Maximum Likelihood Estimation of Multiple Correlations and Canonical Correlations with Categorical Data

Sik-Yum Lee
The Chinese University of Hong Kong

Wai-Yin Poon
University of California, Los Angeles

In the behavioral and social sciences, investigators frequently encounter latent continuous variables which are observable only in polytomous form. This paper considers the estimation of multiple correlations and canonical correlations for these variables. Two approaches, the maximum likelihood and the partitioned maximum likelihood, are established based on the corresponding multivariate polyserial and polychoric correlations. A simulation study was conducted to compare the various kinds of estimators.

In the behavioral and social sciences, scores for many continuous random variables are observable only in dichotomous or polytomous form. Examples of such variables include attitude items and performance items. Typically, respondents are asked to select answers from a scale such as

Agree strongly Agree No opinion Disagree Disagree strongly.

When analyzing this kind of data, a common approach is to proceed as if the data had an appropriate continuous distribution. Fortunately, some statistical methods are fairly robust to this kind of deviation. But there are still many situations that may lead to seriously erroneous results. For example, using the Pearson product-correlation to estimate the bivariate correlation may lead to biased results; moreover, the factor-analytic model is not robust to this kind of deviation (see Olsson, 1979a, 1979b). Because most multivariate methods are closely related to the estimate of the correlation (or covariance) matrix, the first important problem is to provide a statistically sound estimate of the correlation (or covariance) based on these data.

Let $\mathbb{Y} = (Y_1, \ldots, Y_s)'$ be an underlying continuous random vector, and let $\mathbb{Z}$ be an observable discrete random vector whose relation with $\mathbb{Y}$ is given by

$$Z_i = k(i) \quad \text{if} \quad \alpha_{i,k(i)} \leq Y_i < \alpha_{i,k(i)+1} \tag{1}$$

for $i = 1, \ldots, n$, $k(i) = 1, \ldots, h(i)$. Here $h(i)$ is the number of categories corresponding to the ith variable which are defined by a set of thresholds $\{\alpha_{i,1}, \ldots, \alpha_{i,k(i)+1}\}$ with $\alpha_{i,1} = -\infty$ and $\alpha_{i,h(i)+1} = \infty$. The correlation between a pair of variables within $\mathbb{Y}$ computed from random observations of $\mathbb{Z}$ is called the

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 11, No. 3, September 1987, pp. 317–323

0146-6216/87/030317-07$1.60

polychoric correlation. Suppose $\mathbb{X}$ represents another observed continuous vector. The correlations between $\mathbb{X}$ and $\mathbb{Y}$ obtained from random observations of $\mathbb{X}$ and $\mathbb{Z}$ are called the polyserial correlations.

Maximum likelihood (ML) estimation of polychoric and polyserial correlations of a pair of variables, say $\{Z_i,Z_j\}$ or $\{X_k,Z_j\}$, has received a good deal of attention (see, e.g., Cox, 1974; Martinson & Hamdan, 1971; Olsson, 1979a; Olsson, Drasgow, & Dorans, 1982; Tallis, 1962). The recent version of the LISREL computer program (Jöreskog & Sörbom, 1984) also contains options that give the polyserial and polychoric correlations. In practice, to produce the correlation matrix in a multivariate distribution, the ML estimates of the polychoric and polyserial correlations between a pair of variables are repeatedly computed until all of them are obtained. As pointed out by Olsson (1979a), this approach is not theoretically perfect, because (1) the estimates of the thresholds are not unique, and (2) the estimate of the covariance between the estimators is unavailable.

Poon and Lee (in press) developed the ML estimates based on the most general setting. In their approach, the polyserial correlations, the polychoric correlations, and the thresholds are estimated simultaneously, thus providing the optimal ML estimate of the parameters. Because of the length of time required for this full ML approach to converge, Poon and Lee developed the partitioned maximum likelihood (PML) approach, which requires much less computer time. Based on monte carlo results, they found that the ML and the PML estimates are very close to each other.

The studies cited above dealt only with bivariate correlation. As far as is known, no results on multiple correlations and canonical correlations have been published based on this kind of data. In this paper, ML estimates and PML estimates of the multiple correlations and canonical correlations are derived based on the observed data of $\mathbb{X}$ and $\mathbb{Z}$, as well as the variables in $\mathbb{Z}$. Various kinds of estimators are then compared by means of a monte carlo study.

