PARTIAL IDENTIFICATION OF LATENT CORRELATIONS WITH BINARY DATA

STEFFEN GRØNNEBERG, JONAS MOSS, AND NJÅL FOLDNES

Abstract. The tetrachoric correlation is a popular measure of association for binary data and estimates the correlation of an underlying normal latent vector. However, when the underlying vector is not normal the tetrachoric correlation will be different from the underlying correlation. Since assuming underlying normality is often done on pragmatic and not substantial grounds, the estimated tetrachoric correlation may therefore be quite different from the true underlying correlation that is modeled in structural equation modeling. This motivates studying the range of latent correlations that are compatible with given binary data, when the distribution of the latent vector is partly or completely unknown. We show that nothing can be said about the latent correlations unless we know more than what can be derived from the data. We identify an interval constituting all latent correlations compatible with observed data when the marginals of the latent variables are known. Also, we quantify how partial knowledge of the dependence structure of the latent variables affect the range of compatible latent correlations. Implications for tests of underlying normality are briefly discussed.

1. Introduction

An important class of statistical methods for samples from random vectors X with ordinal coordinates follows the perspective of Pearson (1900) by postulating a continuous random vector Z which when discretized produces X. The present study is concerned with the most simple case of ordinal variables, namely the binary case, where we observe samples from $X = (X_1, \ldots, X_d)$ obtained through the following discretization:

$$X_i = 1\{Z_i > \tau_i\}, \qquad i = 1, \ldots, d. \tag{1}$$

Here, $1\{\cdot\}$ is the indicator function, $Z = (Z_1, \ldots, Z_d)$ are latent variables, and $\tau_1, \ldots, \tau_d$ are fixed thresholds. In psychometrics, prominent methods that are based on the discretization framework are factor analysis (Christoffersson, 1975; Muthén, 1978), principal component analysis (Kolenikov & Angeles, 2009), and structural equation models (Jöreskog, 1994; Muthén, 1984), as well as some models usually formulated without direct reference to Z, such as multivariate item response theory models (Takane & de Leeuw, 1987). A crucial ingredient in these methods is the estimation of the covariance matrix Σ of Z. In the present article we investigate, for binary variables, what can be said about the covariance among the latent variables Z_i and Z_j when their joint distribution is not fully known.

It is instructive to contrast factor analysis and structural equation modeling with ordinal data to the approach taken with continuous data. Both approaches achieve parameter estimation by minimizing the distance between the model-implied covariance matrix $\Sigma(\theta)$ and an estimate

of the population covariance matrix. For continuous data, it is straightforward to consistently estimate the population covariance matrix, by computing directly from the data the sample covariance matrix S. The most common estimator for continuous data is normal theory based maximum likelihood (NTML). The likelihood assumes that the observed vector is drawn from a multivariate normal distribution. It would therefore seem that the normality assumption is crucial for NTML estimation and inference, given that maximum likelihood estimators are usually inconsistent if the probability distribution that observables are assumed to follow is misspecified (see e.g. Claeskens & Hjort, 2008, Chapter 2.2). However, NTML estimation for covariance models turns out to be contained within a class of moment based estimators known as minimum discrepancy function estimators (see e.g. Shapiro, 1983), and is therefore consistent and covered by a known inference theory even under non-normality as long as the covariance model itself holds. Due to this fortunate fact, NTML estimation is used in almost all applied work with covariance models, and is the standard estimation method of software packages such as `mplus` (Muthén & Muthén, 2017), `Lisrel` (Jöreskog & Sörbom, 1996), and `lavaan` (Rosseel, 2012).

In the ordinal case, it does not make sense to assume multivariate normality for the observed variables. Instead, the normality assumption has traditionally been made for the unobserved vector Z. By assuming that Z is normally distributed, an assumption originating from Pearson (1900), we may estimate its correlation matrix with polychoric correlations (Olsson, 1979), known as tetrachoric correlations in the binary case. The posited model is then fitted to the polychoric correlation matrix using minimum discrepancy methods. In contrast to the continuous case, the multivariate normality assumption is crucial in the ordinal case. Without this, or a similar distributional assumption regarding Z, we can not obtain a sample estimate of the covariance matrix Σ of Z. The reason is that with only the observed vector X at hand, the available information is limited, taking the form of a contingency table. In the bivariate binary case, the information is contained in a 2×2 table. Therefore, we must make strong assumptions on the distribution of Z in order to identify its correlations.

In the present paper our aim is to investigate what can be learned about the latent correlations when the normality assumption is relaxed: Based on observed data, and partial knowledge of the distribution of Z, what can be said about the correlations of Z? We demonstrate that these correlations will not be identified even under quite strong assumptions on the distribution of Z. This means there are several distributions for Z that are compatible with our knowledge, and that can generate X, and these distributions may have different correlations. Instead, we calculate intervals which contain all possible correlations compatible with observed data and our knowledge of the distribution of Z. In the continuous case, population correlations are always identified, and the consequences of relaxing the normality assumption for NTML is a well-studied problem, and several robust approaches (e.g., Satorra & Bentler, 1988) are available to conduct inference in a valid manner. In contrast, in the ordinal case the correlations are not even identified, which is the starting point for all classical statistical techniques. The distributional assumptions made on Z are often based on pragmatic considerations (this is also argued in Molenaar & Dolan, 2018), and not on what we will call substantial knowledge of the

phenomena involved. In practice, it seems that the use of estimation methods which assume the normality of Z is often based on an earlier consensus that normal theory methods are fairly robust against underlying non-normality. This consensus, based on earlier simulation studies, was questioned in Foldnes and Grønneberg (2019a, 2019b, 2020), who used the non-normal simulation method of Grønneberg and Foldnes (2017) to argue that normal theory methods are not as robust as previously thought. Substantial knowledge of the distribution of Z is therefore required. This is in agreement with the discussion in Pearson and Heron (1913, p. 161–162).

We focus on the simplest case of two binary variables, summarized by a 2×2 table. We relax the normality assumption for the joint distribution of the two underlying continuous variables, and ask what can be known about their correlation when the joint distribution is completely or partially unknown. Such an analysis of parameter sets compatible with the observed data is known as partial identification, and has a long history in statistics and econometrics (Manski, 2003; Tamer, 2010), but is to the best of our knowledge hitherto not used in psychometrics. If we can establish a rather narrow band of possible correlations, this would mean that we may estimate the parameters in θ with at least some degree of certainty. If, on the other hand the set of possible correlations that are compatible with the 2×2 table is wide, we can not proceed to estimate our model without imposing further restrictions on the distribution of Z. If such is the case, the validity of our statistical analysis will depend crucially on the normality assumption, and steps must be taken to test this assumption prior to further analysis together with strong reasons why we would expect Z to be multivariate normal. There are various tests for the distributional assumptions made on Z (e.g., Foldnes & Grønneberg, 2019b; Maydeu-Olivares, 2006).

The remainder of this article is organized as follows. In Section 2.1 we show for the bivariate case that nothing can be said about the correlation of Z unless we take into account substantial knowledge of the distribution of Z, that is, knowledge not derivable from the distribution of the observations X. In Section 2.2, we assume substantial knowledge justifies treating the marginal distributions as known, and identify a set which contains all possible Pearson correlations of Z that are compatible with observed data. A similar analysis is done for Spearman's rho. For Spearman's rho, the resulting sets have, in contrast to the Pearson correlation, lengths less than two also if nothing is known about the distribution of Z. Unfortunately, these sets are always so wide that they contain little to no practical information. In Section 2.3 we illustrate in a simple setting with known marginals that a partial identification analysis of latent correlations can be used to provide a partial identification analysis of latent correlation models. In Section 2.4, we study the case when marginals are known, and a rectangle of the cumulative distribution function of the copula is also known, where the rectangle includes the point of the copula which is shown in Section 2.1 to be deducible from the distribution of X. We interpolate between knowing only this point, which leads to extraordinarily spacious intervals, to fully knowing the copula of Z, which point identifies the latent correlation. In Section 2.5, a partial identification analysis is performed when Z_2 is directly observed, and Z_1 is observed via a binary discretized variable, but has a known marginal distribution. When the full distribution of Z is assumed to be normal, this is the setting of the biserial correlation of Pearson and Pearson (1922).

In Section 3.2 we show that without substantial knowledge, multivariate information cannot help identify the pairwise correlations of Z. In Section 3.3, we discuss tests for underlying normality in light of our results. Our study only derives partial identification sets for a single latent correlation, and Section 3.4 discusses the limitations springing from this focus. Some concluding remarks are given in Section 4.

We ignore sampling error in the paper. The partially identified sets we calculate are intervals, where inference can easily be dealt with when observing independent and identically distributed data (Tamer, 2010, Section 4.4). Proofs of all results are found in Appendix A. The online supplementary material includes an online appendix with additional technical details, as well as several `R` scripts.

2. Partial identification with 2×2 tables

The starting point for most statistical theory is that the parameters of interest are point-identified. This is often achieved only under strong assumptions, and some of these assumptions may be questionable. Partial identification analysis calculates the set of possible parameter values attainable under the subset of assumptions that are seen as unquestionable. An immediate application is a form of sensitivity analysis (Tamer, 2010, Section 1), as the size and shape of the resulting set gives information on the influence from the more questionable assumptions. Tamer (2010) contains a literature review of partial identification while the book Manski (2003) is an introduction to the field.

We briefly summarise the Fréchet–Höffding bounds and the partial identification analysis of the Pearson correlation when only the marginal distributions are assumed known, but the full distribution is not known. This may occur if we have studied two phenomena separately, but not jointly. This partial identification problem was solved by Höffding (1940) and Fréchet (1960), with a modern presentation in Nelsen's book (2007). See also the influential papers by Lehmann (1966) and Whitt (1976). Our results are generalizations of the arguments underlying the Fréchet–Höffding bounds argument.

Suppose F is a bivariate cumulative distribution function with marginal distributions F_1, F_2. Recall that a copula C is a cumulative distribution function with uniform marginals on $[0,1]$. According to Sklar's theorem (Nelsen, 2007; Sklar, 1959, Theorem 2.3.3), there exists a copula C such that for any x_1, x_2 we have

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2)), \tag{2}$$

where the copula is unique on the range of F_1, F_2, and therefore unique if F_1, F_2 are continuous. Moreover, if C is a copula and F_1, F_2 are univariate cumulative distribution functions, then F defined by eq. (2) is a cumulative distribution function with marginals F_1, F_2. The Fréchet–Höffding bound (Nelsen, 2007, Theorem 2.2.3) states that any copula C fulfils $W(u,v) \leq C(u,v) \leq M(u,v)$ for all $u, v \in [0,1]$, where W, M are the copulas $W(u,v) = \max(u+v-1, 0)$ and $M(u,v) = \min(u,v)$. Sklar's theorem implies that for $W[F_1, F_2](x_1, x_2) = W(F_1(x_1), F_2(x_2))$ and $M[F_1, F_2](x_1, x_2) = M(F_1(x_1), F_2(x_2))$, both $W[F_1, F_2]$ and $M[F_1, F_2]$

are distribution functions with marginals F_1, F_2. The Fréchet–Höffding bound gives

$$W[F_1, F_2](x_1, x_2) \leq F(x_1, x_2) \leq M[F_1, F_2](x_1, x_2) \tag{3}$$

for all x_1, x_2. Since the upper and lower bounds are themselves distribution functions with marginals F_1, F_2, this bound cannot be improved.

Let $\rho(F)$ denote the Pearson correlation of F when F is a distribution function. The Höffding (1940) formula for the correlation states that

$$\rho(F) = \operatorname{sd}(F_1)^{-1} \operatorname{sd}(F_2)^{-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(z_1, z_2) - F_1(z_1) F_2(z_2) \, \mathrm{d}z_1 \mathrm{d}z_2, \tag{4}$$

where $\operatorname{sd}(F_1), \operatorname{sd}(F_2)$ are the standard deviations of F_1, F_2, the marginals of F.

