SHORT TECHNICAL NOTE

Iteratively Reweighted Least Squares Method for Estimating Polyserial and Polychoric Correlation Coefficients

Peng Zhang\textsuperscript{a}, Ben Liu\textsuperscript{a}, and Jingjing Pan\textsuperscript{b}

\textsuperscript{a}School of Mathematical Sciences, Zhejiang University, Hangzhou, China; \textsuperscript{b}Zhejiang Super Soul Artificial Intelligence Research Institute, Deqing, China

ABSTRACT
An iteratively reweighted least squares (IRLS) method is proposed for the estimation of polyserial and polychoric correlation coefficients in this article. It calculates the slopes in a series of weighted linear regression models fitting on conditional expected values. For polyserial correlation, conditional expectations of the latent predictor is derived from the observed ordinal categorical variable, and the regression coefficient is obtained with the weighted least squares method. In estimating polychoric correlation coefficient, conditional expectations of the response variable and the predictor are updated in turns. Standard errors of the estimators are obtained using the delta method based on data summaries instead of the whole data. Conditional univariate normal distribution is exploited and a single integral is numerically evaluated in the proposed algorithm, comparing to the double integral computed numerically based on the bivariate normal distribution in the traditional maximum likelihood (ML) approaches. This renders the new algorithm very fast in the estimation of both polyserial and polychoric correlation coefficients. Thorough simulation studies are conducted to compare the performances of the proposed method with the classical ML methods. Real data analyses illustrate the advantage of the new method in computing speed. Supplementary materials for this article are available online.

1. Introduction
In medical research, as well as in behavioral, educational, and psychological studies, it is common that the observed variables are measured on ordinal scales. For example, a kind of measure similar as Likert scale, proposed in Likert, Roslow, and Murphy (1934), is widely used to assess responses in surveys, allowing individuals to express how much respondents agree or disagree with a particular statement on a five (or seven) point scale. These categorical variables can be treated as being discretized from an underlying continuous variable for degree of agreement on the statement. There are also many examples of quantitative variables that are discretized explicitly in social science studies. For instance, when asking questions about sensitive or personal quantitative attributes (income, alcohol consumption), the nonresponse rate may often be reduced by simply asking the respondent to select one of two very broad categories (under $30K/ over $30K, etc.). When analyzing this kind of data, as pointed by Robitzsch (2020), it is common in practice to assign each of the categories would be considered as being included in the same score with a resulting reduction in data variability.

Yet another popular approach is the latent variable model, see Moustaki (2000), which usually requires much smaller number of latent variables than the number of observed variables. The most common choice for the distribution of the latent variables is the normal distribution because all covariances between the latent variables can be fully captured by the covariance matrix and each of its elements can be estimated using a bivariate normal distribution separately. The correlation in the standard bivariate normal distribution is called tetrachoric correlation based on $2 \times 2$ contingency table was suggested by Pearson (1900). The tetrachoric correlation was generalized to the poly-choric correlation for the case where the observed variables $X$ and $Y$ have $r$ and $s$ ordinal categories by Ritchie-Scott (1918) and Pearson and Pearson (1922) in the early 20th century.

Pearson correlation has been considered a less suitable method for studying the degree of association between categorical variables for several reasons. First, from a methodological point of view these variables would imply ordinal scales, whereas the Pearson correlation assumes interval measurement scales. Furthermore, the only information provided by this kind of scale is the number of subjects in each of the categories (cells) in a contingency table; if the Pearson correlation is used in this case the relationship between measures would be artificially restricted due to the restrictions imposed by categorization (Gilley and Uhlig (1993)), since all subjects situated in the interval that limits each of the categories would be considered as being included in the same category and, therefore, they would be assigned the same score with a resulting reduction in data variability.
Holgado-Tello et al. (2010) illustrated the advantages of using polychoric rather than Pearson correlations in exploratory factor analysis (EFA) and confirmatory factor analysis (CFA), taking into account that the latter require quantitative variables measured in intervals, and that the relationship between these variables has to be monotonic. The matrix of estimated polychoric and/or polyserial coefficients is used as inputs for the typical parameter estimation in structural equation modeling (Muthén 1984). More recently, network research has gained substantial attention in psychological sciences, which is called psychological network research. Psychological networks have been used in various different fields of psychology Epskamp, Borsboom, and Fried (2018). The Gaussian graphical model (GGM) was proposed by Lauritzen (1996), in which edges were directly interpreted as partial correlation coefficients. The GGM requires an estimate of the covariance matrix as input, for which polychoric and polyserial correlations can also be used in case the data are ordinal.

Let $X$ be an observed ordinal variable determined by an underlying latent continuous random variable $Z_1$ and $Y$ represent another observed continuous variable. It is assumed that the joint distribution of $Z_1$ and $Y$ is bivariate normal. The product moment correlation between $X$ and $Y$ is called the point polyserial correlation, while the correlation between $Z_1$ and $Y$ is called the polyserial correlation. The maximum likelihood estimate (MLE) MLE of the polyserial correlation was derived by Cox (1974). Olsson (1979) presented two approaches to estimating the MLE of the polychoric correlation: the so-called two-step method which first estimates the unknown thresholds, $a$ and $b$, from the marginal frequencies of the table and then finds the MLE of the $\rho$ conditioning on the estimated thresholds. The second approach is to find the joint MLE of $(\rho, a, b)$ from the likelihood function. Olsson, Dragsov, and Dorans (1982) derived the relationship between the polychoric and point polyserial correlation and compared the MLE of polyserial correlation with a two-step estimator and with a computationally convenient ad hoc estimator.

The estimation of polychoric correlation and polyserial correlation using the ML methods has been implemented in many popular statistical softwares such as Mplus by Muthén and Muthén (2005), PRELIS Jöreskog (2005), PROC CORR in SAS (with POLYCHORIC or OUTPLC = options) see Institute (2017), polycor package in R by Fox (2010) and psych package in R by Revelle (2017), POLYCORR program by Uebersax (2010), and an extensive list of software for computing the polychoric correlation by Uebersax (2006).

A Bayesian approach to estimating tetrachoric and polychoric correlation coefficients was proposed by Albert (1992). The author used a latent bivariate normal distribution to estimate a polychoric correlation coefficient from the Bayesian point of view by using the Gibbs sampler. One attractive feature of this method is that it can be generalized in a straightforward manner to handle a number of nonnormal latent distributions. They generalized their method to handle bivariate lognormal and bivariate t latent distributions in their simulations. Chen and Choi (2009) and Choi et al. (2011) proposed two new Bayesian estimators, maximum a posteriori (MAP) and expected a posteriori (EAP), and compared them to the ML method.

The aforementioned methods are all based on likelihood function assuming a bivariate normal distribution. In order to evaluate the cumulative distribution functions (CDF) of the bivariate normal distributions, they inevitably involve computations of the double integrals numerically, which require significant more computational resources than computing single integrals in the CDF of a univariate normal distribution.

In this article we propose an iteratively reweighted least square method to estimate the polyserial and the polychoric correlation coefficients. It is computationally fast because the complex bivariate integrations are circumvented completely. It is motivated by the fact that the Pearson’s correlation coincides with the slope of the regression line for paired standard normal data. When one of the paired continuous data is discretized, an unbiased estimator of the slope is derived from the generated categorical data. When both of the paired data are discretized, the slope of the regression line, that is, the correlation coefficient of the two normal random variables, will be obtained iteratively from a series of similar estimation procedures. The detail of the algorithm can be found in Section 2. In Sections 3 and 4, we conduct simulation studies and data analyses to compare the proposed algorithm with the ML method. At last, we conclude with discussions and some works can be done in the future to improve the proposed method.

