Data Type Detection for Choosing an Appropriate Correlation Coefficient in the Bivariate Case

Anastasiia Yu. Timofeeva
Novosibirsk State Technical University, Russia
a.timofeeva@corp.nstu.ru

Abstract

The data scientists usually define a data type based on a nature of variables and select an appropriate correlation measure. However, this is not convenient and very time-consuming in data intensive domains. I propose to detect the types of variables and choose the appropriate correlation coefficient in order to automate the statistical procedure of correlation estimating from mixed data. This should lead to a reduction of time spent on correlation analysis and to increase the accuracy of estimation of correlation coefficients. The continuity index is used to detect whether a variable is continuous or ordered categorical. Based on simulation study I have estimated the cutoff level for the continuity index to choose the Pearson correlation, the polychoric, or the polyserial correlation coefficient.

1 Introduction

As a measure of the linear dependence the Pearson correlation coefficient is most commonly used due to computationally ease and good statistical properties of the estimate under standard assumption of normality. Applications of Pearson correlation coefficient are limited to quantitative data in a continuous scale of measurement. In practice, the analyst has to deal with a set of data measured in different scales. Consequently, mixed data is subjected to statistical procedures such as calculating the correlation coefficient. Therefore it is necessary to choose an appropriate correlation measure which allows to handle such data.

Types of input variables can be various, e.g. binary, integer, ordered categorical (e.g. item response), and continuous. For each combination of measurement scales a certain bivariate correlation coefficient is used:

- tetrachoric correlation between two binary variables,
- polychoric correlation between two ordered categorical variables,
- biserial correlation between a continuous variable and a dichotomous variable,
- polyserial correlation between a continuous variable and a ordered categorical variable.

The use of various correlation coefficients for the same set of data may lead to significantly different conclusions. The authors of [Hol2010] have shown that when construct validity is analysed according to ordinal data obtained

In: S. Hölldobler, A. Malikov, C. Wernhard (eds.): *YSIP2 – Proceedings of the Second Young Scientist's International Workshop on Trends in Information Processing, Dombai, Russian Federation, May 16–20, 2017*, published at http://ceur-ws.org.

from Likert scales, the factor results show a better fit to the theoretical model when the factorization is carried out using the polychoric rather than the Pearson correlation matrix.

In carrying out a correlation the scientist himself usually defines the data type and selects an appropriate correlation measure. However, this is not convenient and very time-consuming in data intensive domains. I propose to detect the types of variables and choose the appropriate correlation coefficient in order to automate the statistical procedure of correlation estimating from mixed data. This should lead to a reduction of time spent on correlation analysis and to increase the accuracy of estimation of correlation coefficients.

It is further assumed that the generating data process is based on multivariate normal distribution. The continuous data is discretized, i.e. converted to ordered categorical variables. The number of categories provides information to automatically detect whether a variable is discrete or continuous.

2 Measures of Bivariate Association

In some cases it is difficult to measure precisely the value a variable on a quantitative scale, but very easy to place observation into ordered categories [Dra1988]. So in addition to the well-known measure of correlation between continuous variables Karl Pearson proposed polychoric and polyserial correlations [Pea1913]. Let us take a closer look at their definition and methods of estimation.

2.1 Pearson Correlation Coefficient

Pearson correlation coefficient are widely used both in the practice, and in the sciences. It measures the linear dependence between two variables. The variables should be measured on an interval scale. If you analyse the ordinal data, a simple and naive plug-in strategy would be to use the discrete values as if they were continuous and to calculate Pearson correlation coefficient. However, this approach is inferior to other methods for analyzing discrete data, such as using the polychoric correlations [Hol2010], [Kol2004].

2.2 Polychoric Correlation Coefficient

Briefly, let us suppose that x_1 and x_2 are two ordinal items with n_1 and n_2 categories. It can be assumed that underlying these items are variables ξ_1 and ξ_2. Their joint distribution is a bivariate standard normal distribution with a correlation ρ between random variables ξ_1 and ξ_2.

