Inducing a desired value of correlation between two point-scale variables

Alessandro Barbiero*

Abstract: Focusing on point-scale random variables, i.e., variables whose support space is given by the first m integers, we discuss how a desired value of Pearson's correlation can be induced between two assigned probability distributions, which are linked to a joint distribution via a copula function. After recalling how the value of the desired ρ is not free to vary within $[-1, +1]$, but is bounded to a narrower interval depending on the two marginal distributions, we devise a procedure to recover the same feasible value ρ for different dependence structures, focusing on one-parameter copulas encompassing the entire dependence spectrum.

Keywords: Attainable correlations, Copula, Ordinal variables.

1. Introduction

Datasets arising in the social sciences often contain ordinal variables. In particular, Likert scale items are those where, given a statement, the subject indicates strong agreement, agreement, neutrality, disagreement, or strong disagreement. A relevant example derives from questionnaires about customers' satisfaction. Satisfaction can be regarded as a multidimensional latent (i.e., unobservable) phenomenon, involving several aspects that can be usually measured using graded scales, such as "Very dissatisfied", "Dissatisfied", "Neither satisfied nor dissatisfied", "Satisfied" and "Very satisfied". Likert scales are often treated as interval scales, by scoring the ordered categories using the integers 1, 2, 3, ...; this amounts to assuming that the categories are evenly spaced. Though representing just an arbitrary assumption, it is quite a common and accepted practice as well as proceeding to further multivariate statistical analyses handling them as (correlated) univariate discrete variables.

*Department of Economics, Management and Quantitative Methods, University of Milan, alessandro.barbiero@unimi.it

Now, one may be interested in building and simulating a multivariate random vector whose univariate components are point-scale variables with assigned marginal distributions and whose pairwise correlations are chosen a priori as well. In the following we will limit our analysis to the bivariate case, which is by far easier to deal with, but whose results, with some caution, can be extended to the multivariate context. We consider two point scale random variables (r.v.s), X_1 and X_2, defined over the support spaces $\mathcal{X}_1 = \{1, 2, \ldots, m_1\}$ and $\mathcal{X}_2 = \{1, 2, \ldots, m_2\}$, respectively, with probability mass functions $p_1(i) = P(X_1 = i), i = 1, \ldots, m_1$, and $p_2(i) = P(X_2 = j), j = 1, \ldots, m_2$. We want to determine *some* bivariate probability mass function $p_{ij} = P(X_1 = i, X_2 = j), i = 1, \ldots, m_1; j = 1, \ldots, m_2$ such that its margins are p_1 and p_2 and the correlation ρ_{X_1,X_2} is equal to an assigned ρ. In order to give an answer to this question, we have first to recall two properties of Pearson's correlation, which applies to both the continuous and, to even a larger extent, the discrete case; this is the topic of Section 2. In Section 3, we briefly recall how to build copula-based bivariate discrete distributions. Section 4 is devoted to the description of the proposed procedure for inducing a desired value of correlation between two point-scale variables. Section 5 illustrates an application to CUB distributions.

2. Attainable correlations between two random variables

A first important but often neglected feature of Pearson's correlation is that given two marginal cumulative distribution functions (c.d.f.s) F_1 and F_2 and a correlation value $\rho \in [-1, +1]$, it is not always possible to construct a joint distribution F with margins F_1 and F_2, whose correlation is equal to the assigned ρ. We can state the following result, concerning "attainable correlations" (see McNeil et al. 2005, pp.204-205). Let (X_1, X_2) be a random vector marginal cdfs F_1 and F_2 and an unspecified joint cdf; assume also that $\text{Var}(X_1) > 0$ and $\text{Var}(X_2) > 0$. The following statements hold:

1. The attainable correlations form a closed interval $[\rho_{\min}, \rho_{\max}]$ with $\rho_{\min} < 0 < \rho_{\max}$.

2. The minimum correlation $\rho = \rho_{\min}$ is attained if and only if X_1 and X_2

are countermonotonic. The maximum correlation $\rho = \rho_{\max}$ is attained if and only if X_1 and X_2 are comonotonic.

3. $\rho_{\min} = -1$ if and only if X_1 and $-X_2$ are of the same type, and $\rho_{\max} = 1$ if and only if X_1 and X_2 are of the same type.