### 2. Iteratively Reweighted Least Squares Algorithm

Assume $(Z_1, Z_2)^T \sim N_2(0, R)$ where $0 = (0, 0)^T$ and $R = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, $-1 \leq \rho \leq 1$. Conditioning on $Z_1$, $Z_2|Z_1 \sim N(\rho Z_1, 1-\rho^2)$. Hence,

\[ E(Z_2|Z_1) = \rho Z_1. \]  

(1)

This represents a simple linear regression model fitting $Z_2$ on $Z_1$ and $\rho$ is the slope of the regression line. Therefore, $\rho$, the Pearson correlation of $Z_1$ and $Z_2$, can be estimated from such a linear regression model.

#### 2.1. Polyserial Correlation Coefficients

Consider the case where one of the paired random variables, namely $Z_1$, is discretized into an ordinal polychotomous variable, $X$, and the other is observed as a continuous variable, $Y$. Let $X$ be an observed ordinal variable with $s$ categories, generated from the latent variable $Z_1$ with $X = i$ if $a_{i-1} < Z_1 \leq a_i$, $i = 1, \ldots, s$, where $a_s$ are thresholds with $a_0 = -\infty$ and $a_s = \infty$.

If $Z_1$ were observable, it would have been given from the regression line that $E(Y|Z_1) = \rho Z_1$. Taking expectation with respect to $Z_1$,

\[ E[E(Y|Z_1)] = \rho E(Z_1). \]

It holds for every $Z_1$ such that $a_{i-1} < Z_1 \leq a_i$, or correspondingly, $X = i$, for $i = 1, 2, \ldots, s$. That is,

\[ E[E(Y|Z_1, a_{i-1} < Z_1 \leq a_i)] = E[E(Y|X = i)] = \rho E(Z_1|a_{i-1} < Z_1 \leq a_i), \]  

(2)

for $i = 1, \ldots, s$. 

Denote $E(Y | X = i)$ by $E_{Y_i}$ and $E(Z_1 | a_{i-1} < Z_1 \leq a_i)$ by $e_{x_i}$.

(2) becomes

$$E(E_{Y_i}) = \rho e_{x_i},$$

for $i = 1, \ldots, s$.

This is a regression model without an intercept, in which $E_{Y_i}$ is the response variable and $e_{x_i}$ is the explanatory variable, with $\rho$ being the regression coefficient. Because $E_{Y_i}$'s have unequal variances, $\rho$ cannot be estimated with an ordinary least squares method. However, $\rho$ can be estimated using a weighted least squares method with a diagonal weight matrix.

It is easy to show that the density function of $E_{Y_i}$ is

$$f(y) = \frac{\Phi \left( \frac{a_{i-1} - \mu_i}{\sqrt{1 - \rho^2}} \right) - \Phi \left( \frac{a_i - \mu_i}{\sqrt{1 - \rho^2}} \right)}{P_i},$$

where $P_i = \Pr(X = i)$. The mean and variance of $E_{Y_i}$, $\mu_i$, and $\sigma_i^2$, are given by

$$\mu_i = \rho \frac{\phi(a_{i-1}) - \phi(a_i)}{P_i},$$

$$\sigma_i^2 = 1 + \rho^2 \frac{\phi(a_{i-1}) - a_i \phi(a_i)}{P_i},$$

$$-\rho^2 \frac{\phi(a_{i-1}) - \phi(a_i))^2}{P_i^2}.$$  \hfill(4)

Let $y_{ij}, y_{i2}, \ldots, y_{in_i}$ be the observed response variables associated with $X = i$ and $a_{i-1} < Z_{ij} \leq a_i$ for $j = 1, \ldots, n_i$, where $n_i$ is the size of data with $X = i$, $E_{Y_i}$ is estimated by

$$\hat{E}_{Y_i} = \bar{y}_{X=i} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}.$$  \hfill(5)

Since $Z_1$ has a truncated normal distribution with lower and upper limits $a_{i-1}$ and $a_i$, respectively, $e_{x_i}$ is the expected value of the truncated normal distribution, given by

$$e_{x_i} = \frac{\phi(a_{i-1}) - \phi(a_i)}{P_i},$$  \hfill(6)

where $\phi(\cdot)$ is the density function of the standard normal distribution. Let $CP_i = \Pr(X \leq i) = \Phi(a_i) = \sum_{j=1}^{i} P_j$, $i = 1, \ldots, s$. Then $\hat{a}_i = \Phi^{-1}(\hat{CP}_i)$, and $e_{x_i}$ in (6) is estimated by

$$\hat{e}_{x_i} = \frac{\phi(\hat{a}_{i-1}) - \phi(\hat{a}_i)}{\hat{P}_i} = \frac{\phi(\Phi^{-1}(\hat{CP}_{i-1})) - \phi(\Phi^{-1}(\hat{CP}_i))}{\hat{P}_i}.$$  \hfill(7)

Let $\hat{E}_x = (\hat{e}_{x1}, \hat{e}_{x2}, \ldots, \hat{e}_{xs})^T$, $\hat{E}_y = (\hat{E}_{Y1}, \hat{E}_{Y2}, \ldots, \hat{E}_{Ys})^T$, and

$$\hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_1^2/n_1 & 0 & \cdots & 0 \\ 0 & \hat{\sigma}_2^2/n_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\sigma}_s^2/n_s \end{bmatrix},$$

the regression coefficient is given by the weighted least squares method

$$\hat{\rho} = (\hat{E}_x^T \hat{\Sigma}^{-1} \hat{E}_x)^{-1} \hat{E}_x^T \hat{\Sigma}^{-1} \hat{E}_y = \frac{\sum_{i=1}^{s} n_i \hat{\sigma}_i^2 Z_{ij} \hat{e}_{y_i}}{\sum_{i=1}^{s} n_i \hat{\sigma}_i^2 \hat{e}_{x_i}^2}.$$  \hfill(8)

Algorithm 1 IRLS algorithm to estimate the polyserial correlation coefficient

**Input:** observed continuous variable $y$ and ordinal variable $x$

**Output:** $\hat{\rho}$, polyserial correlation coefficient of $Y$ and $X$; $\sqrt{\text{var}(\hat{\rho})}$, the standard error of $\hat{\rho}$

1. Compute $\hat{P}_i, \hat{CP}_i$ from data $x$. Estimate the thresholds $\hat{a}_i = \Phi^{-1}(\hat{CP}_i)$, $i = 1, \ldots, s$.
2. Compute $\hat{e}_{x_i}, \hat{e}_{y_i}$ using the formula in (7) and (5) for $i = 1, \ldots, s$.
3. Compute the initial $\hat{\rho}$ using the Pearson correlation coefficient between $y$ and $x$.
4. Set $\text{iter} = 0$, $\text{diff} = 1$, $n = 100$, $\epsilon = 1$e$-8$ and $\hat{\rho}_0 = \hat{\rho}$.
5. while $\text{iter} < n$ & $\text{diff} > \epsilon$ do
   6. Compute $\hat{\sigma}_i^2$ using the formula in (4) with $\hat{\rho}_0$, for $i = 1, \ldots, s$.
   7. Update $\hat{\rho}$ using the formula in (8).
   8. Compute $\text{diff} = |\hat{\rho} - \hat{\rho}_0|$.
   9. Update $\hat{\rho}_0 = \hat{\rho}$ and $\text{iter} = \text{iter} + 1$.
10. end while
11. Compute $\text{var}(\hat{\rho})$ using the formula in (9).
12. return $\hat{\rho}$ and $\sqrt{\text{var}(\hat{\rho})}$.