The discrete random variables x_1 and x_2 are obtained by grouping, i.e. the partition of the range of values of random variables ξ_1 and ξ_2 into intervals. It is assumed that x_1 takes values from 1 to n_1, x_2 — from 1 to n_2. The bounds of these intervals $\alpha_{i1},\ i = 0, 1, \ldots, n_1,\ \alpha_{j2},\ j = 0, 1, \ldots, n_2$ are called discretizing thresholds. They are unknown and $\alpha_{0k} = -\infty,\ \alpha_{n_k k} = +\infty$. Then the relation between between x_k and ξ_k is given by the expression

$$x_k = i \text{ if } \alpha_{(i-1)k} < \xi_k < \alpha_{ik}, k = 1, 2. \tag{1}$$

The sample distribution of x_1 and x_2 is given by the contingency table. It contains the relative frequencies d_{ij}, i.e. the number of cases in category i of item 1 and in category j of item 2 to the sample size.

The theoretical probability $p_{ij} = P(x_1 = i,\ x_2 = j)$ corresponding to d_{ij} is defined as

$$\begin{gathered} p_{ij} = P(x_1 = i,\ x_2 = j) = P(\alpha_{(i-1)1} < \xi_1 < \alpha_{i1},\ \alpha_{(j-1)2} < \xi_2 < \alpha_{j2}) = \\ = \Phi_2(\alpha_{i1}, \alpha_{j2}, \rho) - \Phi_2(\alpha_{(i-1)1}, \alpha_{j2}, \rho) - \Phi_2(\alpha_{i1}, \alpha_{(j-1)2}, \rho) + \Phi_2(\alpha_{(i-1)1}, \alpha_{(j-1)2}, \rho) \end{gathered} \tag{2}$$

where $\Phi_2(z_1, z_2, \rho)$ is bivariate standard normal distribution function with correlation ρ between random variables ξ_1 and ξ_2.

The problem is to estimate the unknown parameters of the bivariate distribution of random variables x_1 and x_2 based on observed values d_{ij}. The estimate of ρ of this model is called the polychoric correlation coefficient.

In this study I consider a two-step approach [Ols1979]. The first step is to find estimates for thresholds α_{ik} as quantiles of corresponding marginal distributions:

$$\hat{\alpha}_{i1} = \Phi^{-1}\left(\sum_{l=1}^{i}\sum_{j=1}^{n_2} d_{lj}\right),\ i = 1, \ldots, n_1 - 1,\ \hat{\alpha}_{j2} = \Phi^{-1}\left(\sum_{l=1}^{j}\sum_{i=1}^{n_1} d_{il}\right),\ j = 1, \ldots, n_2 - 1$$

where $\Phi(\cdot)$ is the standard normal distribution function, $\Phi^{-1}(\cdot)$ is the quantile function.

In the second step the estimates of thresholds are substituted into (2) and theoretical probabilities are considered as a function of unknown parameter ρ. For its estimation the maximum likelihood method is used. For the joint discrete distribution of random variables x_1 and x_2 under the assumption of independence of observations the average log-likelihood of the sample [Ols1979] is

$$\hat{\ell} = \sum_{i,j \in U} d_{ij} \ln p_{ij} \tag{3}$$

where a finite set $U = \{i, j : d_{ij} \neq 0 \& p_{ij} \neq 0\}$ is to avoid infinite value of the function $\hat{\ell}$ by $d_{ij} = p_{ij} = 0$. In (3) each d_{ij} is a fixed value for the given sample.

Both Pearson correlation coefficient, and polychoric correlation coefficient have similar properties. The correlation coefficient has a value between $+1$ and -1 inclusive. The coefficient is symmetric. This means between x_1 and x_2 is the same as the correlation between x_2 and x_1.

2.3 Polyserial Correlation Coefficient

If the latent correlation between a continuous variable and a ordered categorical variable is assumed, then the polyserial correlation coefficient is the most appropriate correlation measure. In this case one variable x_1 with underlying standard normal ξ_1 is assumed to be discrete. It is expressed by (1) with $k = 1$. Another observed variable x_2 is considered to be continuous and standard normally distributed. In practice the variable defined as continuous should be standardized, so that its mean becomes zero and its standard deviation becomes one.