For a set $\mathcal{P}$ of bivariate distributions with finite standard deviations, define $\rho(\mathcal{P}) = \{\rho(F) : F \in \mathcal{P}\}$. Let $\mathcal{P}$ be the set of distributions with fixed marginals F_1, F_2 and let $F \in \mathcal{P}$. The Fréchet–Höffding bounds (3) implies that

$$\rho(F) \in [\rho(W[F_1, F_2]), \rho(M[F_1, F_2])].$$

An argument based on convex combinations of the boundary distributions shows that $\rho(\mathcal{P}) = [\rho(W[F_1, F_2]), \rho(M[F_1, F_2])]$, see the proof of Proposition 1 for details.

2.1. **Latent correlations in 2×2 tables.** Now we will handle the discretization model (1) in the bivariate case. Let $Z = (Z_1, Z_2)$ be a bivariate latent variable with distribution function F. Denote its marginal distribution functions by F_1, F_2, and its copula by C. The distribution of X is parametrised by the 2×2 table

$$p = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}, \quad p_{x_1 x_2} = \mathrm{P}(X_1 = x_1, X_2 = x_2).$$

Here $x_1, x_2 \in \{0, 1\}$ are the inputs to a distribution function, as in eq. (2), though in a different domain. We ignore sampling error, and therefore assume that p is known.

We have $\mathrm{P}(X_1 = 0) = \mathrm{P}(Z_1 \leq \tau_1) = F_1(\tau_1)$ and $\mathrm{P}(X_2 = 0) = F_2(\tau_2)$. Therefore, if F_1, F_2 are specified, we get the simple relationship $\tau_1 = F_1^{-1}(p_{01} + p_{00})$ and $\tau_2 = F_2^{-1}(p_{10} + p_{00})$. Without specifying F_1, F_2, nothing can be said about τ_1, τ_2, as only $F_1(\tau_1)$ and $F_2(\tau_2)$ are identified. From the remaining degree of freedom in p, we can derive a restriction on C, the copula of Z. From Sklar's theorem (2) we get

$$p_{00} = \mathrm{P}(Z_1 \leq \tau_1, Z_2 \leq \tau_2) = C[F_1(\tau_1), F_2(\tau_2)] = C[p_{01} + p_{00}, p_{10} + p_{00}]. \tag{5}$$

We are interested in the correlation of Z. This latent correlation is not unique as a function of p unless we place restrictions on the family of distributions for Z. Let $\mathcal{P}$ be a family of probability measures over Z with finite standard deviations. Define the set $\rho(\mathcal{P}; p)$ as the set of latent correlations compatible with p and $\mathcal{P}$. That is,

$$\rho(\mathcal{P}; p) = \{\rho(F) : F \in \mathcal{P}, C_F[p_{01} + p_{00}, p_{10} + p_{00}] = p_{0,0}\} \tag{6}$$

where C_F is the copula of F.

Assume $\mathcal{P}$ is the class of bivariate normal distributions, as done by Pearson (1900). In this case the latent correlation is called the tetrachoric correlation. By a change in threshold

values, we may assume that the marginals are standard normal (Pearson, 1900, eq. (i)-(v)). By Sklar's theorem, $\mathcal{P} = \{C_\rho(\Phi(x_1), \Phi(x_2)) : -1 \leq \rho \leq 1\}$ where Φ is the standard normal cumulative distribution function and C_ρ is the normal copula parametrised by the correlation ρ. From Joe (1997, Section 5.1) and Almeida and Mouchart (2014), we know that $\rho \mapsto C_\rho(u, v)$ is strictly increasing for $0 < u, v < 1$. The tetrachoric correlation is therefore point-identified and solves $C_\rho[p_{01} + p_{00}, p_{10} + p_{00}] = p_{0,0}$. As noted by Almeida and Mouchart (2014), the same argument can yield identifiability when assuming other marginals and other one-dimensional parametric copula classes $\{C_\theta : \theta \in \Theta\}$. We only require that $\theta \mapsto C_\theta(u, v)$ is increasing for each $0 < u, v < 1$, a property fulfilled by many copulas classes, for instance those catalogued in Section 5.1 of Joe (1997).

Theorem 1 calculates $\rho(\mathcal{P}; p)$ when we place no restrictions on $\mathcal{P}$, see page 15 of the appendix for the proof.

Theorem 1. *Suppose $\mathcal{P}$ contains all probability distributions. If none of the elements of p are zero, then $\rho(\mathcal{P}; p) = (-1, 1)$.*

Pearson's correlation depends on the marginals of Z as well as the copula of Z. While equation (5) gives a restriction on the copula of Z, the marginals of Z are unrestricted, and this is what we use to show Theorem 1. In contrast, Spearman's rho, a copula dependency measure, has partially identified sets with lengths less than two, even when nothing is known of the distribution of Z, as we will see in the upcoming Proposition 2.

2.2. **Partial identification for given latent marginals.** From Theorem 1, the point identification of the latent correlation depends crucially on assumptions on the distribution of Z. As discussed by Pearson and Heron (1913), such assumptions must be justified by external information on the variable Z. Let us suppose that relevant external information is available, but that this only specifies the marginal distributions F_1, F_2 and not the full distribution F. Practically, this may occur in situations where the coordinates of Z have been studied separately, and from this the likely distribution can be deduced, but the joint distribution is unknown.

Proposition 1. *Let $\mathcal{P}$ be the set of distributions with marginals F_1, F_2. Then*

$$\rho(\mathcal{P}; p) = [\rho(W[F_1, F_2; p]), \rho(M[F_1, F_2; p])],$$

where $\rho(\mathcal{P}; p)$ is defined in equation (6). Here $M[F_1, F_2; p](x_1, x_2) = M_p(F_1(x_1), F_2(x_2))$ and $W[F_1, F_2; p](x_1, x_2) = W_p(F_1(x_1), F_2(x_2))$ are defined in terms of the copulas

$$\begin{aligned} M_p(u, v) &= \min\left\{u, v, p_{00} + (u - p_{01} - p_{00})^+ + (v - p_{10} - p_{00})^+\right\}, \\ W_p(u, v) &= \max\left\{0, u + v - 1, p_{00} - (p_{01} + p_{00} - u)^+ - (p_{10} + p_{00} - v)^+\right\}. \end{aligned}$$

The proof of Proposition 1 is in Appendix A, page 16.

An important class of applications of normal theory tetrachoric correlations is factor analysis for ordinal data, as well as more general structural equation models. Since the Pearson

correlation depends on the marginal distributions of Z, normal marginals are of special interest as this is the marginal scale of standard methodology.

Computational considerations for how to apply Proposition 1 are given in Appendix A.5, page 22, including considerable computational simplifications when the marginals are normal.

Spearman's rho is the Pearson correlation of a copula (Nelsen, 2007, Section 5.1.2), and is therefore not dependent on the unidentified marginals. Let $\mathcal{R}(p)$ be the set of Spearman's rho values compatible with p. We therefore have $\mathcal{R}(p) = \rho(\mathcal{P}, p)$ when $\mathcal{P}$ is the set of distributions with uniform marginals on $[0, 1]$. We identify the following compact algebraic formula, proved on page 17 in Appendix A.

Proposition 2. *We have* $\mathcal{R}(p) = [6p_{00}p_{11}(p_{00} + p_{11}) - 1, 1 - 6p_{01}p_{10}(p_{01} + p_{10})]$.

2.3. **Some illustrations.** Let the distribution of the binary vector (X_1, X_2) be given by the 2×2 table

$$p = \begin{bmatrix} 0.2 & 0.4 \\ 0.1 & 0.3 \end{bmatrix}. \tag{7}$$

Assuming Z has a normal copula, Spearman's rho is 0.14; assuming Z is bivariate normal, Pearson's correlation is 0.15; assuming normal marginals, $\rho(\mathcal{P}; p) = [-0.88, 0.93]$. The interval of compatible correlations is very wide. On the other hand, if the distribution of Z is totally unknown, $\rho(\mathcal{P}; p) = (-1, 1)$ by Theorem 1 and $\mathcal{R}(p) = [-0.82, 0.88]$ by Proposition 2.

The skew-t family (Azzalini, 2013, Section 4.3) is commonly used to model skewed and heavy-tailed data. For instance, the multivariate skew-t has been studied in structural equation models by Asparouhov and Muthén (2016). In addition to the degree of freedom parameter ν from the t-distribution, it is parametrised by the skewness parameter α. The distribution is skewed to the right if $\alpha > 0$, skewed to the left if $\alpha < 0$, and symmetric if $\alpha = 0$, with the degree of skewness increasing with the absolute value of α. When $\alpha = 0$ and $\nu = \infty$ the distribution is normal.

To investigate how skewness and heavy tails influence the length of the bounds, we calculated the length of the bounds for $\alpha \in (-20, 20)$ and $\nu \in (3, 100)$ when both marginals have the same distribution. The underlying 2×2 table is p in eq. (7). The lengths range from 1.95 to 1.76, with heavier tails (smaller ν) and negative skew being associated with intervals of larger lengths (the length with normal marginals is 1.81). All these lengths are too large to be useful, but the maximal difference of approximately 0.20 is larger than expected. As seen in Figure 1, the relationship between α, ν and the length is quite complicated.

For more illustrative examples, see Appendix A, page 14.

Now we illustrate how the partially identified interval above leads to a partial identification analysis for a factor model. Since the present article has a bivariate focus, our illustrative factor model is necessarily simple. Suppose (Z_1, Z_2) follows the congeneric measurement model

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} \xi + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}$$

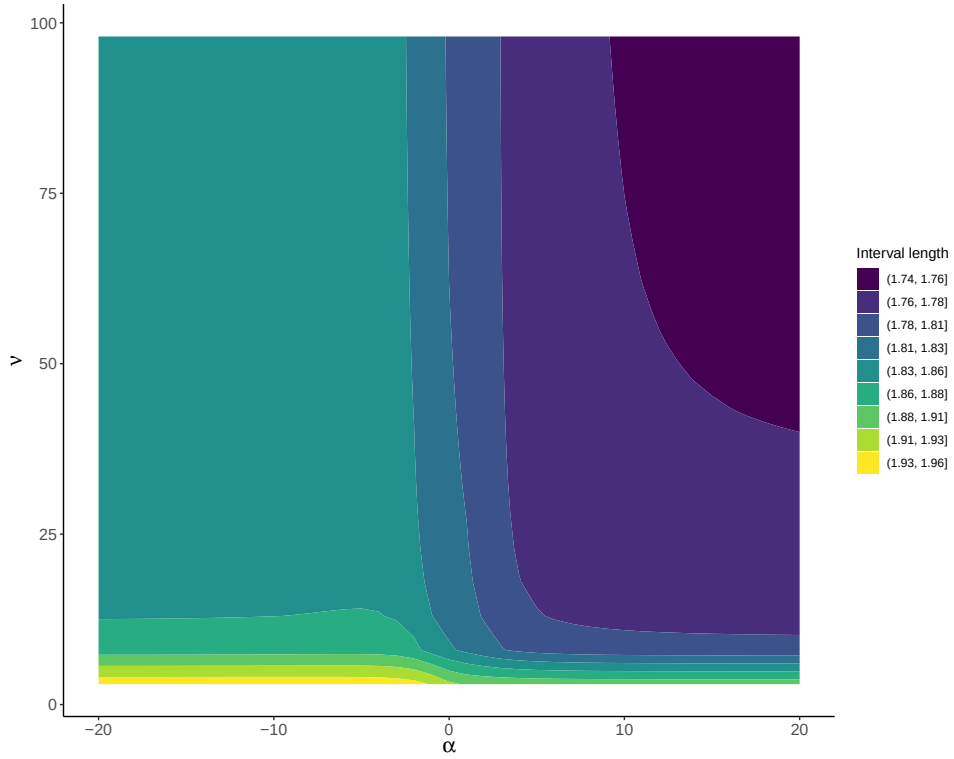

FIGURE 1. Lengths of the identification intervals for the skew-t distribution as a function of the skewness α and degrees of freedom ν.

where ξ is a one-dimensional variable with unit variance, ϵ_1, ϵ_2 are error terms, mutually uncorrelated, and uncorrelated with ξ.