While $\hat{\sigma}_i^2$ in (4) depends on $\rho$, it can be obtained iteratively using the formula in (8), with the Pearson correlation coefficient as the initial value. The variance of $\hat{\rho}$ is given by

$$\text{var}(\hat{\rho}) = \left( \hat{E}_x^T \Sigma^{-1} \hat{E}_x \right)^{-1} \left( \sum_{i=1}^{s} n_i \hat{\sigma}_i^2 \hat{e}_{x_i}^2 \right)^{-1},$$  \hfill(9)

and the standard error of $\hat{\rho}$ is $\sqrt{\text{var}(\hat{\rho})}$.

The details of the IRLS method for estimating polyserial correlation coefficient are given in the following Algorithm 1.

### 2.2. Tetrachoric and Polychoric Correlation Coefficients

When both of the paired normal variables are discritized into ordinal variables, neither the response nor the predictor variable is observable. Both the weights and the response parts in (8) have to be updated after $\hat{\rho}$ is obtained in each iteration. Hence, we call it the iteratively reweighted least squares algorithm. The correlation coefficient between observed categorical variables is smaller than that between the latent continuous variables in magnitude. However, the same procedure given in Section 2.1 will update the estimate of the correlation coefficient, making it closer to the true parameter than the previous one. Therefore, the polychoric correlation coefficient can be estimated iteratively by updating conditional expectations of the two latent variables in turns. As shown subsequently, the estimates are obtained by replacing the conditional expectations with the contingency table statistics, that is sample moments. Therefore, it is a moment method that exhibits the consistency by the law of large number and the series of the estimates will converge to the polychoric correlation coefficient.

Let $(Z_1, Z_2)^T \sim N_2(\mathbf{0}, \mathbf{R})$ as above. $Z_1$ and $Z_2$ are categorized into $X$ and $Y$ with $s$ and $r$ categories, respectively. Assume that $X$ is generated from the latent variable $Z_1$ with $X = i$ if $a_{i-1} < Z_1 \leq a_i$, $i = 1, \ldots, s$, where $a_i$'s are thresholds with $a_0 = -\infty$. 
and $a_i = \infty$. Similarly, $Y$ is generated from $Z_2$ with $Y = j$ if $b_{j-1} < Z_2 \leq b_j, j = 1, \ldots, r$, where $b_j$ are thresholds with $b_0 = -\infty$ and $b_r = \infty$. The data usually consist of an array of observed frequencies $n_{ij}$, the number of $X = i, Y = j$, for $i = 1, \ldots, s; j = 1, \ldots, r$. Let $P_{ij} = P(X = i, Y = j) = P(a_{i-1} < Z_1 \leq a_i, b_{j-1} < Z_2 \leq b_j)$ be the proportion of data in cell $(i,j)$. The cumulative marginal proportions of the table are $P_i = \sum_{k=1}^i \sum_{j=1}^r P_{kj}, i = 1, \ldots, s, P_j = \sum_{i=1}^s \sum_{k=1}^r P_{ik}, j = 1, \ldots, r$.

The IRLS method for estimating the tetrachoric correlation coefficient ($s = r = 2$) is derived first. Suppose that $X$ and $Y$ are two dichotomous variables, generated from $Z_1$ and $Z_2$, respectively. Let $N = \begin{pmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{pmatrix}$ be the contingency table with $N_{ij}, i, j = 1, 2$ the frequencies of $X$ by $Y$ at the category $i$ and $j$, respectively, $N = N_{11} + N_{12} + N_{21} + N_{22}$.

Since $Z_2$ is not observable, the regression model in (2) becomes

$$E(Y|Z_2 = i) = \rho E(Z_1|a_{i-1} < Z_1 \leq a_i), \quad i = 1, 2. \quad (10)$$

Denote the explanatory variables on the right-hand side of (10) by $e_{xi}$. It is given by

$$e_{x1} = E(Z_1|Z_1 \leq a) = \frac{-\phi(a)}{1 - \rho^2}, \quad e_{x2} = E(Z_1|Z_1 > a) = \frac{-\phi(a)}{1 - \rho^2}, \quad (11)$$

where $a = \phi^{-1}(P_1)$ and $b = \Phi^{-1}(P_1)$ are the cutoff points, $P_1 = \frac{N_{11} + N_{21}}{N}$ is the proportion of $X = 1$ and $\tilde{P}_1 = \frac{N_{11} + N_{12}}{N}$ is the proportion of $Y = 1$. Similarly, $\tilde{P}_2$ and $\tilde{P}_2$ are proportions of $X = 2$ and $Y = 2$, respectively. $e_{x1}$ and $e_{x2}$ are estimated with formulas in (11) by plugging in the observed frequencies of the contingency table, and are used as the initial value of predictors in the regression model (10).

Denote the response variable of the regression model in (10) by $E_{Y_i} = E(Y|Z_2 = i), \quad i = 1, 2$. It cannot be calculated directly. But because $Z_2$ is dichotomized by whether $Z_2 \leq b$ or $Z_2 > b$ into two categories, indicated by either $Y = 1$ or $Y = 2$, $E_{Y_i}$ can be obtained based on the binomial distribution of $Y$, or into which of the two intervals $Z_2$ falls. This procedure is presented in the following two steps:

First, $Z_1$ and $Z_2$ are jointly normally distributed, $Z_2|Z_1 \sim N(\rho Z_1, 1 - \rho^2)$. Denote $\frac{1 - \rho^2}{\sqrt{1 - \rho^2}}$ by $Z_b$, the conditional mean of $Z_2$ given $Z_1$ at $e_{x1}$ and $e_{x2}$ for different categories of $Y$ are

$$e_{11} = E(Z_2|Z_1 \leq b, Z_1 = e_{x1}) \quad \rho \rho e_{x1} - \sqrt{1 - \rho^2} \phi \left( \frac{e_{x1}}{\rho} \right),$$

$$e_{21} = E(Z_2|Z_1 \leq b, Z_1 = e_{x2}) = \rho e_{x2} - \sqrt{1 - \rho^2} \phi \left( \frac{e_{x2}}{\rho} \right), \quad (12)$$

$$e_{12} = E(Z_2|Z_1 > b, Z_1 = e_{x1}) = \rho \rho e_{x1} + \sqrt{1 - \rho^2} \phi \left( \frac{e_{x1}}{\rho} \right),$$

$$e_{22} = E(Z_2|Z_1 > b, Z_1 = e_{x2}) = \rho e_{x2} + \sqrt{1 - \rho^2} \phi \left( \frac{e_{x2}}{\rho} \right),$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density function and the cumulative distribution function of the standard normal distribution, respectively.

Second, $E_{Y_i}$ is obtained based on the conditional expected values in (12). Because $Z_2$ is dichotomized into $Y$, with two mass probabilities, then

$$E_{Y_1} = \frac{p_{11}}{p_{11} + p_{12}} e_{11} + \frac{p_{12}}{p_{11} + p_{12}} e_{12},$$

$$E_{Y_2} = \frac{p_{21}}{p_{21} + p_{22}} e_{21} + \frac{p_{22}}{p_{21} + p_{22}} e_{22}, \quad (13)$$

where $p_{ij}, i = 1, 2; j = 1, 2$ is the proportion in cell $(i, j)$ of the probability table:

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}.$$

Similar as in model (2), $E_{Y_i}$ have unequal variances. They are not independent. Yet the regression coefficient can be obtained with a weighted least square method. Denote the covariance matrix of $\hat{P}$ by $B$, which is given by

$$B = \begin{pmatrix} p_{11}(1 - p_{11}) & p_{11}p_{12} & -p_{11}p_{12} & -p_{11}p_{22} \\ -p_{11}p_{12} & p_{12}(1 - p_{12}) & -p_{11}p_{22} & -p_{12}p_{22} \\ -p_{11}p_{12} & -p_{11}p_{12} & p_{12}(1 - p_{12}) & -p_{12}p_{22} \\ -p_{11}p_{12} & -p_{11}p_{12} & -p_{12}p_{22} & p_{12}(1 - p_{12}) \end{pmatrix} / N.$$

$E_{Y_i}$ is multivariate differentiable function of $P$. The partial derivative matrix of $E_{Y_i}$ with respect to $P$ is denoted by $D$, which is presented in the appendix. The covariance matrix of vector $E_Y = (E_{Y_1}, E_{Y_2})^T$, is given by $\Sigma = B^{-1}D^T$, using the delta method. $\hat{P}$ can be obtained by replacing all $P_{ij}$ with the observed frequencies $\hat{P}_{ij}$. Let $e_{x} = (e_{x1}, e_{x2})^T$, the regression coefficient in model (10) can be calculated using a weight least square method as

$$\hat{\rho} = (e_x^T \hat{\Sigma}^{-1} e_x)^{-1} e_x^T \hat{\Sigma}^{-1} \hat{E}_Y, \quad (15)$$

and

$$\text{var}(\hat{\rho}) = (e_x^T \hat{\Sigma}^{-1} e_x)^{-1}. \quad (16)$$

The standard error of $\hat{\rho}$ is $\sqrt{\text{var}(\hat{\rho})}$. The predictors in (10) are not constants. They have to be updated with the improved estimate $\hat{\rho}$ from (15). This also consists of the following two steps. First, we obtain the conditional expected values of $Z_1$ given $Z_2$ for different categories of $Z_1$, using a similar derivation of (12)–(13),

$$E_{Y_{1i}} = \frac{p_{11}}{p_{11} + p_{12}} e_{11} + \frac{p_{21}}{p_{21} + p_{22}} e_{21}, \quad (17)$$

$$E_{Y_{2i}} = \frac{p_{12}}{p_{11} + p_{12}} e_{12} + \frac{p_{22}}{p_{21} + p_{22}} e_{22}. \quad (18)$$

Second, we update $e_{x1}$ and $e_{x2}$ by taking expectations over $Z_2$,

$$e_{x1} = \frac{p_{11}}{p_{11} + p_{12}} e_{x1} + \frac{p_{12}}{p_{11} + p_{12}} e_{x2},$$

$$e_{x2} = \frac{p_{21}}{p_{21} + p_{22}} e_{x1} + \frac{p_{22}}{p_{21} + p_{22}} e_{x2}.$$
where
\[
e_{s11} = \hat{E}(Z_1 | Z_1 \leq a, Z_2 = E_{Y_{s1}}) = \rho E_{Y_{s1}} - \sqrt{1 - \rho^2} \frac{P_i}{\Phi(E_{Y_{s1}})} \\
= \rho E_{Y_{s1}} - \sqrt{1 - \rho^2} \frac{E_{Y_{s1}}^d}{\Phi(E_{Y_{s1}})} ,
\]
\[
e_{s12} = \hat{E}(Z_1 | Z_1 \leq a, Z_2 = E_{Y_{s2}}) = \rho E_{Y_{s2}} - \sqrt{1 - \rho^2} \frac{E_{Y_{s2}}^d}{\Phi(E_{Y_{s2}})} ,
\]
\[
e_{s21} = \hat{E}(Z_1 | Z_1 > a, Z_2 = E_{Y_{s1}}) = \rho E_{Y_{s1}} + \sqrt{1 - \rho^2} \frac{E_{Y_{s1}}^d}{\Phi(E_{Y_{s1}})} ,
\]
\[
e_{s22} = \hat{E}(Z_1 | Z_1 > a, Z_2 = E_{Y_{s2}}) = \rho E_{Y_{s2}} + \sqrt{1 - \rho^2} \frac{E_{Y_{s2}}^d}{\Phi(E_{Y_{s2}})} .
\]

Then we go back to (12) and repeat the previous procedure until \( \hat{\rho} \) converges. Because both the weight matrix and the response vector in (15) have been updated, we call this an iteratively reweighted least squares algorithm. The proposed IRLS algorithm proceeds in sequence: \( \mathbf{e}_s \rightarrow \mathbf{E}_Y, \mathbf{S} \rightarrow \rho \rightarrow \mathbf{e}_s \rightarrow \mathbf{E}_Y, \mathbf{S} \rightarrow \rho, \text{var}(\rho) \). Figure 1 shows the change of the values used in (15).

The red star points are the \( \hat{\mathbf{e}}_s, \hat{\mathbf{E}}_Y \) and the blue points are \( \hat{\mathbf{e}}_s, \hat{\mathbf{E}}_Y \). It can be seen that the algorithm converges fast and stops in several iterations. The slope of the red line is the direct Pearson correlation and is apparently smaller than the true \( \rho \) which is green line. The slope of the black line is the estimated tetrachoric correlation coefficient from the proposed IRLS algorithm. Clearly, as the algorithm proceeds, the results of our algorithm is getting closer to the true value. The details of the IRLS algorithm for estimating tetrachoric correlation coefficient are given in the following Algorithm 2.

Figure 2 illustrate the whole process of estimating the tetrachoric correlation coefficient with the Algorithm 2.

Next, we generalize Algorithm 2 to the case where \( X \) and \( Y \) have \( s \) and \( t \) categories, respectively, that is, to estimate the polychoric correlation coefficient. As in the dichotomous case,

Algorithm 2 IRLS method to estimate the tetrachoric correlation coefficient

Input: observed ordinal data \( y \) and \( x \) or contingency table
Output: tetrachoric correlation coefficient \( \hat{\rho} \) of \( Y \) and \( X \), s.e.(\( \hat{\rho} \))
1: Calculate the frequency table \( \hat{P} = (\hat{P}_{ij}), i,j = 1,2 \), the covariance \( \hat{B} \) using (14).
2: Estimate the thresholds given from \( \hat{P} = \hat{P}_{ij}, i,j = 1,2 \), \( \hat{\rho}_{ij} = \Phi^{-1}(\hat{P}_{ij}) \).
3: Initialization:
4: \( \hat{\rho}_0 = \text{Pearson correlation coefficient of } x \text{ and } y (\text{ordinal data}), \hat{\rho} = \Phi^{-1}(\hat{P}_{ij}) \).
5: \( \hat{\mathbf{e}}_X \) and \( \hat{\mathbf{e}}_Y \) using the formulas in (11)
6: Set \( \text{iter} = 0, \text{diff} = 1, n = 100 \) and \( \epsilon = 1e-8 \).
7: while \( \text{iter} < n \) and \( \text{diff} > \epsilon \) do
8: Compute \( \hat{E}_{Y_1}, \hat{E}_{Y_2} \) using (12) and (13).
9: Compute the derivative matrix \( \hat{D} \) using (24), and \( \hat{\Sigma} = \hat{D} \hat{D} \) using (14).
10: Compute \( \hat{\rho} \) using the formula in (15).
11: Compute \( \text{diff} = |\hat{\rho} - \hat{\rho}_0| \).
12: Update \( \hat{\mathbf{e}}_{X_1} \) and \( \hat{\mathbf{e}}_{X_2} \) using the formula in (17), (19), and (18).
13: Update \( \hat{\rho}_0 = \hat{\rho} \) and \( \text{iter} = \text{iter} + 1 \).
14: end while
15: Compute \( \text{var}(\hat{\rho}) \) using the formula in (16).
16: return \( \hat{\rho}, \sqrt{\text{var}(\hat{\rho})} \).

thresholds are \( a_i = \Phi^{-1}(P_i), i = 1, \ldots, s, b_j = \Phi^{-1}(P_j), j = 1, \ldots, r \).