According to [Dra1988] the log-likelihood function for the joint distribution of random vector (x_1, x_2) from a sample of n observations (x_{i1}, x_{i2}) is

$$\log L = \sum_{i=1}^{n} \log \phi(x_{i2}) + \log P(x_1 = x_{i1} | x_2 = x_{i2}) \tag{4}$$

where $\phi(\cdot)$ is the standard normal density function.

The conditional distribution of ξ_1 given $x_2 = x_{i2}$ is normal with mean ρx_{i2} and variance $1 - \rho^2$. Then if $x_{i1} = j$ with categories $j = 1, \ldots, n_1$, the resulting conditional probability is

$$P(x_1 = j | x_2 = x_{i2}) = \Phi\left(\frac{\alpha_{(j-1)1} - \rho x_{i2}}{\sqrt{1-\rho^2}}\right) - \Phi\left(\frac{\alpha_{j1} - \rho x_{i2}}{\sqrt{1-\rho^2}}\right). \tag{5}$$

A two-step approach to estimation the polyserial correlation assumes that discretizing thresholds in (5) can be computed by formula

$$\hat{\alpha}_{j1} = \Phi^{-1}\left(\sum_{l=1}^{j} d_l\right),\ j = 1, \ldots, n_1 - 1$$

where d_l is the relative frequency, i.e. the number of cases in category l of item 1 to the sample size.

As a result of maximization of the log-likelihood function (4) with argument ρ, the polyserial correlation estimate is obtained. It is clear that the polyserial correlation coefficient is not symmetric. For estimating polyserial correlation it is important which of the variables is assumed to be continuous or discrete.

3 Data Type Detection

All actual sample spaces are discrete, and all observable random variables have discrete distributions [Pit1979]. To detect whether a variable is continuous or discrete you need to understand the nature of the data. The variable can be considered as continuous if there is an infinite number of possible values that the variable can take between any two different points in the range. Any measurement of these variables will be discrete. In actual practice, a variable is often treated as continuous when it can take on a sufficiently large number of different values. It sometimes makes sense to treat continuous variable as ordered categorical. This is usually just another kind of binning.

Discrete data can only take particular values. If a variable can take on one of a limited number of possible values referred to as levels, it is a categorical variable. A categorical data type where the variable has natural,

ordered categories is ordinal (ordered categorical) data. The distance between the categories is considered as a unknown. It is generally not correct to consider ordered categorical data as continuous.

In data intensive analysis it is almost impossible to determine the nature of each variable. It is necessary to formulate a simple rule that would allow to detect whether a variable can be considered as continuous or ordered categorical. It seems logical to count the number of unique values of the variable and relate it to sample size. If a number of different values is a sufficiently large, then a variable can be considered as continuous.

Let me introduce the continuity index of k-th variable defined as the ratio the number n_k of categories (unique values) to sample size n:

$$\gamma_k = \frac{n_k}{n}.$$

The problem is to define cutoff at which the discrete variable will be considered as continuous. The author did not find the detailed recommendations on this subject. Documentation to the package 'treeplyr' of statistical environment R gives a default value of cutoff = 0.1 for deciding if numeric data might actually be descrete. The continuity index γ_k should exceed cutoff, or the data will be classified as discrete. However, this cutoff value is not justified. It is therefore necessary to carry out simulation studies to identify the most appropriate coefficient for different values of the continuity indices γ_1 and γ_2.

4 Software Implementation

Both the polyserial, and the polychoric correlations unfortunately are not typically used in the statistical analysis. Nevertheless the functions (or packages) to calculate their are available in many statistical programs such as SPSS, SAS, Stata, R. The first three of these programs are proprietary software, so the authors decided to focus on free software for statistical computing R.

There are two R-packages, which have the functions to calculate polychoric and polyserial correlations, polycor and psych. The functions polyserial{psych} and polychoric{psych} have the drawback that the calculation is not performed if there are more than 8 categories for any item. This is a very substantial limitation. It is not possible to use these functions to simulation study.

