INDUCING A TARGET ASSOCIATION BETWEEN ORDINAL VARIABLES BY USING A PARAMETRIC COPULA FAMILY

A. Barbiero
Università degli Studi di Milano
Milan, ITALY
e-mail: alessandro.barbiero@unimi.it

Abstract

The need for building and generating statistically dependent random variables arises in various fields of study where simulation has proven to be a useful tool. In this work, we present an approach for constructing ordinal variables with arbitrary marginal distributions and association, expressed in terms of either Goodman and Kruskal's gamma or Pearson's linear correlation.
Keywords: data science, ordinal variables, copula

1 Introduction

The need for building and generating statistically dependent random variables arises in various fields of study where simulation has proven to be a useful tool. The ability to simulate data resembling observed data is fundamental to compare and investigate the behaviour of statistical procedures when analytical results are not derivable or are cumbersome to derive.

Many datasets, especially those arising in the social sciences, often contain ordinal variables. Sometimes they are genuine ordered assessments (judgements, preferences, degree of liking, etc.) whereas in other circumstances they are discretized or categorized for convenience (e.g., age of people in classes or education achievement). There are several statistical models and techniques that can be employed for handling multivariate ordinal data without trying to quantify their ordered categories: [1] gives a thorough treatment. Among them, correlation models and association models both study departures from independence in contingency tables and involve the assignment of scores to the categories of the row and column variables in order to maximize the relevant measure of relationship (the correlation coefficient in the correlation models or the measure of intrinsic association in association models [5]). Alternatively, one can code the ordered categories as integers numbers $(1, 2 \ldots, m)$: This amounts to assuming that the categories are evenly spaced.

In this work, we present an approach for constructing ordinal variables with arbitrary marginal distributions and association, expressed in terms of either Goodman and Kruskal's gamma or Pearson's linear correlation. Similar proposals have been already suggested by [7], when dealing with ordinal variables and Goodman and Kruskal's gamma, and by [2, 8, 4] for ordinal (and count) variables and Pearson's correlation.

2 Statement of the problem

We consider two ordinal random variables (rvs), X and Y, with h and k ordered categories, respectively, with marginal distributions $p_{i\cdot} = P(X = x_i), i = 1, \ldots, h$, and $p_{\cdot j} = P(Y = x_j), j = 1, \ldots, k$. We want to determine *some* joint probability distribution $p_{ij} = P(X = x_i, Y = y_j)$, $i = 1, \ldots, h$, $j = 1, \ldots, k$, such that its margins are actually $p_{i\cdot}$ and $p_{\cdot j}$, and with an assigned level of association.

Being X and Y ordinal variables, the association can be naturally expressed through the Goodman and Kruskal's gamma [6]. Considering two independent realizations (X_s, Y_s) and (X_t, Y_t) of (X, Y), Goodman and Kruskal's gamma is defined as

$$\gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d},$$

where Π_c is the probability of concordance:

$$\Pi_c = Pr\left\{X_s < X_t \text{ and } Y_s < Y_t\right\} + Pr\left\{X_s > X_t \text{ and } Y_s > Y_t\right\}$$

and Π_d the probability of discordance:

$$\Pi_d = Pr\left\{X_s < X_t \text{ and } Y_s > Y_t\right\} + Pr\left\{X_s > X_t \text{ and } Y_s < Y_t\right\}.$$

Π_c and Π_d can be expressed in terms of the joint probabilities p_{ij}. γ take values in the $[-1, +1]$ interval; in particular, the values -1, 0, and $+1$ are attained when $\Pi_c = 0$, $\Pi_c = \Pi_d$, $\Pi_d = 0$, respectively.

If we treat X and Y as point-scale discrete variables, by assigning the first h and k positive integers, respectively, to their ordered categories, then we can use Pearson's correlation coefficient as a measure of association:

$$\rho = (\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y))(\mathrm{Var}(X)\mathrm{Var}(Y))^{-1/2}$$