Initially, define
\[
e_{s1} = \hat{E}(Z_1 | a_{i-1} < Z_1 \leq a_i) = \frac{\phi(a_{i-1}) - \phi(a_i)}{P_i} \tag{20}
\]
for \( i = 1, \ldots, s \).

Conditional on \( Z_1 = e_{s1} \) and \( b_{j-1} < Z_2 \leq b_j \),
\[
e_{s1} = \hat{E}(Z_2 | b_{j-1} < Z_2 \leq b_j, Z_1 = e_{s1})
= \rho e_{s1} + \sqrt{1 - \rho^2} \frac{\phi(e_{s1} b_{j-1}) - \phi(e_{s1} b_j)}{\Phi(e_{s1} b_{j-1}) - \Phi(e_{s1} b_j)} \tag{21}
\]
for \( i = 1, \ldots, s; j = 1, \ldots, r \).

Figure 1. The track of the estimating process of the tetrachoric correlation coefficient.

Figure 2. Process of estimating the tetrachoric correlation coefficient.
Then the conditional expected value of $Z_2$ given $Z_1 = e_i$ is

$$ E_{Y_i} = E(Z_2 | Z_1 = e_i) = \sum_{j=1}^r \frac{P_{ij}}{P_i} e_j $$

for $i = 1, \ldots, s$.

Let frequency table $\hat{P}_{s \times t} = (\hat{P}_{ij}), i = 1, \ldots, s, j = 1, \ldots, r$. Then the covariance matrix of vector $\hat{P}$, by stacking $\hat{P}$ by row, is given by

$$ B = (B_{ij})_{s \times t \times s \times t} $$

$$ B_{ij} = \begin{cases} \frac{1}{N} P_i (1 - P_i), & i = j, \\ -\frac{1}{N} P_i P_j, & i \neq j. \end{cases} $$

(23)

For simplicity, we ignore the derivative brought by thresholds $b$ (This leads to a smaller final variance). The partial derivative matrix can be calculated as follows:

$$ \frac{\partial E_{Y_i}}{\partial P_{ij}} = \left\{ \begin{array}{ll} 0, & k \neq i \\ ((P_k - P_j) e_{kj} - \sum_{n=1 \neq j}^r P_{kn} e_{kn}) / P_k^2, & k = i. \end{array} \right. $$

(24)

Let $E_Y = (E_{Y_i})_{s \times t}, i = 1, \ldots, s, D = (D_{ij})_{s \times t \times s \times t}$ is block diagonal matrix:

$$ \begin{pmatrix} D_1 & 0 & \cdots & 0 \\ 0 & D_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D_s \end{pmatrix} $$

where $D_i$ is the nonzero elements derived by equation above which length $r$, and the covariance matrix is $\Sigma = DBD'$. Similar as in estimating tetrachoric correlation coefficient, $\hat{\rho}$ can be calculated with (15). Variance of $\hat{\rho}$ is calculated with (16). Standard error of $\hat{\rho}$ is obtained by taking the square root of its variance.

Then the predictors $e_{ij}$ are updated in the following two steps. First, compute

$$ E_{Y_{ij}} = E(Z_2 | b_{j-1} < Z_2 \leq b_j) = \sum_{i=1}^s \frac{P_{ij}}{P_j} e_{ij}. $$

Second, obtain

$$ e_{ij} = E(Z_1 | a_{i-1} < Z_1 \leq a_i) = \sum_{j=1}^r \frac{P_{ij}}{P_i} e_{ij}, $$

(25)

for $i = 1, \ldots, s$, where

$$ e_{ij} = E(Z_1 | a_{i-1} < Z_1 \leq a_i, Z_2 = E_{Y_{ij}}) $$

$$ = \rho \frac{\phi \left( E_{Y_{ij}} \right) - \phi \left( E_{Y_{ij}}^{a_{i-1}} \right)}{\phi \left( E_{Y_{ij}}^{a_i} \right) - \phi \left( E_{Y_{ij}}^{a_{i-1}} \right)} $$

(26)

for $j = 1, \ldots, r$.

The procedure is repeated until $\hat{\rho}$ converges.

The details of the IRLS algorithm for estimating polychoric correlation coefficient are given in the following Algorithm 3.

An R package IRLSpoly is developed to implement the Algorithms 1, 2, and 3. It can be installed in R using the command install_github ("encoreus/IRLSpoly").

### 3. Simulation Study

In this section, we conduct a series of simulation studies to compare the proposed algorithm with the standard maximum-likelihood method. All tests are carried out on a computer with an Intel (R) Core (TM) i7-10700 CPU @ 2.90G赫兹 and 64G memory. Following Choi et al. (2011), the sample size is set to $N = 30, 50, 100, 500, 1000$. These numbers were chosen to reflect from small to moderate sample sizes that might be commonly encountered. The population correlation coefficient is set to $\rho = 0.2, 0.4, 0.6, 0.8$, ranging from null to moderate high. The number of categories for each ordinal variable are set to 2 (binary responses), 3, 5 and 7, that is, $r = s = 2, 3, 5, 7$, respectively. Repeat $n = 1000$ times.

We generate data from a bivariate normal distribution with correlation $\rho$ and discretized them into categorical data, following the same procedure as described in Figure 1 in Bollen and Barb (1981), where the mean of the distribution is taken as a reference point and the variable is divided into equally spaced intervals that move away from the mean of the distribution toward the extremes. Therefore, the distribution of the categorized data gets closer to normal as the number of categories increases.

