PARTIAL IDENTIFICATION OF LATENT CORRELATIONS WITH BINARY DATA

Abstract. The tetrachoric correlation is a popular measure of association for binary data and estimates the correlation of a normal latent variable. Several influential simulation studies have concluded that the tetrachoric correlation and its extension to polytomous variables are quite robust against latent non-normality. Many practitioners have therefore used these methods even when latent normality was unknown. Recent papers have identified a flaw in these simulation studies, and shown that these estimators can be highly non-robust against non-normality. This motivates studying what can be said about latent correlations when the distribution of the latent variable is partly or fully unknown. We show that the bivariate case is studied without loss of generality unless knowledge not derivable from data is available. For bivariate data, we show that nothing can be said about the latent correlations unless we know more than what can be derived from the data, and we identify an interval constituting all latent correlations compatible with observed data when the marginals of the latent variables are known. These intervals are always too wide to be useful. Implications for tests for underlying normality are briefly discussed.

1. Introduction

Measures of association between ordinal variables, and models motivated from such measures, are important in statistics in general. A prominent approach since the time of Pearson (1900) postulates that samples from a random vector X with ordinal coordinates have been produced by discretizing an unobserved latent continuous random vector Z. In the dichotomous case, which our paper focuses on, we observe samples from $X = (X_1, \ldots, X_d)$ where

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

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.

The Pearson correlation matrix Σ of Z contains the latent correlations which are taken as measures of association for the ordinal variables X. In factor analysis, this latent variable framework has been widely adopted since Christoffersson (1975). A standard assumption when estimating Σ is that Z follows a multivariate normal distribution. Then Σ, or empirical estimates thereof, are referred to as the tetrachoric correlation matrix in the case of dichotomous data, and as the polychoric correlation matrix in the more general case of polytomous data. Exploratory and confirmatory factor analysis for ordinal data is routinely performed using the tetrachoric or polychoric correlation matrix (Flora et al., 2012). This is the default method in popular software such as `mplus` (Muthén & Muthén, 2012) and the popular R (R Core Team, 2013) packages `lavaan` and `psych` (Rosseel, 2012; Revelle, 2018). Principal component analysis for dichotomous data may also be done using the tetrachoric correlation matrix (Kolenikov & Angeles, 2009; Klapper et al., 2013; Howe et al., 2012). A search in Google Scholar for "polychoric" and "factor analysis" gives 12,400 results, and a search for "tetrachoric" and "factor analysis" gives 6,780 results. Given the popularity of factor analysis, structural equation modelling and principal component analysis based on tetrachoric or polychoric correlation matrices, it is important to study the consequences of weakening the normality assumption, an assumption which is not always justified.

Estimation and inference for Σ assuming that Z is normal will be called normal theory methods. Most simulation studies on whether normal theory methods are robust towards underlying non-normality have relied on the non-normal simulation method of Vale & Maurelli (1983) to generate Z with covariance matrix Σ. Normal theory methods are then applied to the resulting sample from X. This includes the influential studies of Flora & Curran (2004) and Rhemtulla et al. (2012), see Grønneberg & Foldnes (2019) for further references. The resulting consensus based on these studies has been that normal theory methods are fairly robust towards underlying non-normality. Grønneberg & Foldnes (2019) showed that generating Z using the method of Vale & Maurelli (1983) usually produces an X which is numerically equal to the discretization of an exactly normal random variable with a slightly different correlation matrix than Σ. This surprising finding may occur since the distribution of Z is not identified from the distribution of X. Foldnes & Grønneberg (2019a) gives a basic analysis of identification in ordinal models, and suggest an improved non-normal simulation approach using the simulation procedure of Grønneberg & Foldnes (2017). The new simulation approach indicates that the normal theory estimation and inference for latent correlations are highly non-robust towards underlying non-normality (Foldnes & Grønneberg, 2019a,b, 2020).

The present paper studies what can be said about Σ when normality is not assumed. We restrict attention mainly to the case when the coordinates of X are dichotomous, with a special focus on the bivariate case. In the bivariate case we show that nothing can be said about the correlation of Z unless we take into account what we call substantial knowledge of the distribution of Z. Substantial knowledge means knowledge not derivable from the distribution of the observations X. If substantial knowledge justifies treating the marginal distributions as known, we identify a set which contains all possible Pearson correlations of Z that are compatible with observed data, an analysis called partial identification. A similar analysis is done for Spearman's rho, and the resulting sets have 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. A partial identification analysis is also 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 & Pearson (1922). We end our paper by showing that without substantial knowledge, multivariate information cannot help identify the pairwise correlations of Z, and the bivariate problem is studied without loss of generality. Tests for underlying normality are also discussed in light of our results.