To identify the parameter vector $\theta = (\lambda_1, \lambda_2, \sigma)$ as a function of ρ, we assume that Z_1 and Z_2 have unit variance, that the error variances are identical and equal to, say, σ^2, and that $\lambda_1 \geq 0$. We get

$$\operatorname{Cov}\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \operatorname{Cor}\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} \lambda_1^2 + \sigma^2 & \lambda_1\lambda_2 \\ \lambda_1\lambda_2 & \lambda_2^2 + \sigma^2 \end{bmatrix}.$$

From $\lambda_1^2 + \sigma^2 = \lambda_2^2 + \sigma^2 = 1$ and $\lambda_1\lambda_2 = \rho$ we get $\lambda_1 = \sqrt{|\rho|}$, $\lambda_2 = \operatorname{sign}(\rho)\lambda_1$, and $\sigma = \sqrt{1 - |\rho|}$. Given the identification interval $\rho(\mathcal{P}; p) = [-0.88, 0.93]$, the joint identification region for $(\lambda_1, \lambda_2, \sigma)$ becomes

$$\begin{aligned} H(\lambda_1, \lambda_2, \sigma; \rho(\mathcal{P}; p)) = &\{(\sqrt{|\rho|}, -\sqrt{|\rho|}, \sqrt{1 - \rho^2}) \mid \rho \in \rho(\mathcal{P}; p) \cap [-1, 0]\} \\ &\cup \{(\sqrt{\rho}, \sqrt{\rho}, \sqrt{1 - \rho^2}) \mid \rho \in \rho(\mathcal{P}; p) \cap [0, 1]\}. \end{aligned}$$

Figure 2 shows the joint partial identification region for $(\lambda_1, \lambda_2, \sigma)$, a union of two curves. Under bivariate normality, the tetrachoric correlation is $\rho = 0.15$, and $(\lambda_1, \lambda_2, \sigma) = (0.39, 0.39, 0.92)$.

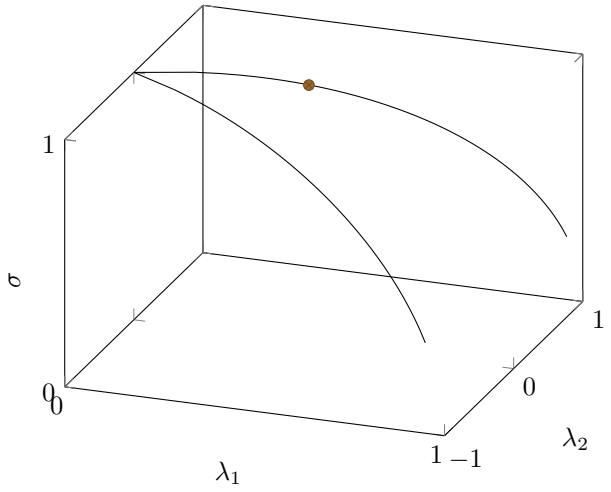

FIGURE 2. Joint partial identification region for $(\lambda_1, \lambda_2, \sigma)$ when $\rho(\mathcal{P}; p) = [-0.83, 93]$. The point at $(0.39, 0.39, 0.92)$ corresponds to the fully identified parameter vector assuming Z is bivariate normal, where $\rho = 0.15$.

2.4. **Quantifying the effect of increasing the degree of knowledge of the latent distribution.** We consider a way to interpolate between only knowing the marginal distributions of Z to completely specifying the distribution of Z, studying how the sets of possible latent correlation values change from being exceedingly wide in the case when only marginals are known, to being point-identified when the full distribution of Z is known.

From Theorem 1 we know that we must be able to specify certain aspects of the distribution of Z in order to say anything about the latent correlation. We have hitherto only specified knowledge of the marginals, but other forms of knowledge may be relevant in some cases. The main ingredient for extending our result to such cases is optimal Fréchet–Höffding distributions that are compatible with what is known.

Now we study partial identification of latent correlations in the case when the marginals are known, and a rectangular region of the copula cumulative distribution function is known to equal the normal copula with a correlation compatible with the generated distribution of X.

From equation (5), we have

$$p_{00} = C(\tilde{u}, \tilde{v}), \qquad \tilde{u} = p_{01} + p_{00},\ \tilde{v} = p_{10} + p_{00}.$$

Let Q be the unique bivariate normal copula that is compatible with this restriction. We consider knowledge of C of the form

$$C(u, v) = Q(u, v) \text{ for all } (u, v) \in \mathcal{H},$$

where $\mathcal{H} = \{(\tilde{u} + \varepsilon_1, \tilde{v} + \varepsilon_2) : 0 \leq \tilde{u} + \varepsilon_1 \leq 1, 0 \leq \tilde{v} + \varepsilon_2 \leq 1, |\varepsilon_1| \leq r_1, |\varepsilon_2| \leq r_2\}$ for some numbers $r_1, r_2 \geq 0$. That is, we specify that we know the copula of Z exactly in a rectangular region.

Optimal Fréchet–Höffding distributions that are compatible with these restrictions are identified in Corollary 2.2 of Bernard, Jiang, and Vanduffel (2012). From the Höffding formula of equation (4), we get formulas for the partial identification interval of the latent correlation by plugging in the resulting upper and lower copulas. As in earlier cases, the property of agreeing with the normal copula on a rectangle is stable under convex combinations, meaning that all values in between the upper and lower correlation limits are attainable, and the partial identified set is an interval.

In Figure 3 we have numerically identified these intervals for the case when the marginals are standard normal, Q is the normal copula with correlation $1/2$, when $(\tilde{u}, \tilde{v}) = (1/2, 1/2)$, and when $r_1 = r_2$ is set to ϵ which varies in the region $[0, 1/2]$. When $\epsilon = 0$, we regain the bounds from Proposition 1. When $\epsilon = 1/2$, we have point-identified the latent correlation. We see that the upper and lower limits of the intervals converge towards each other at $1/2$ in a non-symmetric manner: The upper bound is closer to $1/2$ when $\epsilon = 0$ compared to the lower bound, and the upper bound therefore moves slower towards its endpoint compared to the lower bound. In summary, the figure shows how increasing knowledge of the latent distribution influences the length of the possible values of the latent correlation.

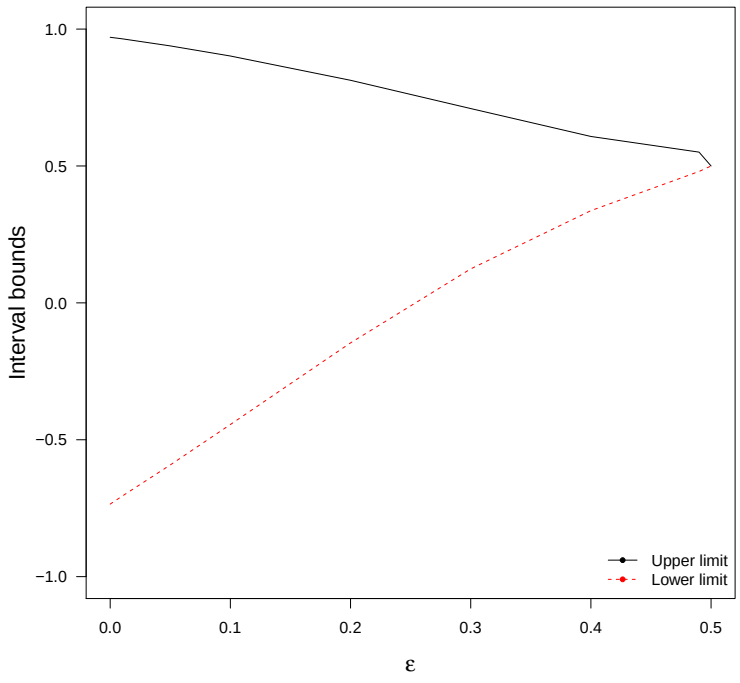

FIGURE 3. The partial identified interval for the latent correlation with growing knowledge of the latent distribution when X is compatible with being generated by an underlying normal variable with correlation 0.5.

2.5. **Partial identification when Z_2 is directly observed.** We now assume that Z_2 is directly observed. When Z is normal, this gives the biserial correlation of Pearson (1909), see also Tate (1955a, 1955b). That is, we observe $X = (1\{Z_1 > \tau_1\}, Z_2)'$. Let the distribution of X be denoted by p. From p, we deduce F_2, $F_1(\tau_1)$ and $C(F_1(\tau_1), v)$ for all v. But we can neither deduce F_1 nor the copula C. The latent correlation is therefore not identified from data alone. Define $\rho(\mathcal{P}; p)$ as the correlations of Z with distribution in $\mathcal{P}$ that can generate X. The next result builds on Tankov (2011). For compactness, we state it in terms of $C(F_1(\tau_1), \cdot)$ and $F_1(\tau_1)$ and not directly via the distribution of X.

Proposition 3. *Let $\mathcal{P}$ be the set of distributions with marginals F_1, F_2. Let $\rho(\mathcal{P}; p)$ be the set of every possible correlation of Z when Z has a distribution in $\mathcal{P}$ that can generate $X = (1\{Z_1 > \tau_1\}, Z_2)'$ with distribution p. Then*

$$\rho(\mathcal{P}; p) = [\rho(W[F_1, F_2; p]), \rho(M[F_1, F_2; p])].$$

Here $M[F_1, F_2; p](x_1, x_2) = M_p(F_1(x_1), F_2(x_2))$ and $W[F_1, F_2; p](x_1, x_2) = W_p(F_1(x_1), F_2(x_2))$ are defined in terms of the copulas

$$\begin{aligned} M_p(u, v) &= \min\left(u, v, C(F_1(\tau_1), v) + (u - F_1(\tau_1))^+\right), \\ W_p(u, v) &= \max\left(0, u + v - 1, C(F_1(\tau_1), v) - (F_1(\tau_1) - u)^+\right). \end{aligned}$$

The proof of Proposition A.3 is in the appendix, page 18.

For a numerical illustration, consider the case when Z is normal with standardized marginals and correlation $\rho = 0.15$, and let $\tau_1 = 0.25$. If Z_2 is also dichotomized with $\tau_2 = -0.52$, this gives the 2×2 table used in the numerical illustration after Proposition 2. Proposition 3 gives $\rho(\mathcal{P}; p) = [-0.49, 0.68]$. This is considerably tighter than the bounds from Propositions 1 and 2. As in Section 2.2, a partial identification analysis of Spearman's rho is given by the above analysis when assuming uniform marginals.

3. The multivariate binary case

3.1. **The distinction between distributional and substantial knowledge.** Consider the case when we observe (X, Y) where Y is a random variable, and $X = (1\{Z_1 > \tau_1\}, 1\{Z_2 > \tau_2\})$. In the upcoming Theorem 2 we show in a more general setting that we cannot learn anything more about the distribution of Z from the joint distribution of (X, Y) compared to what we know from the distribution of X.