For point-scale r.v.s X_1 and X_2, it is then clear that the maximum correlation is $+1$ if and only if they are identically distributed; whereas the minimum correlation can never be -1. The values $\rho_{\min}$ and $\rho_{\max}$ can be computed by building the cograduation and countergraduation tables (see, Ferrari and Barbiero, 2012, for an example of calculation).

A second fallacy of Pearson's correlation can be resumed as follows: Given two margins F_1 and F_2 and a feasible linear correlation ρ, the joint distribution F having margins F_1 and F_2 and correlation ρ is not unique. In other terms, the marginal distributions and pairwise correlations of a r.v. do not univocally determine its joint distribution. Even if this second fallacy may represent a limit from one side, on the other side represents a form of flexibility, since it means that given two point-scale r.v.s and a consistent value of ρ, there are different (possibly, infinite) ways to join them into a bivariate distribution with that value of correlation, as we will see in the next two sections.

3. Generating bivariate discrete distributions via copulas

How can we generate from a bivariate distribution respecting the assigned margins and correlation? Using copulas represent a straightforward solution. A d-dimensional copula is a joint c.d.f. in $[0, 1]^d$ with standard uniform c.d.f.s U_j, $j = 1 \ldots, d$:

$$C(u_1, \ldots, u_d) := P(U_1 \leq u_1, \ldots, U_d \leq u_d).$$

The importance of copulas in the study of multivariate c.d.f.s is summarized by the Sklar's theorem (see McNeil et al., 20005), whose version for $d = 2$ states that if F_1 and F_2 are the c.d.f.s of the point-scale r.v.s X_1 and X_2, the function

$$F(i, j) = C(F_1(i), F_2(j)), i = 1, \ldots, m_1; j = 1, \ldots, m_2 \qquad (1)$$

defines a valid joint c.d.f. over $\mathcal{X}_1 \times \mathcal{X}_2$, whose margins are F_1 and F_2. The only requirement we have to impose is that the copula C is able to encompass the entire range of dependence, from perfect negative dependence ($\rho_{\min}$) to perfect positive dependence ($\rho_{\max}$). Among copulas enjoying this property, we recall the Gauss copula, the Frank copula, and the Plackett copula.

The Gauss copula

The d-variate Gauss copula is the copula that can be extracted from a d-variate normal vector $\boldsymbol{Y}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ and is exactly the same as the copula of $\boldsymbol{X} \sim N_d(\mathbf{0}, P)$, where P is the correlation matrix of $\boldsymbol{Y}$. In two dimensions, it can be expressed, for $\rho \neq \pm 1$, as:

$$C^{Ga}(u_1, u_2) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho_{Ga}^2}} e^{-\frac{s_1^2 - 2\rho_{Ga} s_1 s_2 + s_2^2}{2(1-\rho^2)}} \mathrm{d}s_1 \mathrm{d}s_2.$$

Independence, comonotonicity, and countermonotonicity copulas are special cases of the bivariate Gauss copula (for $\rho_{Ga} = 0$, $\rho_{Ga} = 1$, and $\rho_{Ga} = -1$, respectively).

The Frank copula

The one-parameter bivariate Frank copula is defined as

$$C^F(u_1, u_2; \theta) = -\frac{1}{\kappa} \ln \left[1 + \frac{(e^{-\kappa u_1} - 1)(e^{-\kappa u_2} - 1)}{e^{-\kappa} - 1} \right],$$

with $\kappa \neq 0$. For $\kappa \to 0$, we have that the Frank copula reduces to the independence copula; for $\kappa \to \infty$, it tends to the comonotonicity copula; for $\kappa \to -\infty$, it tends to countermonotonicity copula.

The Plackett copula

The one-parameter bivariate Plackett copula is defined as

$$C^P(u_1, u_2; \kappa) = \frac{1 + (\theta - 1)(u_1 + u_2) - \sqrt{[1 + (\theta - 1)(u_1 + u_2)]^2 - 4\theta(\theta - 1)u_1 u_2}}{2(\theta - 1)},$$

with $\theta > 0$. When $\theta = 1$, it reduces to the independence copula, whereas for $\theta \to 0$ it tends to the countermonotonicity copula and for $\theta \to \infty$ to the comonotonicity copula.

4. Inducing a desired value of correlation between two point-scale random variables

The bivariate p.m.f. corresponding to (1) can be computed as

$$p(i,j) = F(i,j) - F(i-1,j) - F(i,j-1) + F(i-1,j-1) \qquad (2)$$

Computing the correlation coefficient for a bivariate point-scale variable (2) is very easy; since

$$\rho_{x_1 x_2} = (\mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2))(\text{Var}(X_1)\text{Var}(X_2))^{-1/2} \qquad (3)$$

with $\mu_1 = \mathbb{E}(X_1) = \sum_{i=1}^{m_1} i p_1(i)$, $\text{Var}(X_1) = \sum_{i=1}^{m_1}(i-\mu_1)^2 p_1(i)$ (analogous results hold for X_2), and $\mathbb{E}(X_1 X_2) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} ijp(i,j)$.

Once the marginal distributions of X_1 and X_2 are assigned, their correlation coefficient ρ_{X_1,X_2} will depend only on the copula parameter $\theta \in [\theta_{\min}, \theta_{\max}]$; this relationship may be written in an analytical or numerical form, say $\rho_{X_1,X_2} = g(\theta)$. Since the function g is not usually analytically invertible, inducing a desired value of correlation ρ between two point-scale variables, falling in $[\rho_{\min}, \rho_{\max}]$, by setting an appropriate value of the θ, is a task that can be generally done only numerically, by finding the (unique) root of the equation $g(\theta) - \rho_{X_1,X_2} = 0$. If ρ_{X_1,X_2} is a monotone increasing function of the copula parameter, it can be implemented by resorting to the following iterative procedure (see Ferrari and Barbiero, 2012; Barbiero and Ferrari, 2015b):

1. Set $\theta^{(0)} = \theta^{\Pi}$ (with θ^{Π} being the value of θ for which the copula C reduces to the independence copula); $\rho^{(0)} = 0$.
2. Set $t = 1$ and $\theta = \theta^{(t)}$, with $\theta^{(t)}$ some value strictly greater (smaller) than $\theta^{(0)}$ if $\rho > (<)0$
3. Compute $F(i,j;\theta^{(t)})$ using (1)

4. Compute $p(i, j; \theta^{(t)})$ using (2)

5. Compute $\rho^{(t)}$ using (3)

6. If $|\rho^{(t)} - \rho| < \epsilon$ stop; else
   set $t \leftarrow t + 1$,
   $\theta^{(t)} \leftarrow \min(\theta_{\max}, \theta^{(t-1)} + m(\rho - \rho^{(t-1)}))$ if $\rho > 0$, or
   $\theta^{(t)} \leftarrow \max(\theta_{\min}, \theta^{(t-1)} + m(\rho - \rho^{(t-1)}))$ if $\rho < 0$,
   with $m = \dfrac{\theta^{(t-1)} - \theta^{(t-2)}}{\rho^{(t-1)} - \rho^{(t-2)}}$; go back to 3.

The above heuristic algorithm makes sense if g is a monotone increasing function, which is often the case: for the Gauss, Frank, and Plackett copulas, the linear correlation is an increasing function of the dependence parameter θ, keeping fixed the two marginal distributions. The advantage of the proposed algorithm stands in the two following (connected) features: i) in the capacity of finding the appropriate value of θ without making use of any sample from the two marginal distributions, ii) in the possibility of controlling a priori the error ϵ (absolute difference between target and actual values of ρ_{X_1,X_2}); setting ϵ equal to 10^{-7} generally allows to recover θ in a few steps.

Existing procedures for solving the same problem are available in the literature, but do not enjoy the two features above mentioned. For example, the proposal by Demirtas (2006), requires the preliminary generation of a "huge" bivariate sample of binary data.

5. Application to CUB random variables

A CUB r.v. X is defined as the mixture of a shifted Binomial and a discrete Uniform distribution over the support $\{1, 2, \ldots, m\}$, for $m > 3$ (Piccolo, 2003). Its probability mass function is

$$P(X = i) = \pi \binom{m-1}{i-1} \xi^{m-j} (1-\xi)^{j-1} + (1-\pi)\frac{1}{m}$$

with (π, ξ) a parameter vector with the parametric space $(0, 1] \times [0, 1]$.

Corduas (2011) proposed using the Plackett distribution in order to construct a one parameter bivariate distribution from CUB margins; this proposal

was later investigated by Andreis ad Ferrari (2012), also in a multivariate direction. Here, we reprise and extend these attempts of constructing a bivariate CUB r.v. Let suppose we want to build a bivariate model with margins $X_1 \sim \text{CUB}(m_1 = 5, \pi_1 = 0.4, \xi_1 = 0.8)$ and $X_2 \sim \text{CUB}(m_2 = 5, \pi_2 = 0.7, \xi_2 = 0.3)$; we can find the values of the attainable correlations using the function `corrcheck` in `GenOrd` (Barbiero and Ferrari, 2015a). It returns as minimum and maximum correlations the values $\rho_{\min} = -0.952003$ and $\rho_{\max} = 0.8640543$. We can then proceed and select a desired feasible value of correlation between the two CUB variates, say $\rho = 0.6$. We can then recover the values of ρ_{Ga} (for the Gauss copula), κ (for the Frank copula), and θ (for the Plackett copula), according to the iterative procedure illustrated in the previous section. Setting $\epsilon = 10^{-7}$, we obtain $\rho_{Ga} = 0.6898959$, $\kappa = 5.453455$, and $\theta = 11.30106$. The three joint p.m.f.s, sharing the same level of linear correlation, are reported in Table 1. It is easy to notice the differences among them. For example, the probability $P(X_1 = 2, X_2 = 3)$ takes the values 0.0922, 0.0948, and 0.1008, in the three joint distributions.

References

Andreis F., Ferrari P.A. (2013) On a copula model with CUB margins. *Quaderni di Statistica*, 15, 33-51.

Barbiero A., Ferrari P.A. (2015a) GenOrd: Simulation of Discrete Random Variables with Given Correlation Matrix and Marginal Distributions, *R package version 1.4.0*.

Barbiero A., Ferrari P.A. (2015b) Simulation of correlated Poisson variables. *Applied Stochastic Models in Business and Industry*, 31, 669-680.

Corduas M. (2011) Modelling Correlated Bivariate Ordinal Data with CUB Marginals. *Quaderni di statistica*, 13, 109-119.

Demirtas H. (2006) A method for multivariate ordinal data generation given marginal distributions and correlations. *Journal of Statistical Computation and Simulation*, 76(11), 1017-1025.

Ferrari P.A., Barbiero A. (2012) Simulating ordinal data. *Multivariate Behavioral Research*, 47, 566-589.

McNeil A., Frey R., Embrechts P. (2005) *Quantitative risk management. Concepts, Techniques and Tools*. Princeton Series in Finance, Princeton.

Piccolo D. (2003) On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica*, 5, 85-104.

Table 1. Bivariate distribution with margins $X_1 \sim$ ***CUB***$(m_1 = 5, \pi_1 = 0.4, \xi_1 = 0.8)$ *and* $X_2 \sim$ ***CUB***$(m_2 = 5, \pi_2 = 0.7, \xi_2 = 0.3)$ *and* $\rho_{x_1 x_2} = 0.6$*, obtained based on different copulas*

(x_1, x_2)	1	2	3	4	5	total
1	0.0553	0.0711	0.0959	0.0551	0.0065	0.2838
2	0.0088	0.0317	0.0922	0.1178	0.0333	0.2838
3	0.0013	0.0077	0.0377	0.0869	0.0479	0.1814
4	0.0002	0.0020	0.0150	0.0566	0.0565	0.1302
5	0.0000	0.0004	0.0045	0.0319	0.0838	0.1206
total	0.0657	0.1129	0.2452	0.3481	0.2281	1

(a) Gauss copula

(x_1, x_2)	1	2	3	4	5	total
1	0.0498	0.0744	0.1042	0.0483	0.0071	0.2838
2	0.0126	0.0297	0.0948	0.1167	0.0300	0.2838
3	0.0022	0.0060	0.0301	0.0916	0.0515	0.1814
4	0.0007	0.0019	0.0108	0.0548	0.0621	0.1302
5	0.0003	0.0009	0.0053	0.0366	0.0775	0.1206
total	0.0657	0.1129	0.2452	0.3481	0.2281	1

(b) Frank copula

(x_1, x_2)	1	2	3	4	5	total
1	0.0518	0.0775	0.1001	0.0439	0.0105	0.2838
2	0.0093	0.0251	0.1008	0.1221	0.0266	0.2838
3	0.0025	0.0060	0.0276	0.1004	0.0450	0.1814
4	0.0012	0.0026	0.0105	0.0532	0.0627	0.1302
5	0.0008	0.0018	0.0062	0.0285	0.0833	0.1206
total	0.0657	0.1129	0.2452	0.3481	0.2281	1

(c) Plackett copula