4.5 s
			α	$\hat{\alpha} = 0.695$	-22.9%	0.0122	0.0545	43%	1.6 s
	15	6	DT	$\hat{\omega} = 0.880$	-2.4%	0.0005	0.0010	62%, 97%	18.8 s, 4.7 s
			α	$\hat{\alpha} = 0.697$	-23.2%	0.0140	0.0576	45%	1.6 s
4	15	3	ML	$\hat{\omega} = 0.845$	-2.6%	0.0054	0.0058	92%, 96%	21.5 s, 0.2 s
			SMP	$\hat{\omega} = 0.716$	-17.0%	0.0064	0.0273	0%	7.4 s
			α	$\hat{\alpha} = 0.782$	-9.9%	0.0063	0.0133	70%	1.6 s
	30	3	ML	$\hat{\omega} = 0.847$	-1.5%	0.0023	0.0024	94%, 95%	33.9 s, 0.4 s
			SMP	$\hat{\omega} = 0.789$	-8.1%	0.0030	0.0077	23%	11.4 s
			α	$\hat{\alpha} = 0.778$	-9.5%	0.0028	0.0093	52%	1.7 s
	15	6	ML	$\hat{\omega} = 0.837$	-2.8%	0.0043	0.0048	90%, 94%	34.6 s, 0.4 s
			SMP	$\hat{\omega} = 0.758$	-11.8%	0.0047	0.0148	3%	10.3 s
			α	$\hat{\alpha} = 0.767$	-10.6%	0.0043	0.0123	44%	1.7 s
5	15	3	ML	$\hat{\omega} = 0.948$	-0.7%	0.0006	0.0007	95%, 95%	47.8 s, 0.4 s
			SMP	$\hat{\omega} = 0.827$	-14.0%	0.0022	0.0198	0%	7.8 s
			α	$\hat{\alpha} = 0.941$	-1.7%	0.0010	0.0012	77%	1.6 s
	30	3	ML	$\hat{\omega} = 0.948$	-0.4%	0.0003	0.0003	95%, 95%	68.8 s, 0.6 s
			SMP	$\hat{\omega} = 0.895$	-6.1%	0.0006	0.0039	0%	12.0 s
			α	$\hat{\alpha} = 0.944$	-1.1%	0.0004	0.0005	77%	1.7 s
	15	6	ML	$\hat{\omega} = 0.946$	-0.9%	0.0005	0.0006	93%, 95%	76.8 s, 0.6 s
			SMP	$\hat{\omega} = 0.866$	-9.6%	0.0015	0.0099	0%	11.7 s
			α	$\hat{\alpha} = 0.937$	-2.1%	0.0009	0.0013	68%	1.7 s
6	15	3	ML	$\hat{\omega} = 0.583$	-6.2%	0.0200	0.0213	93%, 92%	16.6 s, 0.1 s
			α	$\hat{\alpha} = 0.549$	-11.8%	0.0197	0.0247	84%	1.6 s
	30	3	ML	$\hat{\omega} = 0.592$	-3.2%	0.0097	0.0101	94%, 93%	24.1 s, 0.3 s
			α	$\hat{\alpha} = 0.548$	-9.6%	0.0100	0.0133	85%	1.7 s
	15	6	ML	$\hat{\omega} = 0.566$	-7.1%	0.0123	0.0141	90%, 93%	25.6 s, 0.3 s
			α	$\hat{\alpha} = 0.524$	-13.3%	0.0122	0.0185	73%	1.7 s

Table 5 Results from applying Sklar's ω and Krippendorff's α to the femoral-cartilage data.

Marginal Model	Agreement	Interval	AIC	Model Probability
Gamma	$\hat{\omega} = 0.849$	bootstrap: $\omega \in (0.815, 0.878)$ ML: $\omega \in (0.819, 0.880)$	3,564	≈ 1
Noncentral t	$\hat{\omega} = 0.862$	bootstrap: $\omega \in (0.831, 0.888)$ ML: $\omega \in (0.834, 0.890)$	3,588	≈ 0
Empirical	$\hat{\omega} = 0.846$	$\omega \in (0.810, 0.870)$	–	–
α Ratio	$\hat{\alpha} = 0.850$	$\alpha \in (0.822, 0.874)$	–	–
α Interval	$\hat{\alpha} = 0.837$	$\alpha \in (0.806, 0.864)$	–	–