This may seem counter-intuitive, as Y is arbitrary, and may equal Z. If we knew that $Y = Z$, instead of just the distribution of (X, Y), the distribution of Z would have been identified. We define substantial knowledge as knowledge that is not derivable from the distribution of the observables. For example, that $Y = Z$ is substantial knowledge, as it cannot be deduced from the distribution of (X, Y), as shown in a more general setting in Theorem 2. Theorem 2 also shows that without substantial knowledge, knowledge of the joint distribution of (X, Y) is as informative for identifying the latent correlation as when only knowing the distribution of X.

Hence, without substantial knowledge, the bivariate case is studied without loss of generality when considering a bivariate statistic, such as the correlation.

Underlying normality of Z is substantial knowledge, see Section 3.3. Another example is when $Y = Z_2$ and this relation is known, which leads to the case considered in Section 2.5. An interesting third example is when Z is known to be discretized into a vector of ordinal variables X that have multiple categories. When Z is normal, this leads to the polychoric estimator of Pearson and Pearson (1922). We may represent the coordinates of X by a sequence of binary variables. For example, we could encode $(1\{\tau_{1,1} < Z_1 < \tau_{1,2}\} + 2 \times 1\{Z_1 > \tau_{1,2}\}, 1\{Z_2 > \tau_{2,1}\})$ by $(1\{Z_1 > \tau_{1,1}\}, 1\{Z_2 > \tau_{2,1}\}, Y)$ where $Y = 1\{Z_1 > \tau_{1,2}\}$. Substantial knowledge of the connection between Y and Z is then given from the structure of the problem. The authors are preparing a follow-up paper on this topic. A final example, now from a different context, is the direction and presence of causal effects in structural models, as these cannot always be deduced from observational data (Pearl, 2009). For example, there are many structural equation models for continuous data which has the same covariance matrix as other structural equation models with different causal directions (Bollen, 2014, Chapter 3). Which model is correct therefore cannot be deduced by statistical means, but requires substantial knowledge.

3.2. **Increasing the dimensionality can not help identify parameters when substantial knowledge is lacking.** We here briefly consider a more general problem, which encompasses the problem of latent correlations as a special case as shown in Example 1. For a probability measure P on S, and a random variable X, let P_X denote the distribution of X, defined by $P_X(A) = P(X \in A)$ (Kallenberg, 2006, p.47). The map $P \mapsto P_X$ is not injective in general. That is, there will usually be probabilities $P \neq P'$ such that $P_X = P'_X$. Let $f_\theta, \theta \in \Theta$ be a family of measurable functions. Define two families of measures by

$$\gamma(P_X) = \left\{P_Z \mid P_{f_\theta(Z)} = P_X \text{ for some } \theta\right\},$$
$$\gamma(P_{X,Y}) = \left\{P_Z \mid P_{f_\theta(Z),Y} = P_{X,Y} \text{ for some } \theta\right\}.$$

Here $\gamma(P_X)$ is the family of all distributions P_Z that could have generated some P_X by means of $f_\theta, \theta \in \Theta$. On the other hand, $\gamma(P_{X,Y})$ is the family of all distributions P_Z that could have generated $P_{X,Y}$ by means of $f_\theta, \theta \in \Theta$.

Example 1. *When $f_\theta(z) = (1\{z_1 > \theta_1\}, 1\{z_2 > \theta_2\})$ we regain the case in Section 3.1.*

Suppose we know the distribution $P_{X,Y}$. Can this knowledge be more informative than knowing P_X for deducing aspects of the distribution of Z? The following result shows this not to be possible. It is shown under a mild measure-theoretic assumption stated in the appendix, page 21.

Theorem 2. *We have $\gamma(P_X) = \gamma(P_{X,Y})$.*

3.3. **On the interpretation of tests for underlying normality.** For 2×2 tables, we saw in Section 2.1 that there is a bijection between the table p on one hand, and the normal theory

tetrachoric correlation and τ_1, τ_2 on the other. Underlying normality therefore has no testable implications.

As observed by Vaswani (1950) and Muthén and Hofacker (1988), we may increase the dimensionality, and study trivariate binary variables to reach a testable implication of underlying normality. Similar tests for compatibility with normality have been proposed in the general polychoric case with arbitrary dimensions (Foldnes & Grønneberg, 2019b; Maydeu-Olivares, 2006). While such tests can identify incompatibilities with underlying normality, what are the implications if such incompatibilities are not detected?

If we do not have substantial knowledge about the normality of the latent variables, Theorem 2 shows that compatibility with underlying multivariate normality cannot reduce the bounds found from Proposition 1 even when the marginals are known to be normal: Firstly, the bounds on ρ are optimal when taking into account only the bivariate information in the 2×2 table. Secondly, Theorem 2 shows that we cannot improve the bounds when taking into account multivariate information. Therefore, if a test for underlying normality is not rejected, or even when the exact distribution of X is compatible with having been generated from a multivariate normal Z, this fact is not useful from a partial identification perspective.

3.4. **Limitations originating from focusing on the bivariate case.** We have focused on the partial identification of a single latent correlation. Theorem 2 implies that including multivariate information cannot be used to rule out values attainable by this single correlation identified by a bivariate analysis. While this is a multivariate result, it still deals with the identification of a single bivariate correlation. This bivariate identification does not extend to multivariate identification. That is, we can not use our results to exactly calculate the space $\mathcal{S}$ of latent correlation matrices attainable by a multivariate Z that is compatible with the distribution of X and specified substantial knowledge of the distribution of Z.

The reason for this is that there need not be multivariate probability distributions which simultaneously attain the bivariate copulas identified in e.g., Proposition 1. This is similar to the more familiar setup with confidence regions for population means μ based on a multivariate normal sample. The standard 95% confidence region for μ is an ellipsoid. From this ellipsoid, we may deduce 95% confidence intervals for each coordinate of μ by identifying the values attained by this coordinate in the confidence ellipsoid. But we cannot go from knowing 95% confidence intervals for each coordinate of μ to knowing a 95% confidence region for μ: All we know is the rectangle within which the ellipsoid is contained, and this is not enough information to reconstruct the ellipsoid. For the same reason, we cannot in general deduce $\mathcal{S}$ from knowing how its coordinates vary. While this gives us an upper bound for $\mathcal{S}$, this upper bound is likely to be crude.

4. Concluding remarks

We have shown that a great deal of substantial knowledge is required to usefully analyse binary data through the perspective of latent correlations. As mentioned in Section 2, a partial identification analysis can be seen as a sensitivity analysis. Our analysis shows that the

methodology of tetrachoric correlations is highly sensitive to the assumption of underlying normality.

Our conclusions complement the analyses of Foldnes and Grønneberg (2019a, 2019b, 2020) where it was shown that if one simulates non-normal continuous data and discretize it, normal theory tetrachoric or polychoric correlations estimated from the discretized data can completely miss the underlying correlation, see for example Figure 2 in the introductory example of Foldnes and Grønneberg (2019b). The present paper exactly identifies what can be said about the latent correlation if we only know the discretized data and some specified aspects of the distribution of Z. If no substantial knowledge of the distribution of Z is known, which may often be the case, especially in exploratory studies, we have shown that nothing can be said about the latent correlations.

Even when substantial knowledge allows us to postulate known marginal distributions, the interval of latent correlations that are consistent with the data is still very wide. Smaller and more informative intervals are only available by imposing restrictions on the dependency structure among the underlying latent variables, as we saw in our illustration in Section 2.4. This kind of substantial knowledge seems hard to justify in many practical applications. We therefore must conclude that the normal theory tetrachoric correlation coefficient may not be an informative measure of association for binary variables. We stress that this criticism holds only if underlying normality is not known. If underlying normality is known, there is no problem with the tetrachoric correlation as a measure of association.

An important extension of our investigation is the polychoric case. Most psychometric tests are based on 5-point scales, and the typical size of the set of possible values of latent correlation matrices in this case is practically important. When marginals are known and the number of categories increase, we approach the identified case, and the speed at which this occurs is an interesting subject of investigation. When the marginals are unknown, this convergence does not take place, as the scale of the correlation is undetermined.

Appendix A. Technical proofs and further numerical illustrations

A.1. **Further numerical illustrations for given marginals.** We here give further numerical illustrations of the bounds. Since there is a bijection between 2×2 tables and the dichotomization of standard normal distributions with free correlations and free thresholds τ_0, τ_1, we generate 2×2 tables from proportions of a normal latent variable with varying correlations and chosen threshold parameters. For each table, we compute the bounds from Proposition 1 with standard normal marginals, as well as the bound from Proposition 2. Recall that the bound from Proposition 2 is actually the bound from Proposition 1 with uniform marginals. The lengths of the resulting intervals are shown in Figures 4 and 5. Full computational details are given in the accompanying `R` scripts. In Figure 4, we have $\tau_1 = \tau_2 = 0$, which is a best case scenario. Figure 5 shows a more typical situation, where the length of all bounds are close to the maximal length of such an interval, namely 2. Figure 5 incidentally also illustrates that the bound for the Pearson correlation with standard normal marginals does not always contain the bound for Spearman's rho. In both Figure 4 and Figure 5, points close or at the endpoints

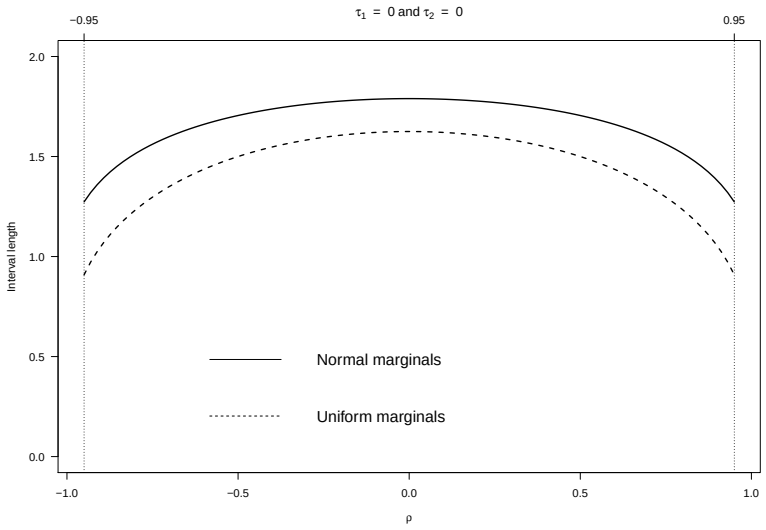

FIGURE 4. Length of bounds for $\tau_1 = 0, \tau_2 = 0$ based on normal or uniform marginal assumptions. The graph does not cover points close to $\rho = \pm 1$.

$\rho = \pm 1$ are not included, as different numerical techniques are needed in this region, as done in an attached `R` file found in the online supplementary material. It is here found that minimum lengths are attained at $\rho = \pm 1$ and $\tau_1 = \tau_2 = 0$, with a length of 0.67 for normal marginals and 0.5 for uniform marginals. In our analysis, we use the `R` (R Core Team, 2020) packages `copula` (Yan & Others, 2007), `cubature` (Narasimhan, Johnson, Hahn, Bouvier, & Kiêu, 2020), and `copBasic` (Asquith, 2020).

A.2. **Proofs for Section 2.** We will sometimes use the following principle of duality, as observed by Tankov (2011, Appendix). The usual matrix of probabilities is

$$P = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}.$$

The swapped matrix is

$$P^\star = \begin{bmatrix} p_{01} & p_{00} \\ p_{11} & p_{10} \end{bmatrix}.$$

This matrix has will have the same upper bound as the negative lower bound of P; this is because it corresponds to the discretized distribution of $(-X, Y)$. Hence we may compute, say, a lower bound via an upper bound by using this duality. Some of the upcoming arguments apply this technique when convenient.