Multiple Correlations and Canonical Correlations

Let $\mathbb{X}$ $(r \times 1)$ and $\mathbb{Y}$ $(s \times 1)$ be continuous random vectors which are jointly distributed according to a multivariate normal distribution, and let $\mathbb{V} = (\mathbb{X}',\mathbb{Y}')'$. Without loss of generality, the mean vector of $\mathbb{V}$ is assumed to be $\mathbb{0}$ and the correlation matrix of $\mathbb{V}$ is given by

$$\mathbb{\Omega} = \begin{bmatrix} \mathbb{\Omega}_{xx} & \mathbb{\Omega}_{xy} \\ \mathbb{\Omega}_{yx} & \mathbb{\Omega}_{yy} \end{bmatrix}, \tag{2}$$

where $\mathbb{\Omega}_{xx}$ is the $r \times r$ correlation matrix of $\mathbb{X}$, $\mathbb{\Omega}_{yy}$ is the $s \times s$ correlation matrix of $\mathbb{Y}$, and $\mathbb{\Omega}_{xy}$ is the $r \times s$ matrix of correlations between $\mathbb{X}$ and $\mathbb{Y}$. Now consider the situation that $\mathbb{Y}$ is unobservable. The information of $\mathbb{Y}$ is provided by the observed discrete random vector $\mathbb{Z}$ defined by Expression 1.

Let $\mathbb{V}_1$ and $\mathbb{V}_2$ be any two random subvectors of $\mathbb{V}$, and let v_j be a component of $\mathbb{V}_1$. Suppose the dimensions of $\mathbb{v}_1$ and $\mathbb{v}_2$ are equal to p_1 and p_2 respectively, with $p_1 \leq p_2$. The goal is to deduce the ML estimates of the multiple correlation between v_j and $\mathbb{V}_2$, and the ML estimates of the canonical correlations between $\mathbb{V}_1$ and $\mathbb{V}_2$, based on an available random sampling from $(\mathbb{X}',\mathbb{Z}')'$.

Let $\hat{\mathbb{\Omega}}$ be the ML estimate of the correlation matrix $\mathbb{\Omega}$ which contains the ML estimates of the polychoric and polyserial correlations. This matrix can be obtained by the method described in Poon and Lee (in press). Let

$$\hat{\mathbb{\Omega}} = \begin{bmatrix} \hat{\mathbb{\Omega}}_{11} & \hat{\mathbb{\Omega}}_{12} \\ \hat{\mathbb{\Omega}}_{21} & \hat{\mathbb{\Omega}}_{22} \end{bmatrix} \tag{3}$$

be a rearranging partition of $\hat{\mathbb{\Omega}}$, where $\hat{\mathbb{\Omega}}_{11}$ and $\hat{\mathbb{\Omega}}_{22}$ are the estimated correlation matrices of $\mathbb{V}_1$ and $\mathbb{V}_2$

respectively, and $\hat{\Omega}_{21}$ is the estimated correlation matrix of $\mathbb{V}_1$ and $\mathbb{V}_2$. Then from Anderson (1958, p. 87), the ML estimate of the multiple correlation between v_j and $\mathbb{V}_2$ is given by

$$\hat{R}_{j\cdot p_1+1,\ \ldots,\ p} = (\hat{\omega}_j \hat{\Omega}_{22}^{-1} \hat{\omega}_j')^{1/2} \quad , \tag{4}$$

where ω_j is the jth row of $\hat{\Omega}_{12}$. Moreover, let $\hat{\lambda}_1, \ldots, \hat{\lambda}_{p_2}$ be the ML estimate of the canonical correlation between $\mathbb{V}_1$ and $\mathbb{V}_2$. These estimates can be obtained as the roots of

$$\begin{vmatrix} -\lambda\hat{\Omega}_{11} & \hat{\Omega}_{12} \\ \hat{\Omega}_{21} & -\lambda\hat{\Omega}_{22} \end{vmatrix} = 0 \quad . \tag{5}$$