with $\mu_X = \mathbb{E}(X) = \sum_{i=1}^{h} ip_{i\cdot}$, $\mathrm{Var}(X) = \sum_{i=1}^{h}(i - \mu_X)^2 p_{i\cdot}$ (analogous results hold for Y), and $\mathbb{E}(XY) = \sum_{i=1}^{h}\sum_{j=1}^{k} ijp_{ij}$. Like γ, also Pearson's correlation takes values in the $[-1, +1]$ interval; however, given two marginal distributions and a value $\rho \in [-1, +1]$, it is not always possible to construct a joint distribution with those assigned margins, whose correlation is equal to the assigned ρ [9]. In more detail, the attainable correlations form a closed interval $[\rho_{\min}, \rho_{\max}]$ with $\rho_{\min} < 0 < \rho_{\max}$. The minimum correlation $\rho = \rho_{\min}$ is attained if and only if X and Y are countermonotonic; the maximum correlation $\rho = \rho_{\max}$ is attained if and only if X and Y are comonotonic. Moreover, $\rho_{\min} = -1$ 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. Given the two margins, a correlation ρ is said "feasible" if it falls within $[\rho_{\min}, \rho_{\max}]$.

3 A solution to the problem employing copulas

Finding a joint probability distribution with assigned margins and a desired (feasible) value of association is equivalent to solving a system in $h \times k$ unknowns, the p_{ij},

belonging to the standard simplex, subject to $h + k - 1$ constraints corresponding to the assigned margins and one further constraint dictated by the desired association. This system, when h or k is greater than 2, has infinite solutions, which can be recovered more easily when using Pearson's correlation as a measure of association, being it a linear function in the p_{ij}.

Here we propose an approach to identify just one solution, i.e., one joint distribution. This procedure relies on one-parameter bivariate copulas, which allow to split the original problem into two sequential steps: first, identifying a class of joint distributions respecting the assigned margins; then, within this class, finding the joint distribution matching the desired level of association.

3.1 Selecting a class of joint distributions having the pre-specified margins

As for the first step, if F_1 and F_2 are the distribution functions of two rvs X and Y, and $C(u, v; \theta)$ is a bivariate parametric copula family, characterized by some scalar parameter θ, the function

$$F(x, y) = C(F_1(x), F_2(y); \theta), \quad x, y \in \mathbb{R}, \tag{1}$$

defines a valid joint distribution function, whose margins are exactly F_1 and F_2. This result keeps holding if X and Y are discrete; in this case, the joint probabilities can be derived from (1) as:

$$p_{ij} = F(i, j) - F(i - 1, j) - F(i, j - 1) + F(i - 1, j - 1),$$

for $i = 1, \ldots, h; j = 1, \ldots, k$. In order to induce any feasible value of association between the two discrete margins, we have further to impose that the copula $C(u, v; \theta)$ is able to encompass the entire range of dependence, from perfect negative dependence to perfect positive dependence.

3.2 Inducing the desired value of association

As for the second step, the association between X and Y now depends only on the copula parameter θ; this relationship may be written in an analytical or numerical form, say $\gamma = f(\theta)$, or $\rho = g(\theta)$. Since the function f (or g) is not usually analytically invertible, inducing a desired feasible value of association, by setting an appropriate value of θ, is a task that can be generally done only numerically, by finding the (unique) root of the equation $f(\theta) - \gamma = 0$ (or $g(\theta) - \rho = 0$). If γ (or ρ) is a monotone increasing function of the copula parameter, and this is often the case (e.g., for the Gauss, Frank, and Plackett copulas), one can implement some iterative procedure that is more efficient than the standard bisection method. For discrete random variables, several proposals have been suggested for matching a desired value of Pearson's correlation, see [2, 8, 4].

Simulating from the selected joint distribution is straightforward, by resorting to preliminary simulation of copulas or more easily to a direct inversion algorithm [3, 7].

References

[1] Agresti A. (2010). *Analysis of ordinal categorical data*. John Wiley & Sons, New York.

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

[3] Devroye L. (1986). *Non-Uniform Random Variate Generation*. Springer, New York.

[4] Ferrari P.A., Barbiero A. (2012). Simulating ordinal data. *Multivariate Behavioral Research*. Vol. **47(4)**, pp. 566-589.

[5] Faust K., Wasserman S. (1993). Correlation and Association Models for Studying Measurements on Ordinal Relations. *Sociological Methodology*. Vol. **23**, pp. 177-215.

[6] Goodman L.A., Kruskal W.H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*. Vol. **49**, pp. 732-764.

[7] Lee A.J. (1997). Some methods for generating correlated categorical variates. *Computational Statistics and Data Analysis*. Vol. **26**, pp. 133-148.

[8] Madsen L., Dalthorp D. (2007). Simulating correlated count data. *Environmental and Ecological Statistics*. Vol. **14(2)**, pp. 129-148.

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