Proof of Theorem 1. We show that $|\rho| \neq 1$ by contradiction. Suppose $|\rho| = 1$. By the Cauchy-Schwarz inequality, $Z_1 = a + bZ_2$ for some numbers a, b. For any thresholds τ_1, τ_2, the probabilities of X equals the probability of observing Z in one of the quadrants $x > \tau_1, y > \tau_2$ or $x < \tau_1, y < \tau_2$ or $x > \tau_1, y < \tau_2$ or $x < \tau_1, y > \tau_2$. Since any two straight lines intersect at

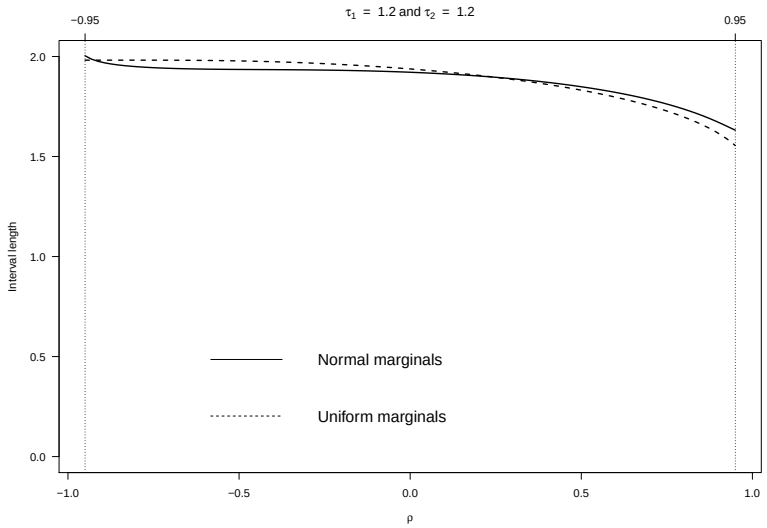

FIGURE 5. Length of bounds for $\tau_1 = 1.2, \tau_2 = 1.2$ based on normal or uniform marginal assumptions. The graph does not cover points close to $\rho = \pm 1$.

either one or zero points, one quadrant will have zero probability, therefore contradicting our assumption that none of the cell probabilities are zero. Therefore, $|\rho| = 1$ is incompatible with the distribution of X.

Now we show that any $\rho \in (0, 1)$ is compatible with X. To do this, let $a, b > 0$ be two positive real numbers and define the random variable

$$Z(a,b) \mid X = \begin{cases} (a,a) & X = (1,1), \\ (b,-b) & X = (1,0), \\ (-a,-a) & X = (0,0), \\ (-b,b) & X = (0,1). \end{cases}$$

Then $\operatorname{pr}[Z(a,b) \in A_{ij}] = \operatorname{pr}[X = (i,j)] = p_{ij}$ when A_{ij} are the quadrants $A_{00} = [-\infty, 0] \times [-\infty, 0]$, $A_{01} = [0, \infty] \times [-\infty, 0]$, $A_{10} = [-\infty, 0] \times [0, \infty]$, and $A_{11} = [0, \infty] \times [0, \infty]$. Thus $Z(a,b)$ induces X through discretization when $\tau_1 = \tau_2 = 0$. We now let $a = 1/b$. When $b \to 0^+$, we get a correlation converging to 1. When $b \to \infty$, we get a correlation converging to -1. This is visually obvious, as the points get closer and closer to a straight line, and is confirmed algebraically in the online appendix accompanying this paper. At the end of the online appendix, we also show that any intermediate value is possible, which is a consequence of the continuity of the correlation of Z as a function of b. □

Proof of Proposition 1. Theorem 3.2.3 of Nelsen (2007, p. 70) shows that all copulas C that fulfil eq. (5) fulfil $W_p(u,v) \leq C(u,v) \leq M_p(u,v)$ and that W_p, M_p are copulas fulfilling the constraint in eq. (5). The Höffding representation in eq. (4) therefore implies $\rho(W_p[F_1, F_2]) \leq \rho(F) \leq \rho(M_p[F_1, F_2])$. Since W_p, M_p are copulas, this bound cannot be improved. We now

show that the interval with limits as in the bound for $\rho(F)$ equals $\rho(\mathcal{P}, p)$. We use an argument that goes back to Fréchet (1958), see (Nelsen, 2007, p. 15, exercise 2.4).

Let $\rho_L = \rho(W_p[F_1, F_2])$ and $\rho_U = \rho(M_p[F_1, F_2])$. Suppose $\rho \in [\rho_L, \rho_U]$. Then there is an $0 \leq \alpha \leq 1$ such that

$$\alpha\rho_L + (1-\alpha)\rho_U = \rho. \tag{8}$$

Let $C_\alpha(u, v) = \alpha W_p(u, v) + (1-\alpha)M_P(u, v)$ which is a convex combination of copulas, and hence a copula (Nelsen, 2007, Exercise 2.3 and 2.4). Let $H_\alpha(x_1, x_2) = C_\alpha(F_1(x_1), F_2(x_2))$. By the second half of Sklar's theorem, H_α is a distribution function with marginals F_1, F_2. Since $F_1(\tau_1) = p_{01} + p_{00}$ and $F_2(\tau_2) = p_{10} + p_{00}$, and $p_{00} = H_\alpha(\tau_1, \tau_2) = C_\alpha(F_1(\tau_1), F_2(\tau_2)) = C_\alpha(p_{01} + p_{00})$ the copula C_α fulfils eq. (5). Therefore, $H_\alpha \in \mathcal{P}$. We now show that $\rho(H_\alpha) = \rho$ using the Höffding representation from eq. (4) in Section 2.

Firstly, we have $F_1(x_1)F_2(x_2) = \alpha F_1(x_1)F_2(x_2) + (1-\alpha)F_1(x_1)F_2(x_2)$, and so by the Höffding representation equation (4), the covariance of H_α equals

$$\begin{aligned}
\rho(H_\alpha) =& \operatorname{sd}(F_1)^{-1}\operatorname{sd}(F_2)^{-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} C_\alpha(F_1(x_1), F_2(x_2)) - F_1(x_1)F_2(x_2)\,\mathrm{d}x_1\mathrm{d}x_2 \\
=& \operatorname{sd}(F_1)^{-1}\operatorname{sd}(F_2)^{-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \alpha C_L(F_1(x_1), F_2(x_2)) - \alpha F_1(x_1)F_2(x_2) \\
&+ \operatorname{sd}(F_1)^{-1}\operatorname{sd}(F_2)^{-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (1-\alpha)C_U(F_1(x_1), F_2(x_2)) \\
&- (1-\alpha)F_1(x_1)F_2(x_2)\,\mathrm{d}x_1\mathrm{d}x_2 \\
=& \alpha\rho(W_p[F_1, F_2]) + (1-\alpha)\rho(M_p[F_1, F_2]) \\
=& \rho
\end{aligned}$$

using equation (8). □

Proof of Proposition 2. Define $a = p_{00}$, $b = p_{00} + p_{01}$, $c = p_{00} + p_{10}$ and $d = c + b - a$. We will calculate the integral $\int_{[0,1]^2} C_U(u, v)\,\mathrm{d}u\mathrm{d}v$. Define the set $A_F = [a, d] \times [a, d]$. Then

$$\int_{[0,1]^2} C_U(u, v)\,\mathrm{d}u\mathrm{d}v = \int_{A_F} C_U(u, v)\,\mathrm{d}u\mathrm{d}v + \int_{A_F^C} C_U(u, v)\,\mathrm{d}u\mathrm{d}v. \tag{9}$$

On A_F^C it holds that $C_U(u, v) = \min(u, v)$. Since $\int_{[0,1]^2} \min(u, v)\,\mathrm{d}u\mathrm{d}v = 1/3$ and

$$\int_a^d\int_a^d \min(u, v)\,\mathrm{d}u\mathrm{d}v = \frac{1}{3}(a + b + c)(a + b - 2c)$$

the second integral in (9) equals

$$\int_{A_F^c} C_U(u, v)\,\mathrm{d}u\mathrm{d}v = \frac{1}{3} - \frac{1}{3}(b - a)(c - a)(b + c).$$

The next part is $\int_{A_F} C_U(u,v)\,\mathrm{d}u\mathrm{d}v$. It is handy to divide A_F into four rectangles

$$\begin{aligned} A_{BL} &= [a,b]\times[a,c]\,, \\ A_{TR} &= [b,d]\times[c,d]\,, \\ A_{TL} &= [a,b]\times[c,d]\,, \\ A_{BR} &= [b,d]\times[a,c]\,. \end{aligned}$$

At A_{BL} we have $C_U(u,v)=a$ and

$$\int_{A_{BL}} C_U(u,v)\,\mathrm{d}u\mathrm{d}v = a(b-a)(c-a)\,.$$

At A_{TR}, $C_U(u,v)=-d+u+v$ and its integral is

$$\int_{A_{TR}} C_U(u,v)\,\mathrm{d}u\mathrm{d}v = \frac{1}{2}(b-a)(c-a)(b+c)\,.$$

At A_{TL}, $C_U(u,v)=\min(u,a-c+v)$ and the integral equals

$$\int_{A_{TL}} C_U(u,v)\,\mathrm{d}u\mathrm{d}v = \frac{1}{3}(b-a)^2(2a+b)\,,$$

and at A_{BR}, $C_U(u,v)=\min(v,a-b+u)$ the integral is

$$\int_{A_{BR}} C_U(u,v)\,\mathrm{d}u\mathrm{d}v = \frac{1}{3}(c-a)^2(2a+c)\,.$$

Add all the expressions together, make the substitutions $b=p_{01}+p_{00}$, $a=p_{10}+p_{00}$ and simplify to get

$$\int_{[0,1]^2} C_U(u,v)\,\mathrm{d}u\mathrm{d}v = \frac{1}{6}\left(2-3p_{01}p_{10}(p_{01}+p_{10})\right)$$

hence

$$12\int_{[0,1]^2} C_U(u,v)\,\mathrm{d}u\mathrm{d}v - 3 = 1-6p_{01}p_{10}(p_{01}+p_{10})$$

as claimed. The lower bound follows by duality. □

The reasoning behind the decomposition can be seen in Figure 6, where each colour correspond to a continuous part of the piece-wise continuous function $C_U(u,v)$.

A.3. **Proofs for Section 2.5.**

Proof of Proposition 3. We follow the structure of the argument of Proposition 1. To help simplify the argument, we structure the argument in a series of lemmas. For easy reference, these lemmas are stated inside the present proof. The proofs of these supporting lemmas follow after the present proof is complete.

Firstly, let us identify what can be said of C when knowing the distribution of X, which is given by the function $p(x_1,y)=\mathrm{P}(X_1=x_1,Z_2\leq y)$, for $x_1=0,1$ and y a real number. We have that $p(0,y)=\mathrm{P}(X_1=0,Z_2\leq y)=\mathrm{P}(Z_1\leq\tau_1,Z_2\leq y)=C(F_1(\tau_1),F_2(y))$. Since $p(0,y)+p(1,y)=F_2(y)$, and therefore $p(1,y)=F_2(y)-p(0,y)$, we do not get new knowledge from similarly expressing $p(1,y)$ in terms of the copula C. Our knowledge of C is therefore that

$$C(u,v)=p(0,F_2^{-1}(v)) \quad ((u,v)\in\mathcal{U}=\{(u,v)\mid u=F_1(\tau_1),0\leq v\leq 1\})\,. \tag{10}$$

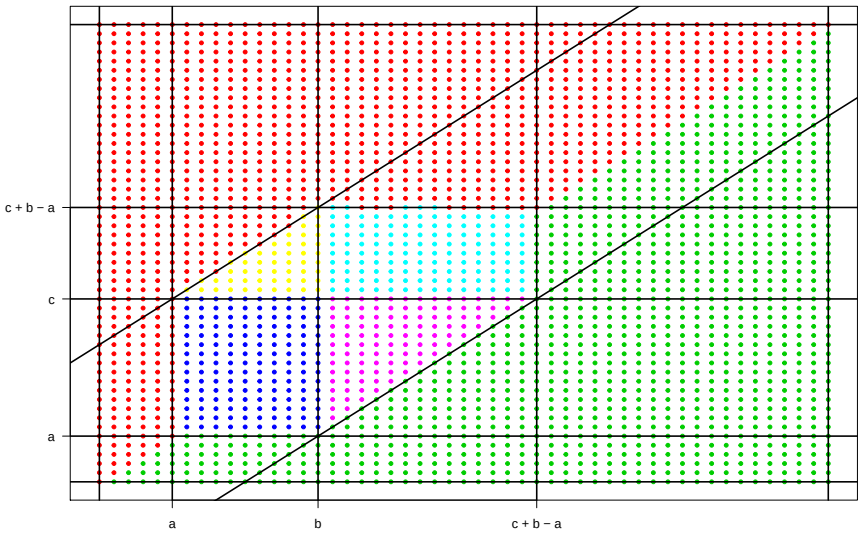

FIGURE 6. Colour-coded graph of the bound copula. Each colour correspond to a continuous part of the piece-wise continuous function $C_U(u, v)$.

We now use a constrained Fréchet–Höffding bound found in Tankov (2011) to take into account this knowledge.

Lemma 1. *Any copula C that satisfies equation (10) also satisfies*

$$C_{L,\mathcal{U}}(u, v) \leq C(u, v) \leq C_{L,\mathcal{U}},$$

where $C_{L,\mathcal{L}}$ and $C_{U,\mathcal{U}}$ are

$$C_{U,\mathcal{U}}(u, v) = \min(u, v, \min_b[C(F_1(\tau_1), b) + (u - F_1(\tau_1))^+ + (v - b)^+]), \tag{11}$$

$$C_{L,\mathcal{U}}(u, v) = \max(0, u + v - 1, \max_b[C(F_1(\tau_1), b) - (F_1(\tau_1) - u)^+ - (b - v)^+]). \tag{12}$$

Moreover, both $C_{L,\mathcal{U}}$ and $C_{U,\mathcal{U}}$ are copulas that satisfy equation (10).

Let us now simplify the expressions for C_L, C_U through identifying the inner minimum or maximum in C_L, C_U respectively. This will show that they are equal to the expressions in the statement of the result. This is achieved in the following lemma.

Lemma 2. *The copulas $C_{L,\mathcal{U}}$ and $C_{U,\mathcal{U}}$ are equal respectively to W_p, M_p from the statement of Proposition 3. That is,*

$$\begin{aligned} C_{U,\mathcal{U}}(u, v) &= \min(u, v, \min_{b\in[0,1]}[C(F_1(\tau_1), b) + (u - F_1(\tau_1))^+ + (v - b)^+]) \\ &= \min(u, v, C(F_1(\tau_1), v) + (u - F_1(\tau_1))^+), \end{aligned} \tag{13}$$

and

$$C_{U,\mathcal{U}}(u,v) = \max(u, u+v-1, \max_{b\in[0,1]}[C(F_1(\tau_1), b) - (F_1(\tau_1)-u)^+ - (b-v)^+]), \tag{14}$$
$$= \max(0, u+v-1, C(F_1(\tau_1), v) - (F_1(\tau_1)-u)^+).$$

From this, the Höffding representation from eq. (3) in Section 2 gives for any $F \in \mathcal{P}$ which is compatible with p that $\rho(W[F_1,F_2;p]) \leq \rho(F) \leq \rho(M[F_1,F_2;p])]$. We now show that any values within this interval can be attained as correlations in $\rho(\mathcal{P},p)$.

As in the proof of Proposition 1, we study convex combinations of W_p and M_p. For $0 \leq \alpha \leq 1$, we study $C_\rho(u,v) = \alpha W_p + (1-\alpha)M_p$. That this class induces all correlation values in the stated interval follows exactly as in the proof of Proposition 1. What is left to show is that the convex combination also fulfil the restriction in eq. (10). Now from Lemma 1, we have that both W_p and M_p fulfil eq. (10), i.e., that $W_p(F_1(\tau_1),v) = M_p(F_1(\tau_1),v) = p(0,F_2^{-1}(v))$. Therefore, we also have $C_\rho(F_1(\tau_1),v) = \alpha C_{L,\mathcal{U}}(F_1(\tau_1),v) + (1-\alpha)C_{U,\mathcal{U}}(F_1(\tau_1),v) = \alpha p(0,F_2^{-1}(v)) + (1-\alpha)p(0,F_2^{-1}(v)) = p(0,F_2^{-1}(v))$. □

We now prove the two lemmas stated within the proof of Proposition 3.

Proof of Lemma 1. Since $\mathcal{U}$ is compact, Theorem 1 (i) of Tankov (2011) shows the claimed bound, and that $C_{L,\mathcal{U}}$ and $C_{U,\mathcal{U}}$ fulfil equation (10).

We now check the conditions of Theorem 1 (ii) of Tankov (2011) which shows that $C_{L,\mathcal{U}}$ and $C_{U,\mathcal{U}}$ are actually copulas. What is required is that $\mathcal{U}$ is both a increasing and a so-called decreasing set, as defined in Tankov (2011, Section 2, bottom of p. 390): A set $S \subset [0,1]^2$ is increasing if for all $(a_1,b_1),(a_2,b_2) \in S$ we have either (i) $a_1 \leq a_2$ and $b_1 \leq b_2$ or (ii) $a_1 \geq a_2$ and $b_1 \geq b_2$. For $S = \mathcal{U}$ this is trivially fulfilled, since if $(a_1,b_1),(a_2,b_2) \in \mathcal{U}$ we have $a_1 = a_2 = F_1(\tau_1)$ as we only have one possible element in the first coordinate, and therefore we trivially also have that either $b_1 \leq b_2$ or $b_1 \geq b_2$ by tautology.

Similarly, recall that a set $S \subseteq [0,1]^2$ is decreasing if for all $(a_1,b_1),(a_2,b_2) \in S$ we have either (i) $a_1 \leq a_2$ and $b_1 \geq b_2$ or (ii) $a_1 \geq a_2$ and $b_1 \leq b_2$. This is again trivially fulfilled. □

For the proof of Lemma 2, we need the following technical result.

Lemma 3. *Let C be a bivariate copula distribution function and $0 \leq a \leq 1$. Then $C(a,v) - v$ is decreasing in v when $0 \leq v \leq 1$.*

Proof. By definition (Nelsen, 2007, p. 8), a bivariate copula satifies $C(1,v) = v$ when $0 \leq v \leq 1$ and

$$C(u_1,v_1) - C(u_2,v_1) \geq C(u_1,v_2) - C(u_2,v_2)$$

when $0 \leq u_1 \leq u_2 \leq 1$ and $0 \leq v_1 \leq v_2 \leq 1$. Now choose $u_1 = a$ and $u_2 = 1$, and $C(a,v_1) - v_1 \geq C(a,v_2) - v_2$ when $0 \leq v_1 \leq v_2 \leq 1$, as claimed. □

Proof of Lemma 2. We start with $C_{U,\mathcal{U}}$. We must show that

$$\begin{aligned} C_{U,\mathcal{U}}(u,v) &= \min(u, v, \min_{b\in[0,1]}[C(F_1(\tau_1), b) + (u - F_1(\tau_1))^+ + (v-b)^+]), \\ &= \min(u, v, C(F_1(\tau_1), v) + (u - F_1(\tau_1))^+), \end{aligned}$$

where the first equality is from Lemma 1 while the second line is the definition of $M_p(u,v)$ from Proposition 3. The second equality holds if, and only if,

$$\min_{b\in[0,1]}[C(F_1(\tau_1), b) + (v-b)^+] = C(F_1(\tau_1), v),$$

which is true if and only if $C(F_1(\tau_1), b) + (u - F_1(\tau_1))^+ + (v-b)^+$ is minimized when $b = v$. Now we show this is indeed the case. For $b \leq v$, we have $0 \leq v - b$, and so $h(b) = C(F_1(\tau_1), b) + v - b$, which is decreasing by Lemma 3 (p. 20). For $b > v$, we have $v - b < 0$, and so $h(b) = C(F_1(\tau_1), b)$, which is increasing. The minimum is therefore attained at $b = v$ and

$$\min_{b\in[0,1]}[C(F_1(\tau_1), b) + (v-b)^+] = C(F_1(\tau_1), v),$$

as claimed.

The case of $C_{L,\mathcal{U}}$ is similar, as we have to show that

$$\begin{aligned} C_{U,\mathcal{U}}(u,v) &= \max(u, u+v-1, \max_{b\in[0,1]}[C(F_1(\tau_1), b) - (F_1(\tau_1) - u)^+ - (b-v)^+]), \\ &= \max(0, u+v-1, C(F_1(\tau_1), v) - (F_1(\tau_1) - u)^+). \end{aligned}$$

Again, the first line is from Lemma 1 and second line is the definition of W_p from Proposition 3. The second equality holds if, and only if,

$$\max_{b\in[0,1]}[C(F_1(\tau_1), b) - (b-v)^+] = C(F_1(\tau_1), v).$$

This equality is true by the same reasoning as above. For $b \leq v$, we have $b - v \leq 0$ and so $g(b) = C(F_1(\tau_1), b)$, which is increasing. For $b > v$, we have $b - v > 0$ and so $g(b) = C(F_1(\tau_1), b) - b + v$, which is decreasing by Lemma 3 (p. 20). Therefore, the maximum is attained at $b = v$, and

$$\max_{b\in[0,1]}[C(F_1(\tau_1), b) - (b-v)^+] = C(F_1(\tau_1), v).$$

as claimed. □

A.4. **Proof for Section 3.2.** Let $S = (\Omega, \Sigma)$ be a measure space. We assume S is an uncountable standard Borel space, i.e., it can be identified with the Borel space over the real numbers. We also assume that S is a *rich Borel space*, meaning it supports an independent uniform random variable that can be used as a randomization device (Kallenberg, 2006, p.112). This assumption can be made with practically no loss of generality.

Proof of Theorem 2. The inclusion $\gamma(P_{X,Y}) \subseteq \gamma(P_X)$ is true for any Z and S. Choose a P_X, a $P_{X,Y}$ compatible with P_X, and a $P_Z \in \gamma(P_X)$. We must show $P_Z \in \gamma(P_{X,Y})$, or $P_{f_\theta(Z),Y} = P_{X,Y}$ for some $\theta \in \Theta$. As a candidate θ choose one of the witnesses of $P_{f_\theta(Z)} = P_X$. By assumption there are two variables X, Y in S with distribution $P_{X,Y}$ such that X is distributed as $f_\theta(Z)$ when Z is distributed according to P_Z. By Corollary 6.11 of Kallenberg (2006), there

is a variable Z' in S such that $X = f_\theta(Z')$ and $P_Z = P'_Z$. But then $P_{f_\theta(Z'),Y} = P_{X,Y}$ and we are done. □

A.5. **Computational simplifications when applying Proposition 1.** The integrals defining the end points of $\rho(\mathcal{P}, p)$ in Proposition 1 can be calculated directly via numerical integration. However, this approach is computationally intensive, as we integrate functions with jumps. We here simplify the integrals in Proposition 1 by splitting the integrals into regions without jumps. This considerably reduces the computational burden of numerical integration. The analysis is analogous to the proof in Proposition 2, except that the integrals at A_{BR} and A_{TR} must be divided in two.

We only treat the upper bound. The lower bound can be found by duality. In the following argument, we assume that F_1, F_2 have variance one, an assumption made without loss of generality, as it can be achieved by re-scaling.

Define

$$g(u,v) = M[F_1, F_2; p](F_1(u), F_2(v)) - \min(F_1(u), F_2(v)).$$

By the Höffding formula for covariance, we have

$$\begin{aligned}\rho(M[F_1, F_2; p]) &= \int_{\mathbb{R}^2} M[F_1, F_2; p](F_1(u), F_2(v)) - F_1(u) F_2(v) \,\mathrm{d}u\mathrm{d}v \\ &= J_1 + \int_{\mathbb{R}^2} g(u,v) \,\mathrm{d}u\mathrm{d}v\end{aligned}$$

where

$$J_1 = \int_{\mathbb{R}^2} \min(F_1(u), F_2(v)) - F_1(u) F_2(v) \,\mathrm{d}u\mathrm{d}v.$$

Here, J_1 is the covariance of the distribution with the Fréchet–Höffding upper bound copula and marginals F_1, F_2. The integral J_1 is seen to be finite by the Cauchy-Schwarz inequality, since it is a covariance where the marginals are assumed to have finite variance. The integral $\int_{\mathbb{R}^2} g(u,v) \,\mathrm{d}u\mathrm{d}v$ can be calculated using a similar decomposition as the one used in Proposition 2. We see that $\rho(M[F_1, F_2; p]) = \sum_{i=1}^8 J_i$ where

$$\begin{aligned}J_2 &= -\int_B \min(F_1(u), F_2(v)) \,\mathrm{d}u\mathrm{d}v, \\ J_3 &= \int_{A_{BL}} M[F_1, F_2; p](F_1(u), F_2(v)) \,\mathrm{d}u\mathrm{d}v, \\ J_4 &= \int_{A_{TR}} M[F_1, F_2; p](F_1(u), F_2(v)) \,\mathrm{d}u\mathrm{d}v, \\ J_5 &= \int_{T_{TL1}} M[F_1, F_2; p](F_1(u), F_2(v)) \,\mathrm{d}u\mathrm{d}v, \\ J_6 &= \int_{T_{TL1}} M[F_1, F_2; p](F_1(u), F_2(v)) \,\mathrm{d}u\mathrm{d}v, \\ J_7 &= \int_{T_{BR1}} M[F_1, F_2; p](F_1(u), F_2(v)) \,\mathrm{d}u\mathrm{d}v, \\ J_8 &= \int_{T_{BR2}} M[F_1, F_2; p](F_1(u), F_2(v)) \,\mathrm{d}u\mathrm{d}v.\end{aligned}$$

The domains of integration can be seen in Figure 6. Here R_{BL} is the bottom-left rectangle, T_{TL1} the first top-left triangle, et cetera.

When the marginals are normal, concrete formulas for the integrals over J_3 and J_4 are possible to derive by using well-known results for normal integrals (Owen, 1980). A simple algebraic formula such as that given in Proposition 2 seems out of reach in this case, as the integrals J_5, J_6, J_7, J_8 are too complicated.

In our numerical implementation, we assume that F_1, F_2 are equal, and are capable of supporting perfect correlations of ± 1, as is well known to hold for normal marginals. As shown in Section 2, the maximum possible correlation with marginals F_1, F_2 equals J_1, and so this assumption amounts to $J_1 = 1$.

References

Almeida, C., & Mouchart, M. (2014). Testing normality of latent variables in the polychoric correlation. *Statistica*, *74*(1), 3–25. https://doi.org/10.6092/issn.1973-2201/4594

Asparouhov, T., & Muthén, B. (2016). Structural equation models and mixture models with continuous nonnormal skewed distributions. *Structural Equation Modeling*, *23*(1), 1–19. https://doi.org/10.1080/10705511.2014.947375

Asquith, W. H. (2020). copBasic—General bivariate copula theory and many utility functions [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=copBasic

Azzalini, A. (2013). *The skew-normal and related families*. Cambridge University Press. https://doi.org/10.1017/CBO9781139248891

Bernard, C., Jiang, X., & Vanduffel, S. (2012). A note on 'Improved Fréchet bounds and model-free pricing of multi-asset options' by Tankov (2011). *Journal of Applied Probability*, *49*(3), 866–875. https://doi.org/10.2139/ssrn.2003462

Bollen, K. A. (2014). *Structural equations with latent variables*. John Wiley & Sons. https://doi.org/10.1002/9781118619179

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*(1), 5–32. https://doi.org/10.1007/BF02291477

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790485

Foldnes, N., & Grønneberg, S. (2019a). On identification and non-normal simulation in ordinal covariance and item response models. *Psychometrika*, *84*(4), 1000–1017. https://doi.org/10.1007/s11336-019-09688-z

Foldnes, N., & Grønneberg, S. (2019b). Pernicious polychorics: The impact and detection of underlying non-normality. *Structural Equation Modeling*, *27*(4), 525–543. https://doi.org/10.1080/10705511.2019.1673168

Foldnes, N., & Grønneberg, S. (2020). The sensitivity of structural equation modeling with ordinal data to underlying non-normality and observed distributional forms. *Psychological Methods*. (Forthcoming)

Fréchet, M. (1958). Remarques de M. Fréchet au sujet de la note précédente. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*(2), 2719–2720. Retrieved from https://gallica.bnf.fr/ark:/12148/bpt6k723q/f661.image

Fréchet, M. (1960). Sur les tableaux de corrélation dont les marges sont données. *Revue de l'Institut International de Statistique*, *28*(1/2), 10–32. https://doi.org/10.2307/1401846

Grønneberg, S., & Foldnes, N. (2017). Covariance model simulation using regular vines. *Psychometrika*, *82*(4), 1035–1051. https://doi.org/https://doi.org/10.1007/s11336-017-9569-60

Höffding, W. (1940). *Maßstabinvariante korrelationstheorie für diskontinuierliche verteilungen* (Unpublished doctoral dissertation). Universität Berlin.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press. https://doi.org/10.1201/b13150

Jöreskog, K. G. (1994). Structural equation modeling with ordinal variables. In *Multivariate analysis and its applications* (pp. 297–310). Institute of Mathematical Statistics. https://doi.org/10.1214/lnms/1215463803

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Scientific Software International.

Kallenberg, O. (2006). *Foundations of modern probability* (2nd ed.). Springer Science & Business Media. https://doi.org/10.1007/978-1-4757-4015-8

Kolenikov, S., & Angeles, G. (2009). Socioeconomic status measurement with discrete proxy variables: Is principal component analysis a reliable answer? *Review of Income and Wealth*, *55*(1), 128–165. https://doi.org/10.1111/j.1475-4991.2008.00309.x

Lehmann, E. L. (1966). Some concepts of dependence. *The Annals of Mathematical Statistics*, *37*(5), 1137–1153. https://doi.org/10.1214/aoms/1177699260

Manski, C. F. (2003). *Partial identification of probability distributions*. Springer Science & Business Media. https://doi.org/10.1007/b97478

Maydeu-Olivares, A. (2006). Limited information estimation and testing of discretized multivariate normal structural models. *Psychometrika*, *71*(1), 57–77. https://doi.org/10.1007/s11336-005-0773-4

Molenaar, D., & Dolan, C. V. (2018). Nonnormality in latent trait modelling. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing* (pp. 347–373). Wiley Online Library. https://doi.org/10.1002/9781118489772.ch13

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*(4), 551–560. https://doi.org/10.1007/BF02293813

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. https://doi.org/10.1007/BF02294210

Muthén, B., & Hofacker, C. (1988). Testing the assumptions underlying tetrachoric correlations. *Psychometrika*, *53*(4), 563–577. https://doi.org/10.1007/BF02294408

Muthén, L., & Muthén, B. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Narasimhan, B., Johnson, S. G., Hahn, T., Bouvier, A., & Kiêu, K. (2020). cubature: Adaptive multivariate integration over hypercubes [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=cubature

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4757-3076-0

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443–460. https://doi.org/10.1007/BF02296207

Owen, D. B. (1980). A table of normal integrals. *Communications in Statistics - Simulation and Computation*, *9*(4), 389–419. https://doi.org/10.1080/03610918008812164

Pearl, J. (2009). *Causality*. Cambridge University Press. https://doi.org/10.1017/cbo9780511803161

Pearson, K. (1900). I. Mathematical contributions to the theory of evolution.—VII. on the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A*, *195*, 1–47. https://doi.org/10.1098/rsta.1900.0022

Pearson, K. (1909). On a new method of determining correlation between a measured character a, and a character b, of which only the percentage of cases wherein B exceeds (or falls short of) a given intensity is recorded for each grade of a. *Biometrika*, *7*(1/2), 96–105. https://doi.org/10.2307/2345365

Pearson, K., & Heron, D. (1913). On theories of association. *Biometrika*, *9*(1/2), 159–315. https://doi.org/10.2307/2331805

Pearson, K., & Pearson, E. S. (1922). On polychoric coefficients of correlation. *Biometrika*, *14*(1/2), 127–156. https://doi.org/10.2307/2331858

R Core Team. (2020). R: A Language and Environment for Statistical Computing [Computer software manual].

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Satorra, A., & Bentler, P. (1988). *Scaling corrections for statistics in covariance structure analysis* (Tech. Rep.). Retrieved from https://escholarship.org/content/qt3141h70c/qt3141h70c.pdf

Shapiro, A. (1983). Asymptotic distribution theory in the analysis of covariance structures. *South African Statistical Journal*, *17*(1), 33–81. Retrieved from https://journals.co.za/content/sasj/17/1/AJA0038271X_800

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

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408. https://doi.org/10.1007/BF02294363

Tamer, E. (2010). Partial identification in econometrics. *Annual Review of Economics*, *2*(1), 167–195. https://doi.org/10.1146/annurev.economics.050708.143401

Tankov, P. (2011). Improved Fréchet bounds and model-free pricing of multi-asset options. *Journal of Applied Probability*, *48*(2), 389–403. https://doi.org/10.1239/jap/1308662634

Tate, R. F. (1955a). Applications of correlation models for biserial data. *Journal of the American Statistical Association*, *50*(272), 1078-1095. https://doi.org/10.1080/01621459.1955.10501293

Tate, R. F. (1955b). The theory of correlation between two continuous variables when one is dichotomized. *Biometrika*, *42*(1/2), 205–216. https://doi.org/10.21236/ad0029741

Vaswani, S. (1950). Assumptions underlying the use of the tetrachoric correlation coefficient. *Sankhyā: The Indian Journal of Statistics*, *10*(3), 269–276. Retrieved from https://www.jstor.org/stable/25048031

Whitt, W. (1976). Bivariate distributions with given marginals. *The Annals of Statistics*, *4*(6), 1280–1289. https://doi.org/10.1214/aos/1176343660

Yan, J., & Others. (2007). Enjoy the joy of copulas: With a package copula. *Journal of Statistical Software*, *21*(4), 1–21. https://doi.org/10.18637/jss.v021.i04

Department of Economics, BI Norwegian Business School, Oslo, Norway 0484
Email address: steffeng@gmail.com

Department of Mathematics, University of Oslo, PB 1053, Blindern, NO-0316, Oslo, Norway
Email address: jonasmgj@math.uio.no

Department of Economics, BI Norwegian Business School, Stavanger, Norway 4014
Email address: njal.foldnes@gmail.com

ONLINE APPENDIX FOR "PARTIAL IDENTIFICATION OF LATENT CORRELATIONS WITH BINARY DATA"

STEFFEN GRØNNEBERG, JONAS MOSS, AND NJÅL FOLDNES

Besides this online appendix, the online supplementary material accompanying the paper "Partial identification of latent correlations with binary data" includes several `R`-scripts. These are described in the text-file `index.txt`.

1. Detailed algebraic verification of Theorem 1

For completeness, we here provide a complete algebraic verification of Theorem 1. The calculations are tedious but elementary.

The distribution of $Z = (Z_1, Z_2)$ is

$$\begin{aligned} \mathrm{P}(Z = (a,a)) &= p_{11}, & \mathrm{P}(Z = (b,-b)) &= p_{10}, \\ \mathrm{P}(Z = (-a,-a)) &= p_{00}, & \mathrm{P}(Z = (-b,b)) &= p_{01}. \end{aligned}$$

From this we compute

$$\mathrm{E}(Z_1 Z_2) = a^2(p_{11} + p_{00}) - b^2(p_{10} + p_{01}).$$

The marginal distributions of Z_1, Z_2 are

$$\begin{aligned} \mathrm{P}(Z_1 = a) &= p_{11} = \mathrm{P}(Z_2 = a), & \mathrm{P}(Z_1 = b) &= p_{10} = \mathrm{P}(Z_2 = -b) \\ \mathrm{P}(Z_1 = -a) &= p_{00} = \mathrm{P}(Z_2 = -a), & \mathrm{P}(Z_1 = -b) &= p_{01} = \mathrm{P}(Z_2 = b). \end{aligned}$$

We therefore have

$$\begin{aligned} \mathrm{E}(Z_1) &= ap_{11} + bp_{10} - ap_{00} - bp_{01} \\ &= a(p_{11} - p_{00}) + b(p_{10} - p_{01}), \\ \mathrm{E}(Z_2) &= ap_{11} - bp_{10} - ap_{00} + bp_{01} \\ &= a(p_{11} - p_{00}) - b(p_{10} - p_{01}) \\ &= \mathrm{E}\, Z_1 - 2b(p_{10} - p_{01}). \end{aligned}$$

Therefore,

$$\begin{aligned} \mathrm{Cov}\,(Z_1, Z_2) &= \mathrm{E}(Z_1 Z_2) - \mathrm{E}(Z_1)\,\mathrm{E}(Z_2) \\ &= a^2(p_{11} + p_{00}) - b^2(p_{10} + p_{01}) \\ &\quad - [a(p_{11} - p_{00}) + b(p_{10} - p_{01})][a(p_{11} - p_{00}) - b(p_{10} - p_{01})] \\ &= a^2(p_{11} + p_{00}) - b^2(p_{10} + p_{01}) - a^2(p_{11} - p_{00})^2 + b^2(p_{10} - p_{01})^2 \\ &= a^2(p_{11} + p_{00} - (p_{11} - p_{00})^2) - b^2(p_{10} + p_{01} - (p_{10} - p_{01})^2). \end{aligned}$$

We also have

$$\begin{aligned}\mathrm{E}(Z_1^2) &= \mathrm{E}(Z_2^2)\\ &= a^2 p_{11} + b^2 p_{10} + a^2 p_{00} + b^2 p_{01}\\ &= a^2(p_{11}+p_{00}) + b^2(p_{10}+p_{01}).\end{aligned}$$

Therefore,

$$\begin{aligned}\operatorname{Var}(Z_1) &= \mathrm{E}(Z_1^2) - \mathrm{E}(Z_1)^2\\ &= a^2(p_{11}+p_{00}) + b^2(p_{10}+p_{01}) - [a(p_{11}-p_{00}) + b(p_{10}-p_{01})]^2\\ &= a^2(p_{11}+p_{00}) + b^2(p_{10}+p_{01}) - a^2(p_{11}-p_{00})^2\\ &\quad - 2ab(p_{11}-p_{00})(p_{10}-p_{01}) - b^2(p_{10}-p_{01})^2\\ &= a^2(p_{11}+p_{00}-(p_{11}-p_{00})^2) + b^2(p_{10}+p_{01}-(p_{10}-p_{01})^2) -\\ &\quad 2ab(p_{11}-p_{00})(p_{10}-p_{01}),\end{aligned}$$

and, using that $\mathrm{E}(Z_2) = \mathrm{E}(Z_1)$, and that $\mathrm{E}(Z_2) = \mathrm{E}(Z_1) - 2b(p_{10}-p_{01})$, we get

$$\begin{aligned}\operatorname{Var}(Z_2) &= \mathrm{E}(Z_2^2) - \mathrm{E}(Z_2)^2\\ &= \mathrm{E}(Z_1^2) - (\mathrm{E}(Z_1) - 2b(p_{10}-p_{01}))^2\\ &= \mathrm{E}(Z_1^2) - \mathrm{E}(Z_1)^2 + 4\,\mathrm{E}(Z_1)b(p_{10}-p_{01}) - 4b^2(p_{10}-p_{01})^2\\ &= \operatorname{Var}(Z_1) + 4[a(p_{11}-p_{00}) + b(p_{10}-p_{01})]\cdot b(p_{10}-p_{01}) - 4b^2(p_{10}-p_{01})^2\\ &= \operatorname{Var}(Z_1) + 4ab(p_{11}-p_{00})(p_{10}-p_{01}) + 4b^2(p_{10}-p_{01})^2 - 4b^2(p_{10}-p_{01})^2\\ &= \operatorname{Var}(Z_1) + 4ab(p_{11}-p_{00})(p_{10}-p_{01}).\end{aligned}$$

We want to calculate

$$\rho = \frac{\operatorname{Cov}(Z_1, Z_2)}{(\operatorname{Var}(Z_1)\operatorname{Var}(Z_2))^{1/2}}.$$

We first calculate the product $\operatorname{Var}(Z_1)\operatorname{Var}(Z_2)$. We now use $a = 1/b$. This simplifies the expressions to

$$\begin{aligned}\operatorname{Var}(Z_1) &= a^2(p_{11}+p_{00}-(p_{11}-p_{00})^2) + b^2(p_{10}+p_{01}\\ &\quad -(p_{10}-p_{01})^2) - 2ab(p_{11}-p_{00})(p_{10}-p_{01})\\ &= q - 2\Delta,\end{aligned}$$

where $q = a^2(p_{11}+p_{00}-(p_{11}-p_{00})^2)+b^2(p_{10}+p_{01}-(p_{10}-p_{01})^2)$ and $\Delta = (p_{11}-p_{00})(p_{10}-p_{01})$. Similarly, $\operatorname{Var}(Z_2) = q + 2\Delta$, and therefore,

$$\begin{aligned}\operatorname{Var}(Z_1)\operatorname{Var}(Z_2) &= (q-2\Delta)(q+2\Delta)\\ &= q^2 - 4\Delta^2\\ &= a^4c_1^2 + b^4c_2^2 + 2a^2b^2c_1c_2 - 4\Delta^2.\end{aligned}$$

Where $c_1 = p_{11}+p_{00}-(p_{11}-p_{00})^2$, and $c_2 = p_{10}+p_{01}-(p_{10}-p_{01})^2$. We note that c_1, c_2, Δ does not vary with a or b.

In terms of the introduced constants, we recognize that

$$\mathrm{Cov}\left(Z_1, Z_2\right) = a^2 c_1 - b^2 c_2.$$

We therefore have

$$\begin{aligned}\rho &= \frac{\mathrm{Cov}\left(Z_1, Z_2\right)}{\left(\mathrm{Var}\left(Z_1\right)\mathrm{Var}\left(Z_2\right)\right)^{1/2}} \\ &= \frac{a^2 c_1 - b^2 c_2}{\sqrt{a^4 c_1^2 + b^4 c_2^2 + 2a^2 b^2 c_1 c_2 - 4\Delta^2}}.\end{aligned}$$

Using $a = 1/b$, we see that

$$\rho = \frac{a^2 c_1 - b^2 c_2}{\sqrt{a^4 c_1^2 + b^4 c_2^2 + d}},$$

where $d = 2a^2 b^2 c_1 c_2 - 4\Delta^2 = d = 2c_1 c_2 - 4\Delta^2$ does not depend on a, b.

Case 1: Letting $b \to \infty$, giving the negative end-point. We use $a = 1/b$ and get

$$\begin{aligned}\rho &= \frac{a^2 c_1 - b^2 c_2}{\sqrt{a^4 c_1^2 + b^4 c_2^2 + d}} \\ &= \frac{b^{-2} c_1 - b^2 c_2}{\sqrt{b^{-4} c_1^2 + b^4 c_2^2 + d}} \\ &= \frac{b^{-4} c_1 - c_2}{\sqrt{b^{-4}(b^{-4} c_1^2 + b^4 c_2^2 + d)}} \\ &= \frac{b^{-4} c_1 - c_2}{\sqrt{b^{-8} c_1^2 + c_2^2 + b^{-4} d)}} \\ &\to \frac{-c_2}{|c_2|}.\end{aligned}$$

If $c_2 > 0$, this shows that $\rho \to -1$. We recall that $c_2^2 = (p_{10} + p_{01} - (p_{10} - p_{01})^2)^2 \geq 0$, and we only need to show that $c_2^2 \neq 0$. We have

$$\begin{aligned}p_{10} + p_{01} - (p_{10} - p_{01})^2 &= p_{10} + p_{01} - p_{10}^2 + 2p_{10}p_{01} - p_{01}^2 \\ &= (p_{10} - p_{10}^2) + (p_{01} - p_{01}^2) + 2p_{10}p_{01}.\end{aligned}$$

Since p_{01} and p_{10} are in $(0, 1)$, we have $p_{10}p_{01} > 0$. We have that $p_{10} > p_{10}^2$ and $p_{01} > p_{01}^2$, and therefore $p_{10} - p_{10}^2 > 0$ and $p_{01} - p_{01}^2 > 0$. Therefore, $c_2^2 \neq 0$.

Case 2: Letting $b \to 0^+$, giving the positive end-point. We use $b = 1/a$ and the exact same steps as above to get that

$$\begin{aligned}\rho &= \frac{a^2 c_1 - b^2 c_2}{\sqrt{a^4 c_1^2 + b^4 c_2^2 + d}} \\ &\to \frac{c_1}{|c_1|}.\end{aligned}$$

If $c_1 > 0$, this shows that $\rho \to 1$. We recall that $c_1^2 = (p_{11} + p_{00} - (p_{11} - p_{00})^2)^2 \geq 0$, and we only need to show that $c_1^2 \neq 0$. We have

$$\begin{aligned} p_{11} + p_{00} - (p_{11} - p_{00})^2 &= p_{11} + p_{00} - p_{11}^2 + 2p_{11}p_{00} - p_{00}^2 \\ &= (p_{11} - p_{11}^2) + (p_{00} - p_{00}^2) + 2p_{11}p_{00}. \end{aligned}$$

Since p_{00} and p_{11} are in $(0, 1)$, we have $p_{11}p_{00} > 0$. We have that $p_{11} > p_{11}^2$ and $p_{00} > p_{00}^2$, and therefore $p_{11} - p_{11}^2 > 0$ and $p_{00} - p_{00}^2 > 0$. Therefore, $c_1^2 \neq 0$.

Let ρ_b be the correlation of $Z(b) = Z(1/b, b)$ for $b > 0$. We recall

$$\rho_b = \frac{b^{-4}c_1 - c_2}{\sqrt{b^{-8}c_1^2 + c_2^2 + b^{-4}d)}}$$

and $c_1, c_2 > 0$. Since this is a continuous function with limits -1 and 1, every correlation in $(-1, 1)$ is attained by the intermediate value theorem.

Department of Economics, BI Norwegian Business School, Oslo, Norway 0484
Email address: steffeng@gmail.com

Department of Mathematics, University of Oslo, PB 1053, Blindern, NO-0316, Oslo, Norway
Email address: jonasmgj@math.uio.no

Department of Economics, BI Norwegian Business School, Stavanger, Norway 4014
Email address: njal.foldnes@gmail.com