The functions polyserial{polycor} and polychor{polycor} do not have explicit restrictions on the dimension of contingency tables. However, polychor works very slow with a large number of categories. The most computationally expensive stage is calculation of the function of two-dimensional normal distribution at each iteration of the optimization algorithm. Here it is performed by internal function binBvn which uses the function pmvnorm{mvtnorm}. The function pmvnorm for a given interval in a p-dimensional space (in our case - a two-dimensional) returns a scalar value, meaning it is not vectorized. For the calculation of probabilities for each combination $\hat{\alpha}_{i1}$, $\hat{\alpha}_{j2}$ the function binBvn uses a nested loop over indices i, j. It is an unfortunate fact and leads to polychor functions work slowdown for a large number of categories, since loops are very slow in interpreted language R.

For this reason, I have implemented the calculation of polychoric correlation coefficient using a standard set of R-packages. To do this, several user-functions have been written.

The function FuncCalcPolychor has inputs: the current value of the correlation coefficient ρ_t, threshold values $\hat{\alpha}_{i1}$, $\hat{\alpha}_{j2}$, frequency table d_{ij}. The algorithm can be divided into two steps.

1. Based on the values of ρ_t, $\hat{\alpha}_{i1}$, $\hat{\alpha}_{j2}$ two-dimensional array is calculated containing the values of the two-dimensional standard normal distribution for all combinations $\hat{\alpha}_{i1}$, $\hat{\alpha}_{j2}$. The last row of the array values $\Phi(\alpha_{j2})$ are added. The last column of the array values $\Phi(\alpha_{i1})$ are added. At their intersection unit is placed. At the beginning of the array one zero row and one zero column are added. On the basis of this array an array of probabilities p_{ij} is constructed according to (2). Negative values of probabilities that may occur due to inaccuracies the calculation, are reset to zero.

2. Based on the values of d_{ij} and obtained in step 1 values p_{ij} calculates and returns the value $\hat{\ell}$.

In Step 1, to calculate the values of the two-dimensional standard normal distribution function you can use pmnorm function. This requires installation of package mnormt. It works by making a suitable call to Fortran-77 routine written by Alan Genz. The function pmnorm is vectorized. It returns a vector of values for input matrix $N \times 2$ where N is the number of points at which probabilities are calculated. In our case $N = n_1 n_2$. To create all combinations a standard function expand.grid is used.

Alternatively (without additional packages), I have implemented a number of user-functions for calculating the two-dimensional matrix of values of the bivariate standard normal distribution function. They are written on the basis of the algorithm proposed in [Mey2013]. I have used the C-code presented in the article [Mey2013] and rewrote it in the language R with the addition of vectorization. In other words, all functions processing scalar values (or vectors) are replaced by functions processing vectors (or matrix), such as ifelse, apply, outer, pmin. Also some errors was corrected, such as division by zero, if any of the values $\hat{\alpha}_{i1}$, $\hat{\alpha}_{j2}$ is zero. Zero is replaced by 10^{-5}.

Finally, the function PolychorEst on the input vectors has sampled values of the observed variables x_1 and x_2. Its work can also be divided into two steps.

1. Filling a table of relative frequencies d_{ij} for given vectors x_1 and x_2. Calculation of vectors $\sum_{l=1}^{i}\sum_{j=1}^{n_2} d_{lj}$, $\sum_{l=1}^{j}\sum_{i=1}^{n_1} d_{il}$. Calculation based on them $\hat{\alpha}_{i1}$, $\hat{\alpha}_{j2}$ using a standard function qnormstats to calculate the quantile of the normal distribution.

2. Optimization of FuncCalcPolychor function with respect to ρ_t for given values $\hat{\alpha}_{i1}$, $\hat{\alpha}_{j2}$, d_{ij} calculated in step 1. A one-parameter optimization was carried out using a basic function optimize{stats} in the interval $\rho \in (-1, 1)$.

In addition, I have implemented the calculation of polyserial correlation coefficient using a standard set of R-packages. The user-function PolyserialEst is quite simple. It standardizes the variable which is assumed to be continuous. Further, it optimizes FuncCalcPolyserial function with respect to ρ_t for given sample values of x_1 and standardized values of x_2. The function FuncCalcPolyserial calculates the log-likelihood function L according to (4). It uses standard functions pnorm, qnorm of calculating distribution function, quantile function for the normal distribution.

5 Simulation Study

For simulation study the following model example was used. The random variable ξ_1 was simulated from a standard normal distribution. The random variable ξ_2 was defined as

$$\xi_2 = \xi_1 + \varepsilon$$

where ε is normally distributed random variable with mean zero and standard deviation σ_ε. The value of σ_ε depended on the value of the correlation coefficient ρ specified by the scheme of the experiment. These values are related by the relationship

$$\sigma_\varepsilon = \sigma_x \sqrt{\frac{1}{\rho^2} - 1}.$$

Further grouping of variables ξ_1 and ξ_2 was carried out. The equidistant intervals were used with boundaries defined by sample quantiles at probabilities $0, \frac{1}{n_k}, \ldots, \frac{n_k-1}{n_k}, 1$. As the value of the variable x_k at the i-th level the sample mean of all the values in i-th interval is taken.

The value of ρ is set to 0.5. The value of n is set to 500. The values of n_k were taken from $5, 10, 25, 50, 100, 250$, and the correlation coefficients were estimated. Results were averaged $1,000$ repetitions. Figure 1 shows the average value $\hat{\rho}$ of Pearson correlation coefficient. Figure 2 shows the average value $\hat{\rho}$ of polyserial correlation coefficient when the first variable is considered as ordered categorial and second variable is assumed to be continuous. Figure 3 shows the average value $\hat{\rho}$ of polyserial correlation coefficient when the second variable is considered as ordered categorial and first variable is assumed to be continuous.

The average values of polychoric correlation are very close to 0.5 for all combinations of the continuity index values. It means that the polychoric correlation coefficient gives an estimate that is the closest to the true value. However, when the continuity index γ_k is high the calculation is very slow. Table 1 shows the user time the running R-functions for calculating one value of polychoric correlation. It is clear that developed user function PolychorEst runs six times faster than the function polychor{polycor} in the case where $\gamma_k \geq 0.05$. However, when $\gamma_2 = 0.05, \gamma_1 = 0.50$ calculations take nearly a minute. Therefore it is recommended to use a simple correlation coefficients (in particular, Pearson correlation) if it provides acceptable accuracy for results.

Figure 1 shows that by fixed value of γ_2 and γ_1 exceeding 0.05 the change in γ_1 values does not significantly decrease a bias in the Pearson correlation coefficient. The values of $\gamma_1 > 0.05$ and $\gamma_2 > 0.05$ lead to a small bias

Table 1: The total user CPU time of the R process, in seconds, $\gamma_2 = 0.05$

Function	$\gamma_1 = 0.01$	$\gamma_1 = 0.02$	$\gamma_1 = 0.05$	$\gamma_1 = 0.10$	$\gamma_1 = 0.20$	$\gamma_1 = 0.50$
polychor{polycor}	0.22	0.78	3.07	15.01	53.75	319.21
PolychorEst	0.17	0.22	0.52	1.95	7.92	52.75

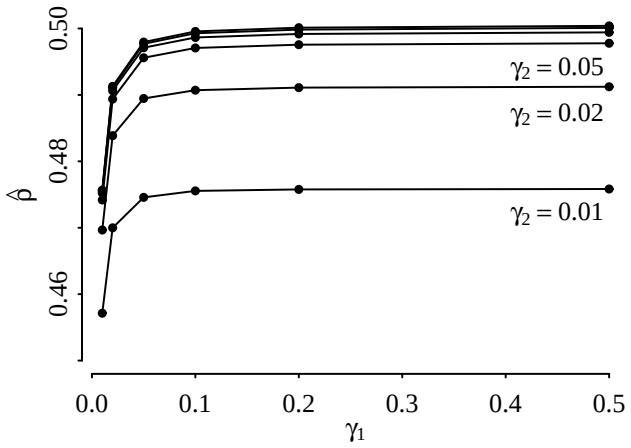

Figure 1: Pearson correlation coefficient

in the Pearson correlation coefficient. Thereby exceeding the continuity index cutoff of 0.05 allows to estimate the Pearson correlation coefficient.

Figure 2 shows that the values of polyserial correlation are weakly dependent on the values of γ_1. This is because the first variable x_1 is assumed categorical. Therefore the quality of correlation estimation depends on the continuity of the second variable. The results shown in Fig. 3 does not depend on the values of γ_2. In these cases also, if the continuity index exceeds cutoff of 0.05 a bias in correlation coefficient is quite small.

6 Conclusions

It proposed to choose an appropriate correlation measure based on data type detection. The continuity index γ_k allows to detect whether variable is continuous or ordered categorical. Simulation study revealed that exceeding the γ_k cutoff of 0.05 for both variables allows to classify the variables as continuous, and the Pearson correlation coefficient can be calculated. If only γ_1 should exceed cutoff of 0.05 then you can use the polyserial correlation of ordered categorical x_2 and continuous x_1. If only γ_2 should exceed cutoff of 0.05 it makes sense to estimate the polyserial correlation of ordered categorical x_1 and continuous x_2. The polychoric correlation coefficient provides best quality of estimation regardless of continuity index. But the calculation of the polychoric correlation coefficient is very slow if the number of categories is large. Therefore it is recommended to use the polychoric correlation if both variables have the continuity index less than 0.05.

Acknowledgements

The reported study was funded by Russian Ministry of Education and Science, according to the research project No. 2.2327.2017/ПЧ.

References

[Hol2010] F. P. Holgado-Tello, et. al. Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity*, 44(1):153–166, 2010.

[Dra1988] F. Drasgow. *Encyclopedia of statistical sciences / Polychoric and polyserial correlations.* John Wiley & Sons, 1988.

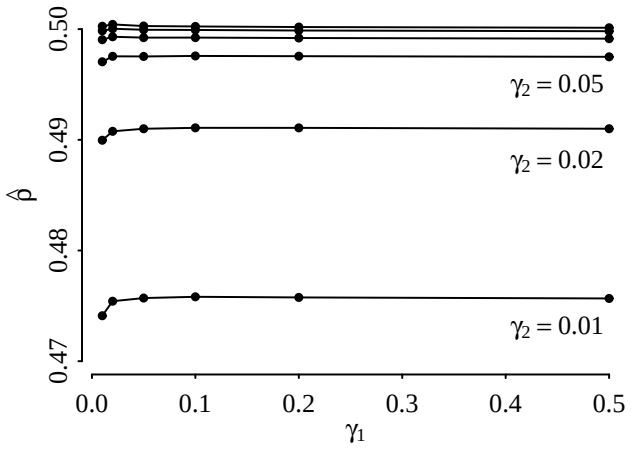

Figure 2: Polyserial correlation of ordered categorical x_1 and continuous x_2

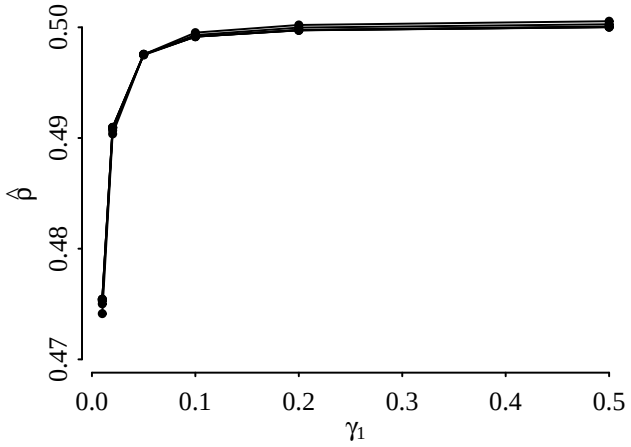

Figure 3: Polyserial correlation of ordered categorical x_2 and continuous x_1

[Pea1913] K. Pearson, and D. Heron. On theories of association. *Biometrika*, 9(1/2):159–315, 1913.

[Ols1979] U. Olsson. Maximum Likelihood Estimation of the Polychoric Correlation Coefficient. *Psychometrica*, 44(4):443–460, 1979.

[Kol2004] S. Kolenikov, and G. Angeles. *The use of discrete data in PCA: theory, simulations, and applications to socioeconomic indices.* Chapel Hill, Carolina Population Center, University of North Carolina, 2004.

[Pit1979] E. J. G. Pitman. *Some basic theory for statistical inference.* London, Chapman and Hall, 1979.

[Mey2013] C. Meyer. Recursive Numerical Evaluation of the Cumulative Bivariate Normal Distribution. *Journal of Statistical Software*, 52(10):1–14, 2013.