To measure the performance of the two methods, we use the following criteria:

$$ \text{MEAN} = \sum_{i=1}^n \hat{\rho} / n, \text{MRB} = \sum_{i=1}^n (\hat{\rho}_i - \rho) / \sqrt{n} $$

$$ \text{RMSE} = \sqrt{\sum_{i=1}^n (\hat{\rho} - \rho)^2} / \sqrt{n} = \sqrt{\sum_{i=1}^n (\hat{\rho} - \text{MEAN})^2} / \sqrt{n} $$

$$ \text{MSD} = \sum_{i=1}^n \sqrt{\text{var}(\hat{\rho})} / n $$

### Algorithm 3 IRLS method to estimate the polychoric correlation coefficient

**Input:** observed ordinal data $y$ and $x$ or contingency table

**Output:** polychoric correlation of $Y$ and $X$, s.e.$(\hat{\rho})$

1. Calculate $\hat{P} = (\hat{P}_{ij}), i = 1, \ldots, s, j = 1, \ldots, r, the covari ance $B$ using (23)
2. Estimate the thresholds $\hat{a}_i$ and $\hat{b}_j$ from the cumulative marginal proportions of $\hat{P}$.
3. Initialization:
4. $\hat{\rho}_0$ = Pearson correlation coefficient of $x$ and $y$ (ordinal data), $0$ (contingency table)
5. Initialize $\hat{e}_{ij}$ using the formula in (20) for $i = 1, \ldots, s$
6. Set $\text{iter} = 0, \text{diff} = 1, n = 100$ and $\epsilon = 1 \times 10^{-8}$
7. while $\text{iter} < n$ & $\text{diff} > \epsilon$
8. Compute $e_{ij}$ for different categories using the formula in (21) for all $i,j$
9. Compute the derivation matrix $\hat{D}$ using (24), $\hat{E} = \hat{D} \hat{B} \hat{D}'$
10. Compute $\hat{E}_{yij}$ using the formula in (22) for $i = 1, \ldots, s$
11. Compute $\hat{\rho}$ using formula in (15)
12. Compute diff $= |\hat{\rho} - \hat{\rho}_0|$
13. Compute $E_{Y_{ij}}, e_{ij}$ using the formula in (26) for all $i,j$
14. Update $e_{ij}$ using the formula in (25) for $i = 1, \ldots, s$
15. Update $\hat{\rho}_0 = \hat{\rho}$ and $\text{iter} = \text{iter} + 1$
16. end while
17. Compute var$(\hat{\rho})$ using the formula in (16)
18. return estimate $\hat{\rho}$ and $\sqrt{\text{var}(\hat{\rho})}$ as the s.e. of $\hat{\rho}$. 


| $\rho$ | MEAN | SD | MRB | RMSE | MEAN | SD | MRB | RMSE | MEAN | SD | MRB | RMSE | MEAN | SD | MRB | RMSE |
|-------|------|----|-----|------|------|----|-----|------|------|----|-----|------|------|----|-----|------|------|
| 0.1   | 0.0012 | 0.0003 | 0.3721 | 0.3773 | 0.0017 | 0.0004 | 0.3698 | 0.3779 | 0.0011 | 0.0002 | 0.3692 | 0.3784 | 0.0016 | 0.0001 | 0.3689 | 0.3792 | 0.0016 | 0.0001 | 0.3686 | 0.3796 |
| 0.2   | 0.0015 | 0.0001 | 0.3711 | 0.3760 | 0.0017 | 0.0001 | 0.3698 | 0.3779 | 0.0011 | 0.0002 | 0.3692 | 0.3784 | 0.0016 | 0.0001 | 0.3689 | 0.3792 | 0.0016 | 0.0001 | 0.3686 | 0.3796 |
| 0.4   | 0.0017 | 0.0001 | 0.3702 | 0.3749 | 0.0017 | 0.0001 | 0.3698 | 0.3779 | 0.0011 | 0.0002 | 0.3692 | 0.3784 | 0.0016 | 0.0001 | 0.3689 | 0.3792 | 0.0016 | 0.0001 | 0.3686 | 0.3796 |
| 0.6   | 0.0019 | 0.0001 | 0.3696 | 0.3735 | 0.0017 | 0.0001 | 0.3698 | 0.3779 | 0.0011 | 0.0002 | 0.3692 | 0.3784 | 0.0016 | 0.0001 | 0.3689 | 0.3792 | 0.0016 | 0.0001 | 0.3686 | 0.3796 |

Table 1. Simulation results of polynomial correlation comparing the BLS algorithm with the ML method.
which, respectively, represent mean, mean relative bias, root mean squared error, standard deviation and mean value of the estimates of the variance of $\rho$ (Monte Carlo). We compare MSD with SD to verify the accuracy of the algorithm variance calculation formula.

For polyserial correlation, the results with true value of $\rho = 0, 0.2, 0.4, 0.6, 0.8$ are shown in Table 1. Function polyserial defined in the polycor package in R by Fox (2010) is invoked to obtain the MLEs.

It can be seen that the results of the two methods are similar, especially with $\rho = 0, 0.2, 0.4$. When $\rho = 0.6, 0.8$, for small sample size such as $N = 30, 50$, the bias of the new method is slightly larger than that of the ML method. However, as the sample size increases, it performs as well as the ML method. In addition, the mean value of the standard deviation calculated by the new method (MSD) is very close to that of Monte Carlo (SD).

The slightly larger biases from the IRLS are mainly due to the fact that only the means of the data, instead of the whole data, are used in estimating polyserial correlations. On the other hand, the advantages of the proposed method for polyserial correlation include: (a) it is obtained from data summaries. If only summaries were reported, for example, in a report or a paper, the proposed method can be applied to estimate the polyserial correlation, but the traditional ML method cannot; (b) In the simulation, the ML method may not be able to calculate because of the small $n$ and $s$ (leading a null category), while the proposed method still works.

For polychoric simulation, a quick two step maximum likelihood (Olsson 1979) procedure is adopted in calling polychor function in the polycor package of R. The regular ML method in which the thresholds are estimated simultaneously is too slow to run the simulations. We use the same settings as in simulations for polyserial correlations (Table 2).

The MEANs of the estimators of both IRLS and ML are close to the true values of $\rho$. IRLS can be considered unbiased. The SD of IRLS is slightly larger than that of ML, and the difference decreases as $\rho$, $s$, $r$, and $N$ increases, by the law of large numbers. The difference between these SDs is negligible when $N \geq 0.4$, $N \geq 100$ or $s, r \geq 5$.

The estimated standard derivation from the IRLS is smaller than the true value. However, the biases decrease with the increase of $\rho, s, r, N$.

Because the estimator of polychoric correlation from the IRLS algorithm is determined by proportions, not the frequencies of the contingency table, increasing sample size of the dataset does not reduce the speed of the estimation. In comparison, the log-likelihood function in the traditional ML methods is calculated by adding those of all data points. Additionally, only single integrals are evaluated in the IRLS when the cumulative distribution function $\Phi$ is invoked. But the traditional ML methods need to evaluated double integrals in calculating likelihood functions. Therefore, the IRLS is a much faster algorithm in estimating polychoric correlation comparing to the ML methods. Our simulations show that IRLS requires significantly less running time than the ML methods. Table 3 presents the running time of the simulation studies comparing with the ML and 2-steps method (IRLS calculate variance but the others not because they are too slow), with $\rho = 0.4, N = 500$ and
Table 2. Simulation results of polychoric correlation comparing the regression method with the ML method.

| N  | RLS | SD  | MB  | RMSE | MEAN | SD | MB | RMSE | MEAN | SD | MB | RMSE |
|----|-----|-----|-----|------|------|----|----|------|------|----|----|------|
| 30 | 0.323 | 0.230 | 0.175 | 0.001 | 0.2866 | 0.005 | 0.2252 | 0.001 | 0.1934 | 0.005 | 0.1653 | 0.001 |
| 50 | 0.323 | 0.230 | 0.175 | 0.001 | 0.2866 | 0.005 | 0.2252 | 0.001 | 0.1934 | 0.005 | 0.1653 | 0.001 |
| 100 | 0.323 | 0.230 | 0.175 | 0.001 | 0.2866 | 0.005 | 0.2252 | 0.001 | 0.1934 | 0.005 | 0.1653 | 0.001 |
| 500 | 0.323 | 0.230 | 0.175 | 0.001 | 0.2866 | 0.005 | 0.2252 | 0.001 | 0.1934 | 0.005 | 0.1653 | 0.001 |
| 1000 | 0.323 | 0.230 | 0.175 | 0.001 | 0.2866 | 0.005 | 0.2252 | 0.001 | 0.1934 | 0.005 | 0.1653 | 0.001 |