The ith canonical variates $\hat{\alpha}_i$, $\hat{\gamma}_i$ corresponding to $\hat{\lambda}_i$ are solutions to

$$\begin{bmatrix} -\hat{\lambda}_i\hat{\Omega}_{11} & \hat{\Omega}_{12} \\ \hat{\Omega}_{21} & -\hat{\lambda}_i\hat{\Omega}_{22} \end{bmatrix} \begin{bmatrix} \hat{\alpha}_i \\ \hat{\gamma}_i \end{bmatrix} = 0 \quad , \tag{6}$$

subject to $\hat{\alpha}_i'\hat{\Omega}_{11}\hat{\alpha}_i = 1$ and $\hat{\gamma}_i'\hat{\Omega}_{22}\hat{\gamma}_i = 1$ (see Anderson, 1958, p. 299). Thus, it is easy to obtain the estimates of multiple correlations and canonical correlations after $\hat{\Omega}$ has been obtained. Of course, these estimates will possess the desirable statistical properties of the general ML estimates.

When the dimension of $\mathbb{Z}$ is large, it requires a great deal of computer time to obtain $\hat{\Omega}$. To remedy this situation, a procedure called the partitioned maximum likelihood (PML) method has been proposed by Poon and Lee (in press). In this method, $\mathbb{Z}$ is partitioned into $\{Z_1, \ldots, Z_n\}$. For each $i = 1, \ldots, n$, the polyserial correlations of $\mathbb{X}$ and Y_i are estimated based on the observed random observations corresponding to $(\mathbb{X},Z_i)$ (see Lee & Poon, 1986). The polychoric correlation between any pair of random variables in $\mathbb{Y}$ is estimated based on a two-way contingency table of the corresponding observed frequencies. It has been shown by Poon and Lee (in press) that this method requires much less computer time than the full ML method.

Let $\tilde{\Omega}$ be the PML estimate of Ω and let

$$\tilde{\Omega} = \begin{bmatrix} \tilde{\Omega}_{11} & \tilde{\Omega}_{12} \\ \tilde{\Omega}_{21} & \tilde{\Omega}_{22} \end{bmatrix} \tag{7}$$

be a partition of $\tilde{\Omega}$ as in Equation 2. The PML estimate of the multiple correlation between v_j and $\mathbb{V}_2$ is defined by

$$\tilde{R}_{j\cdot p_1,\ \ldots,\ p_2} = (\tilde{\omega}_j \tilde{\Omega}_{22}^{-1} \tilde{\omega}_j')^{1/2} \quad , \tag{8}$$

where $\tilde{\omega}_j$ is the jth row of $\tilde{\Omega}_{12}$. The PML estimates $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{p_1}$ of the canonical correlation are defined as the roots of the corresponding characteristic equation

$$\begin{vmatrix} -\lambda\tilde{\Omega}_{11} & \tilde{\Omega}_{12} \\ \tilde{\Omega}_{21} & -\lambda\tilde{\Omega}_{22} \end{vmatrix} = 0 \quad . \tag{9}$$

Similarly, the ith canonical variates $\tilde{\alpha}_i$, $\tilde{\gamma}_i$ corresponding to $\tilde{\lambda}_i$ are defined as solutions to

$$\begin{bmatrix} -\tilde{\lambda}_i\tilde{\Omega}_{11} & \tilde{\Omega}_{12} \\ \tilde{\Omega}_{21} & -\tilde{\lambda}_i\tilde{\Omega}_{22} \end{bmatrix} \begin{bmatrix} \tilde{\alpha}_i \\ \tilde{\gamma}_i \end{bmatrix} = 0 \quad , \tag{10}$$

subject to $\tilde{\alpha}_i\tilde{\Omega}_{11}\tilde{\alpha}_i = 1$ and $\tilde{\gamma}_i\tilde{\Omega}_{22}\tilde{\gamma}_i = 1$.

