# A Generalization of the Polychoric Correlation Coefficient

Annarita Roscino and Alessio Pollice

Dipartimento di Scienze Statistiche,
Università degli Studi di Bari, Italy
{aroscino;apollice}@dss.uniba.it

**Abstract.** The polychoric correlation coefficient is a measure of association between two ordinal variables. It is based on the assumption that two latent bivariate normally distributed random variables generate couples of ordinal scores. Categories of the two ordinal variables correspond to intervals of the corresponding continuous variables. Thus, measuring the association between ordinal variables means estimating the product moment correlation between the underlying normal variables (Olsonn, 1979). When the hypothesis of latent bivariate normality is empirically or theoretically implausible, other distributional assumptions can be made. In this paper a new and more flexible polychoric correlation coefficient is proposed assuming that the underlying variables are skew-normally distributed (Roscino, 2005). The skew normal (Azzalini and Dalla Valle, 1996) is a family of distributions which includes the normal distribution as a special case, but with an extra parameter to regulate the skewness. As for the original polychoric correlation coefficient, the new coefficient was estimated by the maximization of the log-likelihood function with respect to the thresholds of the continuous variables, the skewness and the correlation parameters. The new coefficient was then tested on samples from simulated populations differing in the number of ordinal categories and the distribution of the underlying variables. The results were compared with those of the original polychoric correlation coefficient.

## 1 Introduction

Data in the social and medical sciences are often based on ordinal measurements and represented in contingency tables. A first approach to the analysis of this kind of variables is to measure their association in order to know if some relationship exists and to quantify its strength. To achieve this purpose, it is possible either to estimate the concordance between the scores of each ordinal variable or to assume that those variables derive from the categorization of some continuous variables.
The first type of measure includes Kendall's $\tau$, Somers' $e$, Goodman and Kruskal's $\gamma$ and many others more (Agresti, 2004). They estimate the association between ordinal variables comparing the frequencies of each category without any distributional assumption. The polychoric correlation coefficient, instead, is based on the assumption that the ordinal variables derive from partitioning the range of some continuous normally distributed variables into

categories. Consequently, it does not compare two sets of scores, but rather estimates the correlation between two unobserved continuous variables underlying the two ordinal variables assuming a bivariate normal distribution with means zero and variances one.

Some studies have been carried out in order to compare the most important measures of association. Joreskog and Sorbon (1988) performed an experiment based on a bivariate normal distribution for the underlying variables and showed that, under this condition, the polychoric correlation coefficient is always closer to the real correlation than all measures evaluated in the same study. Moreover, the matrix of the polychoric correlation coefficients is largely used to replace the covariance matrix in order to estimate the parameters of structural equation models when the observed variables are ordinal. On the other hand, experience with empirical data (Aish and Joreskog, 1990) shows that the assumption of underlying bivariate normality seldom holds. It is also believed that this assumption is too strong for most ordinal variables used in the social sciences (Quiroga, 1991). Therefore, there is a need to find a shape of the underlying variables more plausibly compatible with the real data.

Many studies performed in order to analyse the distributions of underlying variables showed that asymmetric distributions are very frequent. Muthen (1984) proved that the distributions of underlying variables can be highly skewed, causing lack of convergence and/or negative standard errors when estimating structural equation model parameters. Moreover, Muthen and Kaplan (1985) noticed that the presence of asymmetric latent distributions can bias the results of chi square tests used to assess the goodness of fit of structural equation models. The former studies suggest a need to find a distribution that takes into account the potential asymmetry of the underlying variables. In this paper a new polychoric correlation coefficient is proposed, based on the hypothesis that underlying variables have a bivariate skew normal distribution (Roscino, 2005). The bivariate skew normal distribution (Azzalini and Dalla Valle, 1996) belongs to a family of distributions which includes the normal distribution as a special case, but with two extra parameters to regulate the skewness. As for the polychoric correlation coefficient, maximum likelihood was used in order to estimate the new polychoric correlation coefficient under the assumption of underlying skew normally distributed variables. A simulation study was then carried out in order to compare the performance of the new coefficient with that of the original polychoric correlation coefficient. In the first section of this paper, the generalised polychoric correlation coefficient is defined and estimated. In the second section the simulation study is presented and in the third section the results of the simulation study are shown and the efficacy of the new polychoric correlation coefficient is discussed.

## 2 An Extension of the Polychoric Correlation Coefficient

The polychoric correlation coefficient (Olsonn, 1979) is based on the assumption that underlying a pair of ordinal variables there is a couple of continuous latent variables which have a bivariate normal distribution. Ordinal variables $X$ and $Y$, with $I$ and $J$ categories each, are thus assumed to be related to underlying continuous variables $Z_1$ and $Z_2$ by

$$\begin{cases} X = i \text{ if } a_{i-1} \leq Z_1 \leq a_i, \; i = 1, 2, \ldots, I \\ Y = j \text{ if } b_{i-1} \leq Z_2 \leq b_i, \; j = 1, 2, \ldots, J \end{cases} \tag{1}$$

where $Z_1$ and $Z_2$ have a bivariate normal distribution with correlation coefficient $\rho$ and $a_i$ and $b_j$ are referred to as thresholds. Measuring the polychoric correlation means estimating the product moment correlation $\rho$ between underlying normal variables. This correlation is estimated by the maximum likelihood method, assuming a multinomial distribution of the cell frequencies in the contingency table. If $n_{ij}$ is the number of observations in cell $(i, j)$, and $K$ is a constant, the likelihood of the sample is given by

$$L = K \prod_{i=1}^{I} \prod_{j=1}^{J} P_{ij}^{n_{ij}}, \tag{2}$$

where

$$P_{ij} = \Phi_2(a_i, b_j) - \Phi_2(a_{i-1}, b_j) - \Phi_2(a_i, b_{j-1}) + \Phi_2(a_{i-1}, b_{j-1}) \tag{3}$$

and $\Phi_2$ is the bivariate normal distribution function with unknown correlation coefficient $\rho$. The estimator of the polychoric correlation coefficient between variables $X$ and $Y$ corresponds to the value of $\rho$ which maximizes equation 2, where the choice of the number of categories $I$ and $J$ has a crucial influence on the dimensionality of the likelihood function.

A problem with the polychoric correlation coefficient concerns the robustness of the method to departures from symmetric distributional assumptions. Quiroga (1991) carried out a Monte Carlo study in order to analyze the effects of the departure from the normal assumption on the estimation of the polychoric correlation coefficient. The author simulated samples from underlying distributions affected by asymmetry and showed that the polychoric correlation underestimates the association between ordinal variables, particularly when the sample size is large and the categories are few. Such results reveal that there could be an advantage in considering a latent distribution more compatible with real data. As discussed in the Introduction, the underlying variables are often asymmetric (Muthen, 1984), therefore the bivariate skew normal distribution was chosen.

A random variable $Z = (Z_1, Z_2)$ is said to be distributed according to a bivariate skew normal $SN(\alpha_1, \alpha_2, \omega)$ if its density function is given by

$$g(z_1, z_2) = 2\phi(z_1, z_2; \omega)\Phi(\alpha_1 z_1 + \alpha_2 z_2), \tag{4}$$

where $\phi(\cdot;\omega)$ is the bivariate normal distribution with null mean, unit variance and correlation $\omega$ and $\Phi(\cdot)$ is the univariate standard normal distribution function. Skewness $\alpha_1$ and $\alpha_2$ can vary in $(-\infty, \infty)$ and imply bivariate normality when they are both null. The correlation coefficient associated to the bivariate skew normal distribution is given by:

$$\rho_{SN} = \frac{\omega - 2\pi^{-1}\delta_1\delta_2}{\{(1 - 2\pi^{-1}\delta_1^2)(1 - 2\pi^{-1}\delta_2^2)\}^{1/2}}, \tag{5}$$

where $\delta_1$ and $\delta_2$ are linked to $\alpha_1$ $\alpha_2$ and $\omega$ by expressions:

$$\begin{aligned} \alpha_1 &= \frac{\delta_1 - \delta_2\omega}{\{(1-\omega^2)(1-\omega^2-\delta_1^2-\delta_2^2+2\delta_1\delta_2\omega)\}^{1/2}} \\ \alpha_2 &= \frac{\delta_2 - \delta_1\omega}{\{(1-\omega^2)(1-\omega^2-\delta_1^2-\delta_2^2+2\delta_1\delta_2\omega)\}^{1/2}}, \end{aligned} \tag{6}$$

with $\delta_1$ and $\delta_2$ in $[-1, 1]$.
Under the new assumption, the joint distribution of the underlying variables $Z_1$ and $Z_2$ is bivariate skew normal, as given in (4). Thus the product moment correlation of $Z_1$ and $Z_2$, $\rho_{SN}$ estimates the polychoric correlation coefficient between $X$ and $Y$.
As for the original polychoric correlation coefficient, the new coefficient is estimated by maximization of the log-likelihood function $L$ (see 2) with respect to the thresholds, the skewness and the correlation parameters, where the new expression of the probability $P_{ij}$ in the likelihood of the sample is equal to:

$$P_{ij} = P[X = i \wedge Y = j] = 2\int_{a_{i-1}}^{a_i}\int_{b_{j-1}}^{b_j} \phi(z_1, z_2, \Omega)\Phi(\alpha_1 z_1 + \alpha_2 z_2)dz_1 dz_2. \tag{7}$$

In order to work with standardized parameters, a different parametrization of the skew normal distribution was considered (Azzalini and Dalla Valle, 1996). The correlation parameter $\omega$ was replaced by $\psi$, where

$$\psi = (\omega - \delta_1\delta_2)[(1 - \delta_1)(1 - \delta_2)]^{-1/2} \tag{8}$$

and the skewness parameters $\alpha_1$and $\alpha_2$ were replaced by $\delta_1$ and $\delta_2$ (see 6).
The function `sn.polychor` (Roscino, 2005) was written in R to perform the maximization of the log-likelihood function using a numerical optimization method, according to Nelder and Mead (1965). This method works reasonably well for non-differentiable functions as it uses only function values and does not require to evaluate the gradient of the log-likelihood.
The function `sn.polychor` first computes the maximum likelihood estimates of $\psi$, $\delta_1$ and $\delta_2$ and their standard errors. Then, after replacing $\psi$, $\delta_1$ and $\delta_2$ with their estimated values in 8, it calculates $\widehat{\omega}$ and $\widehat{\rho}_{SN}$ (see 5).
The function `sn.polychor` is available on request by emailing the first author.

## 3 The Simulation Study

The new polychoric correlation coefficient $\widehat{\rho}_{SN}$ was calculated for 360 samples from simulated bivariate populations differing in the number of ordinal categories, correlation and skewness parameters of the underlying distributions, as shown in Table 1. One sample was considered for each combination of the parameters $\psi$, $\delta_1$, $\delta_2$, $I$, $J$ and $n$. The sampling distribution of $\widehat{\rho}_{SN}$ and $\widehat{\rho}$ was analysed only for the combination ($\psi = 0.5, \delta_1 = 0.7, \delta_2 = -0.7, I = 3, J = 3, n = 400$) where 100 samples were extracted and means and standard errors of $\widehat{\rho}_{SN}$, $\widehat{\rho}$, $\widehat{\delta_1}$, $\widehat{\delta_2}$ were computed. The analysis of the sampling distributions associated with the remaining combinations of parameters is currently being undertaken and the results will be presented in the near future.
The `R` library `MASS` was used to produce samples from bivariate normal distributions, while a new function called `sn.simul` (Roscino, 2005) was implemented in order to generate samples from bivariate skew normal distributions. The generated samples of the underlying variables $(Z_1, Z_2)$ were grouped according to intervals and each interval was associated with a category of the corresponding ordinal variable. The values of $\psi$, $\delta_1$ and $\delta_2$ were chosen to

| $\psi$ | 0.3 | 0.5 | 0.8 | | | |
|---|---|---|---|---|---|---|
| $n$ | 250 | 400 | 600 | 800 | | |
| $(I,\ J)$ | (2,2) | (2,3) | (3,3) | (3,5) | (5,5) | |
| $(\delta_1,\ \delta_2)$ | (0,0) | (0,0.7) | (0.7,0.7) | (0.7,-0.7) | (0.4,0.4) | (0.4,-0.4) |

**Table 1.** Parameters of the simulated distributions

include as many different shapes of the distributions of underlying variables as possible. In particular, when $\delta_1$ and $\delta_2$ are equal to zero, the underlying variables have bivariate normal distribution with correlation coefficient equal to $\psi$. For all the other cases, the simulated distributions are bivariate skew normals and the associated values of the polychoric correlation coefficient can be found in Table 2.
The `R` functions `sn.polychor` and `polychor` (Johnson, 2004) were used to compute $\widehat{\rho}_{SN}$ and $\widehat{\rho}$ respectively. While the output of `polychor` consists of the estimators of $\rho_{SN}$, $a_i$, $b_j$ (for $i = 1, ..., I$ and $j = 1, ..., J$) with their standard errors, the function `sn.polychor` estimates the additional parameters $\psi$, $\delta_1$ and $\delta_2$ and their standard errors, together with $\rho_{SN}$, $a_i$, $b_j$ (for $i = 1, ..., I$ and $j = 1, ..., J$) and their standard errors.
The values of $\widehat{\rho}_{SN}$ and $\widehat{\rho}$ were compared with the true value of the polychoric correlation coefficient for each of the simulated samples. The performance of both estimators with respect to the value of the polychoric correlation coefficient in the underlying population (as the absolute value of the difference)was

| | $\psi$ | | |
|---|---|---|---|
| $(\delta_1, \delta_2)$ | 0.3 | 0.5 | 0.8 |
| (0, 0) | 0.3 | 0.5 | 0.8 |
| (0, 0.7) | 0.2582 | 0.4304 | 0.6887 |
| (0.7, 0.7) | 0.4811 | 0.6293 | 0.8517 |
| (0.7, -0.7) | -0.0364 | 0.1118 | 0.3341 |
| (0.4, 0.4) | 0.3453 | 0.5323 | 0.8129 |
| (0.4, -0.4) | 0.2158 | 0.4028 | 0.6834 |

**Table 2.** Values of $\rho_{SN}$

evaluated only for the combination ($\psi = 0.5, \delta_1 = 0.7, \delta_2 = -0.7, I = 3, J = 3, n = 400$).

## 4 Some Results

In this section some results of the simulations are summarized. It is clear that a complete evaluation of the performance of the estimator would need a more extensive simulation study which is currently being undertaken.
The simulations involved one sample for each combination of parameters and showed that $\widehat{\rho}_{SN}$ is always closer to the real correlation than $\hat{\rho}$ when:

1. The number of categories of ordinal variables is small, ie. less than or equal to 3 (See Figure 1a, where the solid line represents the real polychoric correlation coefficient while the dashed and the dotted lines are respectively $\widehat{\rho}$ and $\widehat{\rho}_{SN}$) or
2. The sample size is large - 400 units or above, or
3. The skewness parameters are discordant, that is when they have opposite signs (See Figure 1b).

Furthermore, under these conditions the estimators of the skewness parameters are always very close to their values in the population.
These results are confirmed by the analysis of the sampling distribution of $\widehat{\rho}_{SN}$ for the combination of parameters ($\psi = 0.5, \delta_1 = 0.7, \delta_2 = -0.7, I = 3, J = 3, n = 400$). The mean and standard deviation of $\widehat{\rho}_{SN}$ were equal to 0.1173 and 0.0072 respectively while the mean and standard deviation of $\widehat{\rho}$ were 0.1003 and 0.0651. The mean of $\widehat{\rho}_{SN}$ is closer to $\rho$ than the mean of $\widehat{\rho}$ (see Table 2) and the standard deviation is lower than the standard deviation of $\widehat{\rho}$ by a factor of almost ten.
On the other side, the polychoric correlation coefficient is closer to the real correlation when:

1. The sample size is small, or
2. The number of categories is large.

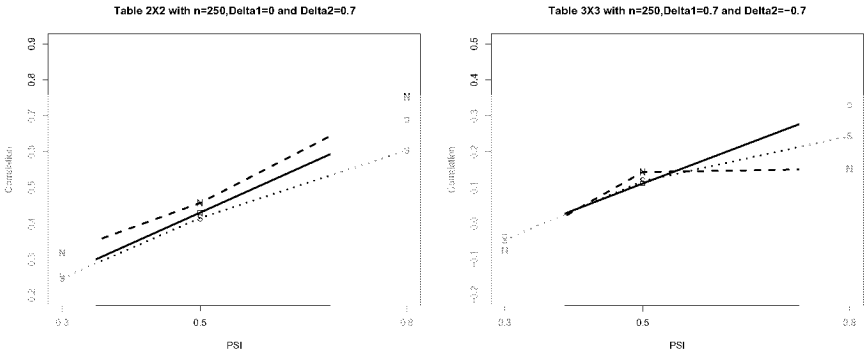


**Fig. 1.** Value of the estimated correlation coefficients for each value of $\psi$.

For $(I, J) = (5, 5)$, the results showed a high degree of variability and were therefore of limited use. This case is currently being studied to improve the quality of the results.
When the sample size is small, the poor results of the generalised polychoric correlation coefficient could be determined by an irregularity in the likelihood function of the bivariate skew normal distribution. Azzalini and Capitanio (1999) showed that for small sample sizes, the maximum likelihood estimators of the parameters of a skew normal distribution can overestimate their real values. This is due to the analytical expression of the likelihood function and cannot be modified using a different parametrisation.
The conclusions of the paper of Joreskog and Sorbon (1988) hold when the number of categories of the ordinal variables is large. The authors compared six measures of ordinal association and found that the polychoric correlation coefficient is more robust to departures from normality in the presence of ordinal variables with a large number of categories.

## 5 Conclusions

In this paper we propose a new polychoric correlation coefficient based on the assumption that the underlying continuous variables are skew normally distributed. By definition, $\rho_{SN}$ is equal to $\rho$ when the underlying variables are normally distributed, but it is more flexible than $\rho$ as it takes into account the potential skewness of the underlying variables.
An `R` function was written in order to compute $\widehat{\rho}_{SN}$ and and 360 samples were generated with the aim of comparing $\widehat{\rho}_{SN}$ and $\widehat{\rho}$.
The examples presented in the simulation study indicates that $\widehat{\rho}_{SN}$ is more appropriate than $\widehat{\rho}$ when the sample size is large or the number of categories of ordinal variables is small or the skewness parameters have opposite signs.

On the other hand, the simulation study has shown that some further developments are needed in particular when the sample sizes are small or the number of ordinal categories is large.

## References

AGRESTI, A. (2004): Categorical Data Analysis, 2nd ed. John Wiley & Sons. New York

AISH, A. M. and JORESKOG, K. G. (1990): A panel model for political efficacy and responsiveness: An application of LISREL 7 with weighted least squares. *Quality and Quantity, 24, 405–426.*

AZZALINI, A. and CAPITANIO, A. (1999): Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society, Series B, 61, 579–602.*

AZZALINI, A. and DALLA VALLE, A. (1996): A multivariate skew-normal distribution. *Biometrika, 83, 715–726.*

JOHNSON, T.R. (2004): *Polychor function for* R. http://www.webpages.uidaho.edu/ trjohns/polychor.R.

JORESKOG, K. G. and SORBON, D. (1988): A *program for multivariate data screening and data summarization.* A *processor for* LISREL. Scientific Software Inc. Mooresville, Indiana.

MUTHEN, B. (1984): A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika, 49, 115–132.*

MUTHEN, B. and KAPLAN, D. (1985): A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical Statistical Psychology, 38, 171–189.*

NELDER, J. A. and MEAD, R. (1965): A simplex method for function minimization. *Computer Journal, 7, 308–317.*

OLSONN, U. (1979): Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44, 443–460.*

QUIROGA, A. M. (1991): Studies of the Polychoric Correlation and other Correlation Measures for Ordinal Data. *Uppsala University, Department of Statistics. PhD Thesis.*

ROSCINO, A. (2005): Una generalizzazione del coefficiente di correlazione policorica. *Università degli Studi di Chieti - Pescara. Tesi di Dottorato di Ricerca.*