(Continued)
In this section, two real data analyses are conducted in comparing the performance of the proposed method and the ML method, including running time recorded. The first two datasets were obtained from Kirk (1973) and were also listed in the psych package as examples. The Big Five Inventory (bfi) data, include 25 personality self report items taken from 2800 subjects. The items were collected using a 6 point response scale: 1 for very inaccurate, 2 for moderately inaccurate, 3 for slightly inaccurate, 4 for slightly accurate, 5 for moderately accurate and 6 for very accurate. Three additional demographic variables (sex, education, and age) are also included. Figure 3 shows a scatterplot of matrices (SPLOM), with bivariate scatterplots below the diagonal, histograms on the diagonal, and the polychoric correlation coefficients with standard errors above the diagonal of the first five variables. Correlation ellipses are drawn in the same graph. The red lines below the diagonal are the LOESS smoothed lines, fitting a smooth curve between two variables.

It can be seen from Figure 3 that the estimates from the proposed method are very close to those from the ML method, with the largest difference being 0.016, smaller than its standard error estimated with both methods.

The running times are listed in Table 4. The user time is the CPU time spent in the execution of user instructions of the calling process. The system time is the CPU time charged for execution by the system on behalf of the calling process. Clearly, the proposed method requires significantly less CPU time than the ML method in estimating the polychoric correlation coefficients.

The second dataset came from the study of investigation on parenteral nutrition in hospitalized patients with digestive tract tumors in China. The study was initiated by the Committee of Cancer Rehabilitation and Palliative Care of China in 2017. In this survey, the nutritional risk and incidence of malnutrition of newly admitted hospitalized cancer patients were obtained through the nutritional risk screening and assessment. The survey consists of many questionnaires, including Quality of Life, KPS, Nutritional Risk Screening (NRS-2002). Patient-Generated Subjective Global Assessment (PG-SGA). The data had 23 categorical variables from 1086 patients. The proposed IRLS and the traditional ML methods are applied on this data.

Table 4 presents the running times of three competing methods. With moderate sample size (1086) and number of variables 1000 replicates. The user time is the CPU time charged for the execution of user instructions of the calling process. The system time is the CPU time charged for execution by the system on behalf of the calling process.

These results show that the IRLS method spends less running time than the ML and 2-steps methods. In addition, as the number of categories getting larger, the ML and 2-steps methods require significantly more running time, while the proposed IRLS algorithm does not slow down much.

4. Data Analysis

In this section, two real data analyses are conducted in comparing the performance of the proposed method and the ML method, including running time recorded. The first two datasets were obtained from Kirk (1973) and were also listed in the psych package as examples. The Big Five Inventory (bfi) data, include 25 personality self report items taken from 2800 subjects. The items were collected using a 6 point response scale: 1 for very inaccurate, 2 for moderately inaccurate, 3 for slightly inaccurate, 4 for slightly accurate, 5 for moderately accurate and 6 for very accurate. Three additional demographic variables (sex, education, and age) are also included. Figure 3 shows a scatterplot of matrices (SPLOM), with bivariate scatterplots below the diagonal, histograms on the diagonal, and the polychoric correlation coefficients with standard errors above the diagonal of the first five variables. Correlation ellipses are drawn in the same graph. The red lines below the diagonal are the LOESS smoothed lines, fitting a smooth curve between two variables.

It can be seen from Figure 3 that the estimates from the proposed method are very close to those from the ML method, with the largest difference being 0.016, smaller than its standard error estimated with both methods.

The running times are listed in Table 4. The user time is the CPU time spent in the execution of user instructions of the calling process. The system time is the CPU time charged for spent in the execution by the system on behalf of the calling process. Clearly, the proposed method requires significantly less CPU time than the ML method in estimating the polychoric correlation coefficients.

The second dataset came from the study of investigation on parenteral nutrition in hospitalized patients with digestive tract tumors in China. The study was initiated by the Committee of Cancer Rehabilitation and Palliative Care of China in 2017. In this survey, the nutritional risk and incidence of malnutrition of newly admitted hospitalized cancer patients were obtained through the nutritional risk screening and assessment. The survey consists of many questionnaires, including Quality of Life, KPS, Nutritional Risk Screening (NRS-2002). Patient-Generated Subjective Global Assessment (PG-SGA). The data had 23 categorical variables from 1086 patients. The proposed IRLS and the traditional ML methods are applied on this data.

Table 4 presents the running times of three competing methods. With moderate sample size (1086) and number of variables 1000 replicates. The user time is the CPU time charged for the execution of user instructions of the calling process. The system time is the CPU time charged for execution by the system on behalf of the calling process.

These results show that the IRLS method spends less running time than the ML and 2-steps methods. In addition, as the number of categories getting larger, the ML and 2-steps methods require significantly more running time, while the proposed IRLS algorithm does not slow down much.
Table 3. Running time of IRLS and ML methods.

| k   | IRLS User | IRLS System | IRLS Elapsed | ML method User | ML method System | ML method Elapsed | 2-step method User | 2-step method System | 2-step method Elapsed |
|-----|-----------|-------------|--------------|----------------|-----------------|-------------------|-------------------|---------------------|----------------------|
| 2   | 0.64      | 0.03        | 0.67         | 5.56           | 0.02            | 5.56              | 1.11              | 0.00                | 1.10                 |
| 3   | 0.67      | 0.02        | 0.69         | 9.21           | 0.05            | 9.28              | 1.02              | 0.03                | 1.07                 |
| 5   | 0.68      | 0.03        | 0.72         | 22.59          | 0.06            | 22.66             | 1.13              | 0.03                | 1.18                 |
| 7   | 0.70      | 0.05        | 0.73         | 45.97          | 0.11            | 46.11             | 1.05              | 0.01                | 1.06                 |

Polychoric correlation

| k   | IRLS User | IRLS System | IRLS Elapsed | ML method User | ML method System | ML method Elapsed | 2-step method User | 2-step method System | 2-step method Elapsed |
|-----|-----------|-------------|--------------|----------------|-----------------|-------------------|-------------------|---------------------|----------------------|
| 2   | 4.68      | 0.05        | 4.63         | 31.01          | 0.09            | 31.16             | 5.93              | 0.00                | 5.83                 |
| 3   | 4.42      | 0.05        | 4.52         | 113.31         | 0.06            | 113.35            | 11.80             | 0.01                | 11.82                |
| 5   | 4.18      | 0.03        | 4.16         | 1756.90        | 1.40            | 1758.69           | 31.65             | 0.06                | 31.61                |
| 7   | 4.70      | 0.00        | 4.74         | 2479.88        | 1.74            | 2482.21           | 61.34             | 0.04                | 61.29                |

Figure 3. Polychoric correlation coefficients of the bfi data.

Table 4. Running time of the IRLS and the ML, 2-steps method.

| N = 2800 | IRLS User | IRLS System | IRLS Elapsed | ML method User | ML method System | ML method Elapsed | 2-steps method User | 2-steps method System | 2-steps method Elapsed |
|----------|-----------|-------------|--------------|----------------|-----------------|-------------------|---------------------|----------------------|------------------------|
| 1.55     | 0.00      | 1.55        | 987.68       | 1.00           | 988.90          | 43.60             | 0.77                | 44.10                |

5. Conclusion and Discussion

In this article, we propose a new approach to estimating the polyserial and polychoric correlation coefficients. Simulation studies and data analyses show that the proposed IRLS algorithm can estimate polyserial and polychoric correlations consistently and efficiently. It also takes much less time to compute than the traditional ML methods. This prominent aspect of the new method may help in modern research with huge datasets to analyze. It makes studies on big data using polyserial or polychoric correlation plausible, such as in network data and text mining.

The basic idea of the new method is that the correlation coefficient of two continuous variables can be obtained from the slope of the derived regression models. The regression coefficient is not necessarily estimated from the whole data. Instead, it can be calculated from points at the conditional expected values and through which the regression line pass in theory. Hence, the regression coefficient, that is, the correlation coefficient, can be estimated with a derived new regression model, with updated weights for different response variables. Therefore, an iteratively reweighted least squares method is proposed to estimate the regression coefficient. This method is in effect a generalized

(a) Estimates and standard errors with the IRLS method

(b) Estimates and standard errors with the ML method

(23), the advantage of the proposed method in speed is manifested in that the running time of ML and 2-steps method is much longer than the IRLS.
moment method based on summary statistics that always generates consistent estimators.

The advantage of this method is obvious. It is applicable to summary data so it can be used for meta analysis that combines different studies with only summaries of data reported. Meta analysis on the polyserial and polychoric correlation will be another future work.

One limitation of the proposed method is that it only considers the estimation and inference of one polychoric/polyserial correlation. In some situations, it is common to use a matrix of estimated polychoric and/or polyserial correlation coefficients as inputs for estimating the parameters. Additionally, an estimate of the asymptotic covariance matrix of the estimated polychories/polyserials is used for inference, for example, for construction of an overall test statistic of correctness of the model. The methods presented in this article certainly can be extended to the multivariate cases, which will be one of our future works. Another possible aspects of the proposed method left to future study is its robustness to the distributional assumption. We assumed that the latent variables have an underlying bivariate normal distribution. Whether this is true is testable. However, it is out of the scope of this article. It is worthwhile to examine what degree of departures from the normality assumption has any effects on the correlation estimation in the future.

Supplementary Materials

IRLSpoly: An R package IRLSpoly is developed to implement the Algorithms 1, 2, and 3. It can be installed in R using the command `install_github("encoreus/IRLSpoly")`. More details can be seen in https://github.com/encoreus/IRLSpoly.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Funding

The authors gratefully acknowledge the Natural Science Foundation of Zhejiang Province (Project No. LDT23E05015A0, titled Isogeometric Analysis and Nonlinear Time Series Modeling of Error Field in Processing Domain of CNC Machine Tool).

References

Albert, J. H. (1992), "Bayesian Estimation of the Polychoric Correlation Coefficient," *Journal of Statistical Computation and Simulation*, 44, 47–61. [317]

Bollen, K. A., and Barb, K. H. (1981), "Pearson’s R and Coarsely Categorized Measures," *American Sociological Review*, 46, 232–239. [321]

Chen, J., and Choi, J. (2009), "A Comparison of Maximum Likelihood and Expected A Posteriori Estimation for Polychoric Correlation Using Monte Carlo Simulation," *Journal of Modern Applied Statistical Methods*, 8, 32. [317]

Choi, J., Kim, S., Chen, J., and Dannels, S. (2011), "A Comparison of Maximum Likelihood and Bayesian Estimation for Polychoric Correlation Using Monte Carlo Simulation," *Journal of Educational and Behavioral Statistics*, 36, 523–549. [317,321]

Cox, N. (1974), "Estimation of the Correlation between a Continuous and a Discrete Variable," *Biometrics*, 30, 171–178. [317]

Epksamp, S., Borsboom, D., and Fried, E. I. (2018), "Estimating Psychological Networks and their Accuracy: A Tutorial Paper," *Behavior Research Methods*, 50, 195–212. [317]
Fox, J. (2010), *Polycor: Polychoric and Polyserial Correlations* r package version 0.7-8. Online [http://www.cran.r-project.org/web/packages/polycor/index.html](http://www.cran.r-project.org/web/packages/polycor/index.html) (Accessed on 31 August 2010). [317,323]

Gilley, W. F., and Uhlig, G. E. (1993), "Factor Analysis and Ordinal Data," *Education*, 114, 258–265. [316]

Holgado-Tello, F. P., Chacón-Moscoso, S., Barbero-García, L., and Vila-Abad, E. (2010), "Polychoric versus Pearson Correlations in Exploratory and Confirmatory Factor Analysis of Ordinal Variables," *Quality & Quantity*, 44, 153–166. [317]

Institute, S. (2017), *Base SAS 9.4 Procedures Guide: Statistical Procedures*, SAS Institute. [317]

Jöreskog, K. G. (2005), "Structural Equation Modeling with Ordinal Variables Using Lisrel." Technical report. [317]

Kirk, D. B. (1973), "On the Numerical Approximation of the Bivariate Normal (tetrachoric) Correlation Coefficient," *Psychometrika*, 38, 259–268. [325]

Lauritzen, S. L. (1996), *Graphical Models* (Vol. 17), Oxford: Clarendon Press. [317]

Likert, R., Roslow, S., and Murphy, G. (1934), "A Simple and Reliable Method of Scoring the Thurstone Attitude Scales," *The Journal of Social Psychology*, 5, 228–238. [316]

Moustaki, I. (2000), "A Latent Variable Model for Ordinal Variables," *Applied Psychological Measurement*, 24, 211–223. [316]

Muthén, B. O. (1984), "A General Structural Equation Model with Dichotomous, Ordered Categorical, and Continuous Latent Variable Indicators," *Psychometrika*, 49, 115–132. [317]

Muthén, L. K., and Muthén, B. O. (2005), *Mplus: Statistical Analysis with Latent Variables: User’s Guide*, Los Angeles: Muthén & Muthén. [317]

Olsson, U. (1979), "Maximum Likelihood Estimation of the Polychoric Correlation Coefficient," *Psychometrika*, 44, 443–460. [317,323]

Olsson, U., Drasgow, F., and Dorans, N. J. (1982), "The Polyserial Correlation Coefficient," *Psychometrika*, 47, 337–347. [317]

Pearson, K. (1900), "Mathematical Contribution to the Theory of Evolution. VII. On the Correlation of Characters not Quantitatively Measurable," *Philosophical Transactions of the Royal Society of London*, 195, 1–47. [316]

Pearson, K., and Pearson, E. S. (1922), "On Polychoric Coefficients of Correlation," *Biometrika*, 14, 127–156. [316]

Revelle, W. R. (2017), "psych: Procedures for Personality and Psychological Research." Software. [317]

Ritchie-Scott, A. (1918), "The Correlation Coefficient of a Polychoric Table," *Biometrika*, 12, 93–133. [316]

Robitzsch, A. (2020), "Why Ordinal Variables Can (Almost) Always Be Treated as Continuous Variables: Clarifying Assumptions of Robust Continuous and Ordinal Factor Analysis Estimation Methods," *Frontiers in Education*, 5, 589965. [316]

Uebersax, J. S. (2006), "Introduction to the Tetrachoric and Polychoric Correlation Coefficients," Obtengido de [http://www.john-uebersax.com/stat/tetra.htm](http://www.john-uebersax.com/stat/tetra.htm). [Links]. [317]  
_______ (2010), "User Guide for polycorr 1.1" (advanced version). [317]