Monte Carlo Comparison

Method

A monte carlo study was conducted to compare the effectiveness of the various estimates for the multiple correlation and the canonical correlations. The study was based on a simulated dataset with sample sizes $N = 100$ and $N = 200$ from a multivariate normal distribution with mean vector $\mathbf{0}$ and population correlation matrix

$$\begin{bmatrix} 1.0 & .2 & .6 & .3 \\ .2 & 1.0 & .3 & .2 \\ .6 & .3 & 1.0 & .3 \\ .3 & .2 & .3 & 1.0 \end{bmatrix}.$$

Two situations were considered. In Case I, the first two components of the random vector were taken to form the continuous observable random vector $\mathbf{X}$, and the remaining two variables were taken as the latent random vector $\mathbf{Y}$. The standardized random vector $(\mathbf{x}_i', \mathbf{y}_i')'$ was transformed to $(\mathbf{x}_i', \mathbf{z}_i')'$ according to the following pre-assigned thresholds:

$$\boldsymbol{\alpha}_1 = [-\infty, -1.1, .5, \infty]$$
$$\boldsymbol{\alpha}_2 = [-\infty, .6, 1.2, \infty]$$

Thus, in this case, $\boldsymbol{\Omega}$ involved both the polyserial and polychoric correlations. In Case II, all components in the random vector were taken to form the 4×1 latent random vector $\mathbf{Y}$, and there was no continuous observable random vector. The standardized random vector $\mathbf{y}_i$ was transformed to $\mathbf{z}_i$ according to the following pre-assigned thresholds:

$$\boldsymbol{\alpha}_1 = [-\infty, -1.0, .5, \infty]$$
$$\boldsymbol{\alpha}_2 = [-\infty, 0, 1.2, \infty]$$
$$\boldsymbol{\alpha}_3 = [-\infty, -1.1, .5, \infty]$$
$$\boldsymbol{\alpha}_4 = [-\infty, .6, 1.2, \infty]$$

In this case, only polychoric correlations were involved in $\boldsymbol{\Omega}$. For each sample, $\hat{\boldsymbol{\Omega}}$ and $\tilde{\boldsymbol{\Omega}}$ were computed by the program developed by Poon and Lee (in press). Then, from these matrices, the ML estimates and the PML estimates of the multiple correlation and the canonical correlation were respectively computed. Also, for comparison, the sample correlation matrix was computed based on the original continuous data $\{\mathbf{x}_i, \mathbf{y}_i\}$ in the first case, and $\{\mathbf{y}_i\}$ in the second case; the corresponding multiple correlation estimates and canonical correlation estimates were then computed, resulting in the ML estimates when continuous data are available. These were called the continuous maximum likelihood (CML) estimates. Additionally, the multiple correlation and canonical correlations were computed directly, based on the sample Pearson correlation matrices obtained from the discrete data $\{\mathbf{x}_i, \mathbf{z}_i\}$ and $\{\mathbf{z}_i\}$. These were called the direct method (DM) estimates. As expected, these DM estimates are inferior; they are compared below with the other estimates.

For each sample size, 50 replications were generated. The means of the estimates and various root mean squared errors, such as

$$\text{RMSE} = \left[\sum_{i=1}^{50} (\hat{\theta}_i - \tilde{\theta}_i)^2/50\right]^{1/2}, \tag{11}$$

were computed, where θ is either a multiple correlation or a canonical correlation.

Results

Table 1 reports the simulation results on the multiple correlation $R_{1.34}$. Other multiple correlation estimates behaved similarly, and thus are not reported. The RMSE columns A-B, A-C, and A-D were used to examine the discrepancy between the various estimates and the CML estimate. This discrepancy is more noteworthy than the discrepancy between the population value and the other estimates, because if the simulated sample of size N is considered based on the continuous variables, its multiple correlation cannot be exactly equal to the population value due to sampling error. The last RMSE column was used to examine the difference between the ML and the PML estimates. From Table 1, the following interesting phenomena can be observed:

1. The ML estimates performed very well. The true population value of $R_{1.34}$ is .613, hence the mean of the estimates is very close to the population value and it is almost identical to the mean of the CML estimates. Moreover, the corresponding RMSE values are very low.
2. The PML estimates also performed very well. In fact, there is very little bias evidence for the PML estimates and the ML estimates.
3. The DM estimates are significantly worse than the other estimates. From the results on means, the DM estimates are seen to be negatively biased. Moreover, in the RMSE column A-D, the differences between the CML and the DM estimates are substantial.
4. As expected, increasing the sample size decreases the RMSE. Also, results for Case I are better than for Case II, because more continuous variables are unobserved in Case II.