The normality assumption made by Pearson (1900) has to be made based on substantive knowledge of Z, see the discussion in Pearson & Heron (1913, p. 161–162). Tetrachoric correlations are routinely used in cases where such substantial knowledge is lacking. This is especially striking for exploratory factor analyses, where the aim is to discover correlational factors in data precisely in cases without much knowledge of the underlying latent variable. Our results indicate that such practice may lead to unwarranted conclusions.

Before a final section with concluding remarks we discuss the interpretation of tests for underlying normality in light of our results, and argue that when underlying normality is not known, these tests can only be used to dismiss the plausibility of underlying normality. If the tests indicate compatibility with normality, our results imply that mere compatibility with normality has no implications without substantial knowledge.

The present paper shows that the assumption of underlying normality has to be justified through substantial knowledge, and that if such knowledge is lacking, the results of empirical investigations can be strongly driven by this groundless assumption.

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, as explained in “Appendix B” in the online appendix.

2. Partial identification with 2 × 2 tables

2.1. Introduction to partial identification, and an introductory example.

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. For a thorough literature review, see Tamer (2010) and the book Manski (2003).

We here 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. This partial identification problem was solved by Höffding (1940) and Fréchet (1951). A modern presentation is given in Nelsen (2007). See also the influential papers Lehmann (1966) and Whitt (1976).

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 $[0,1]$. Sklar's theorem, see Sklar (1959) and Nelsen (2007, Theorem 2.3.3), states that there exists a copula C such that for any x, y 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 for $M[F_1,F_2](x_1,x_2) = M(F_1(x_1),F_2(x_2))$, 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.

For a set $\mathcal{P}$ of bivariate distributions with finite standard deviations, let $\rho(\mathcal{P}) = \{\rho(F) : F \in \mathcal{P}\}$ where $\rho(F)$ is the Pearson correlation of F. Let $\mathcal{P}$ be the set of distributions with marginals F_1, F_2. We now calculate $\rho(\mathcal{P})$. The Höffding (1940) formula for correlation gives

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

where $\operatorname{sd}(F_1), \operatorname{sd}(F_2)$ are the standard deviations of F_1, F_2. Eq. (3) then implies that $\rho(F) \in [\rho(W[F_1,F_2]), \rho(M[F_1,F_2])]$. An additional 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.

2.2. Latent correlations in 2×2 tables.

We now turn to eq. (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 parameterized by the 2×2 table

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

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, see eq. (2) above, 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. As an extension of the notation in Section 2.1, define the set $\rho\left(\mathcal{P}; p\right)$ as the set of latent correlations compatible with p and $\mathcal{P}$. That is,

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

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, C_ρ is the normal copula, parameterized by ρ, the Pearson correlation of C_ρ when combined with normal marginals. From Joe (1997, Section 5.1) and Almeida & 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 & Mouchart (2014), the same argument yields identifiability when assuming other marginals and other one-dimensional parametric copula classes $\{C_\theta : \theta \in \Theta\}$. We then require that $\theta \mapsto C_\theta(u, v)$ is increasing for each $0 < u, v < 1$, a property fulfilled by all copulas catalogued in Section 5.1 of Joe (1997).

Theorem 1 calculates $\rho(\mathcal{P}; p)$ when $\mathcal{P}$ has no restrictions.

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 eq. (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. See the upcoming Proposition 2.

2.3. 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 & Heron (1913), such assumptions must be justified by external information on the variable Z. Let us now 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 when the coordinates of Z have been studied separately, and from this the likely distribution can be deduced, but the joint distribution is unknown.

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.

The following result calculates $\rho(\mathcal{P};p)$. Similar to the introductory calculations in Section 2.1, the following result follows from a bound on the copula of members of $\mathcal{P}$ (Nelsen, 2007, Theorem 3.2.2.).

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 $M[F_1,F_2;p](x_1,x_2) = M_p(F_1(x_1),F_2(x_2))$ 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*

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

Spearman's rho is the Pearson correlation of a copula (Nelsen, 2007, Theorem 5.1.6), and is therefore not dependent on the unidentified marginals. Let $\mathcal{R}(p)$ be the set of Spearman's rho values compatible with p. Numerically, $\mathcal{R}(p) = \rho(\mathcal{P}, p)$ when $\mathcal{P}$ is the set of distributions with uniform marginals on $[0,1]$. We identify the following compact algebraic formula.

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

We now give a numerical illustration. Assume $p_{00} = 0.2, p_{01} = 0.4$ and $p_{10} = 0.1$. Assuming Z has a normal copula, Spearman's rho is 0.14. Assuming Z is fully normal, Pearson's correlation is 0.15. If only the marginals are assumed to be normal, $\rho(\mathcal{P};p) = [-0.88, 0.93]$. If the distribution of Z is fully unknown, $\rho(\mathcal{P};p) = (-1,1)$

by Theorem 1, and $\mathcal{R}(p) = [-0.82, 0.88]$ by Proposition 2. Further illustrations are found in "Appendix A".

2.4. **Partial identification when Z_2 is directly observed.**

We now assume Z_2 is directly observed. When Z is normal, this gives the biserial correlation of Pearson (1909), see also Tate (1955b,a). 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 not F_1 nor the full 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. We only study the case when the marginal distribution of Z_1 is assumed given. 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. Then $\rho(\mathcal{P}; p) = [\rho(W[F_1, F_2; p]), \rho(M[F_1, F_2; p])]$ where $M[F_1, F_2; p](x_1, x_2) = M_p(F_1(x_1), F_2(x_2))$ 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*

$$M_p(u, v) = \min\left(u, v, C(F_1(\tau_1), v) + (u - F_1(\tau_1))^+\right)$$
$$W_p(u, v) = \max\left(0, u + v - 1, C(F_1(\tau_1), v) - (F_1(\tau_1) - u)^+\right).$$

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, but still too wide to be useful. As in Section 2.3, a partial identification analysis of Spearman's rho is given by the above analysis when assuming uniform marginals.

3. The multivariate dichotomous 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 with knowing just 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.

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 2.4. 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 & 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).

3.2. Increasing the dimensionality cannot help identify parameters when substantial knowledge is lacking.

In this sub-section we substantially increase the mathematical generality of our discussion. Specializing to the setting of previous sections is easy, as shown in the coming Example 1.

Let $S = (\Omega, \Sigma)$ be a measure space. We assume S is a Borel space, so that there exists a bijection f between Ω and a Borel subset of $[0, 1]$ such that both f and f^{-1} are measurable (Kallenberg, 2006, p. 7). 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).

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.

Theorem 2. *Assume S is a rich Borel space. Then $\gamma\left(P_X\right) = \gamma\left(P_{X,Y}\right)$.*

3.3. On the interpretation of tests for underlying normality.

For 2×2 tables, we saw in Section 2.2 that there is a bijection between the table on the 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 & Hofacker (1988), we may increase the dimensionality, and study trivariate dichotomous 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 (Maydeu-Olivares, 2006; Foldnes & Grønneberg, 2019b). 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: 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, even if a test or even the exact distribution of X is compatible with having been generated from an underlying normal latent variable, we cannot from this conclude that Z actually is normal. In fact, in our partial identification results, there is nothing special that happens when X has a distribution compatible with normality, and the resulting sets still are too large to be informative.

4. Concluding remarks

We have shown that a great deal substantial knowledge is required to usefully analyse binary data through the perspective of latent correlations. As mentioned in Section 2.1, 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 & Grønneberg (2019b,a) 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 especially Figure 2 in the introductory example of Foldnes & Grønneberg (2019b). The present paper exactly identifies what can be said about the latent correlation if we only know the discretized data, which is the situation in a real investigation. If no substantial knowledge of the distribution of Z is known, which is often 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 large. Therefore, smaller and more informative intervals are only available by imposing restrictions on the dependency structure among the underlying latent variables. This kind of substantial knowledge seems hard to justify in most practical applications. We therefore must conclude that the normal-theory tetrachoric correlation coefficient may not be an informative measure of association for binary variables.

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. Another extension would be to take into account other types of knowledge of Z, such as the structural equations between the coordinates of Z assumed in factor analysis and item response model.

Appendix A. Technical proofs and numerical illustrations

A.1. **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 1 and 2. Full computational details are given in the accompanying `R` scripts. In Figure 1, we have $\tau_1 = \tau_2 = 0$, which is a best case scenario. Figure 2 shows a more typical situation, where the length of all bounds are close to the maximal length of such an interval, namely 2. Figure 2 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 1 and Figure 2, points very close or at the endpoints $\rho = \pm 1$ 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

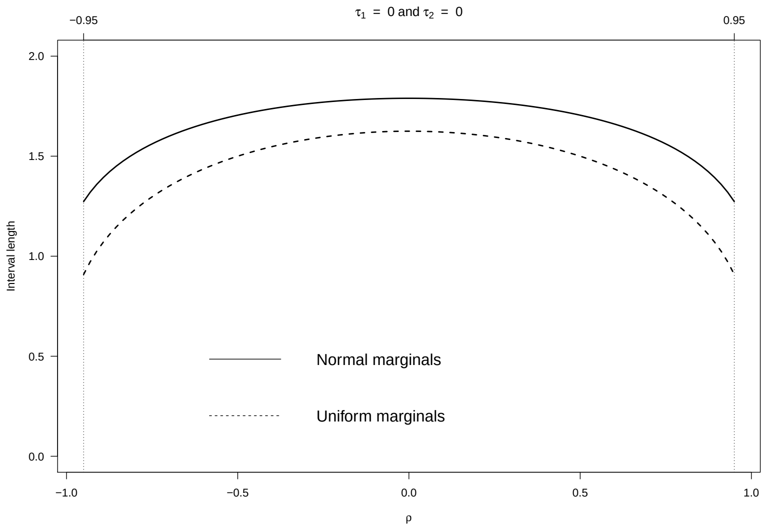

FIGURE 1. Length of bounds for $\tau_1 = 0, \tau_2 = 0$ based on normal or uniform marginal assumptions. The graph does not cover points close to $\rho = \pm 1$.

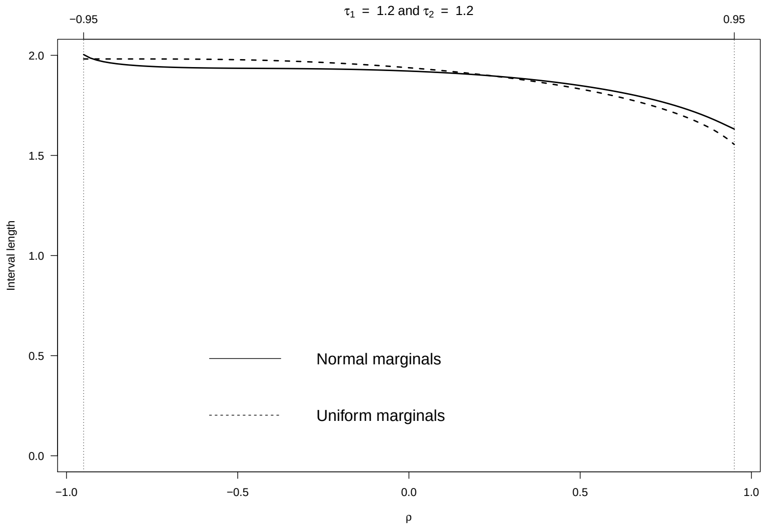

FIGURE 2. Length of bounds for $\tau_1 = 1.2, \tau_2 = 1.2$ based on normal or uniform marginal assumptions. The graph does not cover points close to $\rho = \pm 1$.

Core Team, 2018) packages `copula` (Hofert et al., 2013), `cubature` (Narasimhan et al., 2018), and `copBasic` (Asquith, 2019).

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 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

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

Then $\text{pr}[Z(a,b) \in A_{ij}] = \text{pr}[X = (i,j)] = p_{ij}$ 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 confirmed algebraically in Section C on page 1. At the end of that section, 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

$$\alpha\rho_L + (1-\alpha)\rho_U = \rho. \tag{6}$$

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 2.1.

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

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

using equation (6). □

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 $\int_{[0,1]^2} C_U(u, v)\,dudv$. Define the set $A_F = [a, d] \times [a, d]$. Then

$$\int_{[0,1]^2} C_U(u, v)\,dudv = \int_{A_F} C_U(u, v)\,dudv + \int_{A_F^C} C_U(u, v)\,dudv \tag{7}$$

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

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

the second integral in (7) equals

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

The next part is $\int_{A_F} C_U(u,v)\,dudv$. 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

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

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

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

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

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

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

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

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

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

hence

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

as claimed. The lower bound follows by duality. □

The reasoning behind the decomposition can be seen in Figure 3, where each colour correspond to a continuous part of the piece-wise continuous function $C_U(u,v)$.

A.3. Proofs for Section 2.4.

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 $p(x_1,y) = \mathrm{P}(X_1 = x_1, Z_2 \le y_2)$, for $x_1 = 0,1$ and y a real number. We have that $p(0,y) = \mathrm{P}(X_1 = 0, Z_2 \le y) = \mathrm{P}(Z_1 \le \tau_1, Z_2 \le y) = C(F_1(\tau_1), F_2(y))$. Since $p(0,y) + p(1,y) = F_2(y)$, and therefore $p(1,y) =$

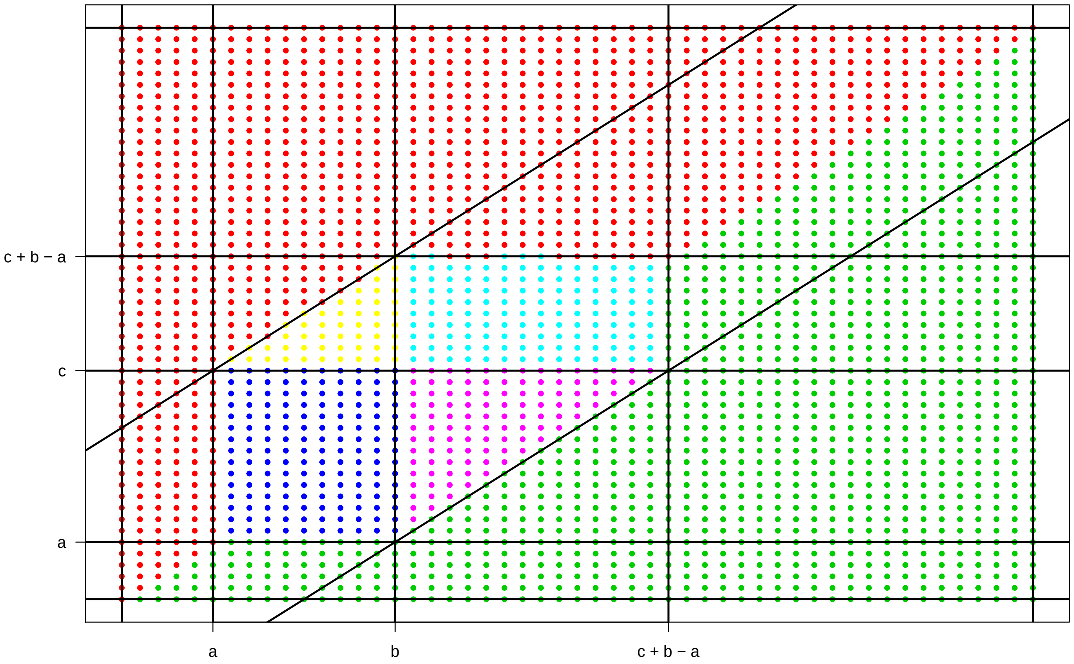

FIGURE 3. Colour-coded graph of the bound copula. Each colour correspond to a continuous part of the piece-wise continuous function $C_U(u, v)$.

$F_2(y) - p(0, y)$, 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

$$C(u, v) = p(0, F_2^{-1}(v)) \quad ((u, v) \in \mathcal{U} = \{(u, v) \mid u = F_1(\tau_1), 0 \leq v \leq 1\}) . \tag{8}$$

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* (8) *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

$$C_{U,\mathcal{U}}(u,v) = \min(u, v, \min_{b}[C(F_1(\tau_1), b) + (u - F_1(\tau_1))^+ + (v-b)^+]), \tag{9}$$

$$C_{L,\mathcal{U}}(u,v) = \max(0, u+v-1, \max_{b}[C(F_1(\tau_1), b) - (F_1(\tau_1) - u)^+ - (b-v)^+]). \tag{10}$$

Moreover, both $C_{L,\mathcal{U}}$ and $C_{U,\mathcal{U}}$ are copulas that satisfy equation (8).

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,*

$$C_{U,\mathcal{U}}(u,v) = \min(u, v, \min_{b\in[0,1]}[C(F_1(\tau_1), b) + (u - F_1(\tau_1))^+ + (v-b)^+]), \tag{11}$$

$$= \min(u, v, C(F_1(\tau_1), v) + (u - F_1(\tau_1))^+), \tag{12}$$

and

$$C_{L,\mathcal{U}}(u,v) = \max(u, u+v-1, \max_{b\in[0,1]}[C(F_1(\tau_1), b) - (F_1(\tau_1) - u)^+ - (b-v)^+]), \tag{13}$$

$$= \max(0, u+v-1, C(F_1(\tau_1), v) - (F_1(\tau_1) - u)^+). \tag{14}$$

From this, the Höffding representation from eq. (3) in Section 2.1 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. (8). Now from Lemma 1, we have that both W_p and M_p fulfil eq. (8), 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 (8).

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 we show 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 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. 17). 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. 17). 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.**

Proof of Theorem 2. The inclusion $\gamma\left(P_{X,Y}\right)\subseteq\gamma\left(P_X\right)$ is true for any Z and S. Choose a P_X, a $P_{X,Y}$ compatible with P_X, and a $P_Z\in\gamma\left(P_X\right)$. We must show $P_Z\in\gamma\left(P_{X,Y}\right)$, 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 $f_\theta\left(Z\right)$ when Z is distributed according to P_Z. By Corollary 6.11 of Kallenberg (2006), there is a variable Z' in S such that $X=f_\theta\left(Z'\right)$ 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

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

By the Höffding formula for covariance, we have

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

where

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

Here, J_1 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 $\int_{\mathbb{R}^2} g\left(u,v\right) dudv$ 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

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

The domains of integration can be seen in Figure 3. 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 2.1, 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**, 3–25.

Asquith, W. (2019). *copBasic—General Bivariate Copula Theory and Many Utility Functions*. R package version 2.1.4.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika* **40**, 5–32.

Flora, D., LaBrish, C. & Chalmers, R. (2012). Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. *Frontiers in Psychology* **3**, 55.

Flora, D. B. & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods* **9**, 466–491.

Foldnes, N. & Grønneberg, S. (2019a). On identification and non-normal simulation in ordinal covariance and item response models. *Psychometrika* **84**, 1000–1017.

Foldnes, N. & Grønneberg, S. (2019b). Pernicious polychorics: The impact and detection of underlying non-normality. *Structural Equation Modeling* Forthcoming.

Foldnes, N. & Grønneberg, S. (2020). Structural equation modeling with ordinal data: The impact of thresholds and underlying non-normality. *Psychological Methods* Submitted.

Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Annales de l'Université de Lyon, Section A, Series 3* **14**, 53–77.

Fréchet, M. (1958). Remarques au sujet de la note précédente. *Compte Rendus des Séances Academie des Sciences* **246**, 2719–2720.

Grønneberg, S. & Foldnes, N. (2017). Covariance model simulation using regular vines. *Psychometrika* , 1–17.

Grønneberg, S. & Foldnes, N. (2019). A problem with discretizing Vale-Maurelli in simulation studies. *Psychometrika* **84**, 554–561.

Hofert, M., Kojadinovic, I., Maechler, M. & Yan, J. (2013). *copula: Multivariate Dependence with Copulas*. R package version 0.999-7.

Höffding, W. (1940). Masstabinvariante korrelationstheorie. *Schriften des Mathematischen Instituts und Instituts fur Angewandte Mathematik der Universitat Berlin* **5**, 181–233.

Howe, L. D., Galobardes, B., Matijasevich, A., Gordon, D., Johnston, D., Onwujekwe, O., Patel, R., Webb, E. A., Lawlor, D. A. & Hargreaves, J. R. (2012). Measuring socio-economic position for epidemiological studies in low- and middle-income countries: a methods of measurement in epidemiology paper. *International Journal of Epidemiology* **41**, 871–886.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. Chapman & Hall/CRC.

Kallenberg, O. (2006). *Foundations of Modern Probability*. Springer, 2nd ed.

Klapper, L., Lusardi, A. & Panos, G. A. (2013). Financial literacy and its consequences: Evidence from Russia during the financial crisis. *Journal of Banking & Finance* **37**, 3904–3923.

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**, 128–165.

Lehmann, E. L. (1966). Some concepts of dependence. *The Annals of Mathematical Statistics* **37**, 1137–1153.

Manski, C. F. (2003). *Partial identification of probability distributions.* Springer Science & Business Media.

Maydeu-Olivares, A. (2006). Limited information estimation and testing of discretized multivariate normal structural models. *Psychometrika* **71**, 57–77.

Muthén, B. & Hofacker, C. (1988). Testing the assumptions underlying tetrachoric correlations. *Psychometrika* **53**, 563–577.

Muthén, B. & Muthén, L. (2012). Mplus version 7: User's guide.

Narasimhan, B., Johnson, S. G., Hahn, T., Bouvier, A. & Kiêu, K. (2018). *cubature: Adaptive Multivariate Integration over Hypercubes.* R package version 2.0.3.

Nelsen, R. B. (2007). *An Introduction to Copulas.* Springer Science & Business Media.

Owen, D. B. (1980). A table of normal integrals: A table. *Communications in Statistics-Simulation and Computation* **9**, 389–419.

Pearl, J. (2009). *Causality.* Cambridge University Press.

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* **196**, 1–47.

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**, 96–105.

Pearson, K. & Heron, D. (1913). On theories of association. *Biometrika* **9**, 159–315.

Pearson, K. & Pearson, E. S. (1922). On polychoric coefficients of correlation. *Biometrika* **6**, 127–156.

R Core Team (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

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

Revelle, W. (2018). *psych: Procedures for Psychological, Psychometric, and Personality Research.* Northwestern University, Evanston, Illinois. R package version 1.8.12.

Rhemtulla, M., Brosseau-Liard, P. É. & Savalei, V. (2012). When can categorical variables be treated as continuous? a comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods* **17**, 354–373.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software* **48**, 1–36.

Sklar, M. (1959). *Fonctions de Répartition à n Dimensions et Leurs Marges.* Université Paris 8.

Tamer, E. (2010). Partial identification in econometrics. *Annual Review of Economics* **2**, 167–195.

Tankov, P. (2011). Improved Fréchet bounds and model-free pricing of multi-asset options. *Journal of Applied Probability* **48**, 389–403.

Tate, R. F. (1955a). Applications of correlation models for biserial data. *Journal of the American Statistical Association* **50**, 1078–1095.

Tate, R. F. (1955b). The theory of correlation between two continuous variables when one is dichotomized. *Biometrika* **42**, 205–216.

Vale, C. D. & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika* **48**, 465–471.

Vaswani, S. (1950). Assumptions underlying the use of the tetrachoric correlation coefficient. *Sankhyā: The Indian Journal of Statistics* , 269–276.

Whitt, W. (1976). Bivariate distributions with given marginals. *The Annals of Statistics* **4**, 1280–1289.

ONLINE SUPPLEMENTARY MATERIAL: APPENDIX

Appendix B. Explanation of included R-scripts

In the online supplementary material, we include six enumerated R-scripts, and some additional supporting R material. Code to reproduce the numerical illustration of Section 2.3 is found in 1.R. A verification of the computational simplifications used in the utility functions is found in 2.R and 3.R. Computations for the numerical illustration of Proposition 3 is found in 4.R. Code for a numerical exploration of the lengths of the bounds from Proposition 1 with standard normal and uniform marginals is given in 5.R. Some output from and comments on this analysis are found in the upcoming Section A.1, p.10. Code for an approximation of the minimal lengths of the intervals (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) is found in 6.R. In the code we use the R packages copula (Hofert et al., 2013), cubature (Narasimhan et al., 2018), and copBasic (Asquith, 2019).

Appendix C. 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

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

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,

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

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,

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

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

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

We want to calculate

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

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

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

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, $\operatorname{Cov}(Z_1) = q + 2\Delta$, and therefore,

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

Where $d = (p_{11} + p_{00} - (p_{11} - p_{00})^2)(p_{10}p_{01} - (p_{10} - p_{01})^2) - 4\Delta^2$, $c_1^2 = (p_{11} + p_{00} - (p_{11} - p_{00})^2)^2$, and $c_2^2 = (p_{10} + p_{01} - (p_{10} - p_{01})^2)^2$. In terms of the introduced constants, we recognize that

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

We therefore have

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

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

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

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

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

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.