Table 2 reports the simulation results on the canonical correlations between the first two variables and the last two variables. The first and the second population canonical correlations are equal to .645 and .050, respectively. Results on the first principal canonical correlation estimates are very similar to the results of the multiple correlation estimate. Therefore, analogous interpretations are appropriate. For the less important second canonical correlation, the means of the DM estimates are closer to the population value and the means of the CML estimates than are those from the ML and PML estimates. However, their corresponding RMSEs are significantly worse. Therefore, it is still possible to conclude that the ML and the PML estimates are the better estimates.

Discussion

On the basis of the results presented above, the following conclusions are offered. For estimating the multiple correlation and the canonical correlations of latent continuous variables, the ML and PML

Table 1
Simulation Results for Multiple Correlations

Sample Size	Mean CML (A)	Mean ML (B)	Mean PML (C)	Mean DM (D)	RMSE A-B	RMSE A-C	RMSE A-D	RMSE B-C
Case I								
N = 100	.623	.625	.625	.356	.038	.038	.075	.004
N = 200	.628	.626	.626	.560	.027	.027	.072	.002
Case II								
N = 100	.619	.634	.631	.467	.068	.071	.163	.021
N = 200	.624	.615	.619	.456	.057	.059	.166	.012

Table 2
Simulation Results for Canonical Correlations

Canonical Variate	Sample Size	Mean CML (A)	Mean ML (B)	Mean PML (C)	Mean DM (D)	RMSE A-B	RMSE A-C	RMSE A-D	RMSE B-C
1	Case I								
	N = 100	.651	.655	.656	.581	.079	.077	.150	.009
	N = 200	.659	.655	.656	.583	.061	.060	.119	.005
	Case II								
	N = 100	.645	.662	.658	.498	.083	.086	.152	.028
	N = 200	.652	.646	.649	.498	.065	.068	.114	.021
2	Case I								
	N = 100	.088	.124	.123	.093	.035	.035	.077	.003
	N = 200	.068	.095	.094	.069	.026	.026	.080	.002
	Case II								
	N = 100	.091	.124	.127	.081	.065	.066	.159	.019
	N = 200	.070	.084	.089	.056	.051	.052	.160	.014

estimators have very small biases to the CML estimator. The ML and PML estimates are extremely close to each other. Therefore, because implementation of the ML approach requires a great deal of computer time, the PML approach represents an extremely attractive method for obtaining these correlations. The estimates computed using the DM method might be acceptable if the available data are symmetric. However, if the data are not symmetric, as in the situation considered here, estimates computed by the DM method are misleading. Furthermore, if the data are not symmetric, the DM method gives bias estimates for the polychoric and polyserial correlations even with 5 or 7 categories. Hence it can also be expected to give bias estimates for the corresponding multiple correlations and canonical correlations.

It should be noted that the estimates obtained from the ML approach possess optimal statistical properties. Hence, tets of significance of the canonical correlation are available in the usual manner (see Bock, 1975, p. 391). However, the analogous test statistics corresponding to the PML approach require further investigation.

References

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. New York: Wiley.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.

Cox, N. R. (1974). Estimation of correlation between a continuous and a discrete variable. *Biometrics, 30*, 171–178.

Jöreskog, K. G., & Sörbom, D. (1984). *LISREL VI user's guide*. Chicago: National Educational Resources.

Lee, S.-Y., & Poon, W. Y. (1986). Maximum likelihood estimation of polyserial correlations. *Psychometrika, 51*, 113–121.

Martinson, E. O., & Hamdan, M. A. (1971). Maximum likelihood and some other asymptotically efficient estimators of correlation in two way contingency tables. *Journal of Statistical Computation and Simulation, 1*, 45–54.

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

Olsson, U. (1979b). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research, 14*, 485–500.

Olsson, U., Drasgow, F., & Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika, 47*, 337–347.

Poon, W. Y., & Lee, S.-Y. (in press). Maximum likelihood estimation of multivariate polyserial and polychoric correlation coefficients. *Psychometrika*.

Tallis, G. (1962). The maximum likelihood estimation

of correlation from contingency tables. *Biometrics, 18*, 342–352.

Acknowledgments

This research was supported by a research grant (DA01070) from the U.S. Public Health Service. The assistance of J. Speckart in manuscript production is also gratefully acknowledged.

Author's Address

Send requests for reprints or further information to S. Y. Lee, Department of Statistics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong.